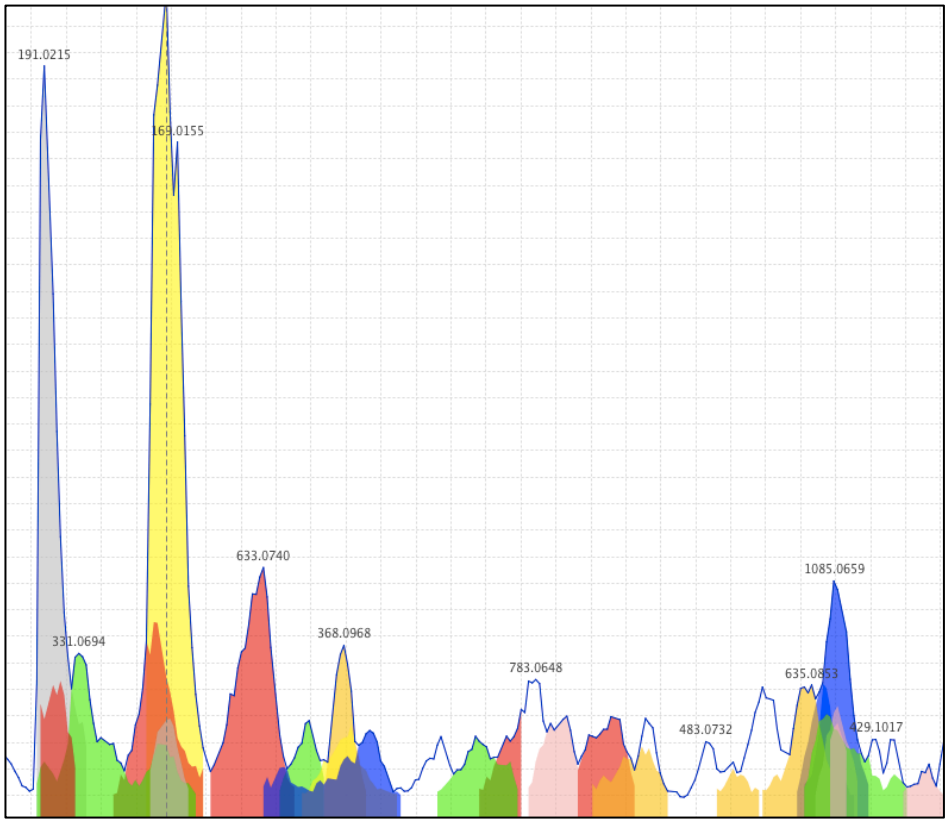
UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Section des sciences pharmaceutiques

Ecole de Pharmacie
EPGL
Genève – Lausanne

UNIL | Université de Lausanne

MZmine: a tutorial



Joëlle Houriet, PhD
Pierre-Marie Allard, PhD

oaa accredited in Switzerland

Section des sciences pharmaceutiques
Prof. Jean-luc Wolfender
Phytochimie et Produits naturels bioactifs
Faculté des sciences
30 quai Ernest-Ansermet
CH-1211 Genève 4

Université de Genève, Université de Lausanne
Tél. 41 22 379 33 85 - Fax 41 22 379 3399
www.unige.ch/sciences/pharm/fasie

Table of contents

This file contains several chapters, to help you to use the MZmine software. It's an addition to the official **User's manual** and **Tutorial of MZmine** and doesn't intend to substitute them.

The chapters are:

- Installation of the softwares on Windows (MZmine, Java3D, R)
- MS-Data Conversion
- Example of the data processing on MZmine

Installation of MZmine on Windows

Here are only the points that are not into the User's manual of MZmine. The User's manual and the official explanations are found in the website:
http://mzmine.sourceforge.net/download.shtml.
This tutorial explains details that are not or that are different in the manual. As new versions are regularly available, maybe the explanations are not updated. Visit the website to solve your eventual problems. The mailing list can be consulted in case of problem:
http://sourceforge.net/p/mzmine/mailman/search/?q=&mail_list=mzmine-devel

Java3D

Download this file:
J3d-xxx-windows-amd64.zip
The zip goes normally in this folder:
Program files\java\java3D\1.5.2
Program file\java is divided in two parts: **Jre$_x$** and **Java3D**.
Jre$_x$ runs java3D, but **Jre$_x$** doesn't know how to find Java3D. To solve this, we have to copy a few files of Java3D into **Jre$_x$**:
From java3D\xxx\lib\ext to jerx\lib\ext, it concerns three jar files:
j3dcore.jar
j3dutils.jar
vecmath.jar
From java3D\xxx\bin to jerx\bin :
J3dcore-ogl.dll

Software R

Install it and enter the command given in the manual:

```
install.packages(c("rJava", "ptw", "gplots"))
source("http://bioconductor.org/biocLite.R")
biocLite("xcms")
biocLite("CAMERA")
```

Save when R asks it.
To check if R works, try the deconvolution step with Wavelets (XCMS) algorithm (see further).
If R doesn't work, edit the file **startMZmine_windows.bat** with a text editor and modify it as followed. The lines to modify are in blue, the comments are in red. Be careful, each character has its place!

```
@echo off

rem Obtain the physical memory size
for /f "skip=1" %%p in ('wmic os get totalvisiblememorysize') do if not defined TOTAL_MEMORY=set TOTAL_MEMORY=%%p

rem The HEAP_SIZE variable defines the Java heap size in MB.
rem That is the total amount of memory available to MZmine 2.
rem By default we set this to the half of the physical memory
rem size, but feel free to adjust according to your needs.
set /a HEAP_SIZE=%TOTAL_MEMORY% / 4096 / 2

rem Check if we are running on a 32-bit system.
rem If yes, force the heap size to 4096 MB.
for /f "skip=1" %%x in ('wmic cpu get addresswidth') do if not defined ADDRESS_WIDTH set ADDRESS_WIDTH=%%x
if %ADDRESS_WIDTH%==32 (
  set HEAP_SIZE=1024
)

rem The TMP_FILE_DIRECTORY parameter defines the location where temporary
rem files (parsed raw data) will be placed. Default is %TEMP%, which
rem represents the system temporary directory.
set TMP_FILE_DIRECTORY=%TEMP%

rem Set R environment variables.
set R_HOME=C:\Program Files\R\R-3.1.0
set R_SHARE_DIR=C:\Program Files\R\R-3.1.0
set R_INCLUDE_DIR=C:\Program Files\R\R-3.1.0\include
set R_DOC_DIR=C:\Program Files\R\R-3.1.0\doc
set R_LIBS_USER=C:\Users\xxxxxx\documents\R\win-library\3.1  OR  C:\Program Files\R\R-3.1.0\library (Look where is your library)

rem Include R DLLs in PATH.
set PATH=%PATH%;%R_HOME%\bin\x64

rem The directory holding the JRI shared library (libjri.so).
set JRI_LIB_PATH=%R_LIBS_USER%\rJava\jri\x64

rem It is usually not necessary to modify the JAVA_COMMAND parameter, but
rem if you like to run a specific Java Virtual Machine, you may set the
rem path to the java command of that JVM
set JAVA_COMMAND=java

rem It is not necessary to modify the following section
set JAVA_PARAMETERS=-XX:+UseParallelGC -Djava.io.tmpdir=%TMP_FILE_DIRECTORY% -Xms%HEAP_SIZE%m -Xmx%HEAP_SIZE%m -Djava.library.path="%JRI_LIB_PATH%"
```

```
set CLASS_PATH=lib\MZmine-2.10.jar
set MAIN_CLASS=net.sf.mzmine.main.MZmineCore

rem Show java version, in case a problem occurs
%JAVA_COMMAND% -version

rem This command starts the Java Virtual Machine
%JAVA_COMMAND% %JAVA_PARAMETERS% -classpath %CLASS_PATH% %MAIN_CLASS% %*

rem If there was an error, give the user chance to see it
IF ERRORLEVEL 1 pause
```

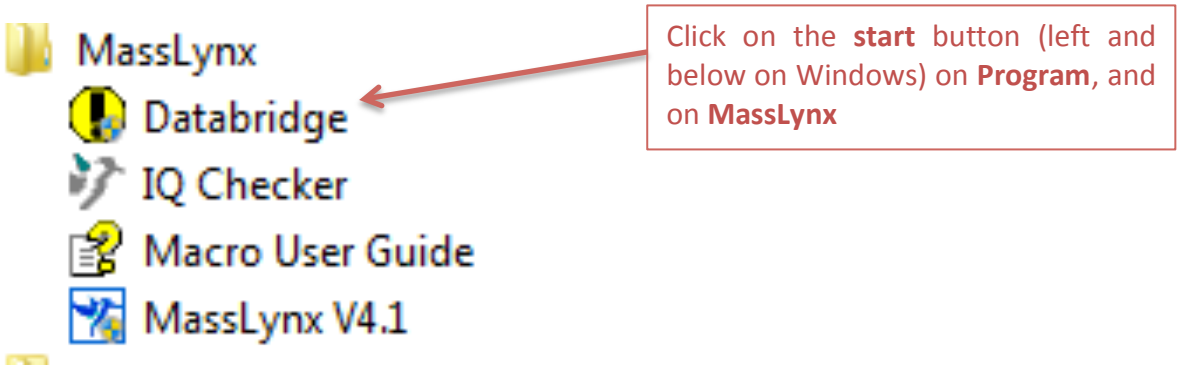
MS-Data conversion.

The idea is to pass from proprietary format (Waters, Thermo, Agilent etc …) to open mass format. The most common actually is **.mzXML.** (see details about MS data formats here http://en.wikipedia.org/wiki/Mass_spectrometry_data_format) This conversion step offers various advantages:

- Treat data coming from different mass spectrometers on the same analysis program.
- Exchange your MS data with other labs
- Use a range of open-source software for the data analysis.

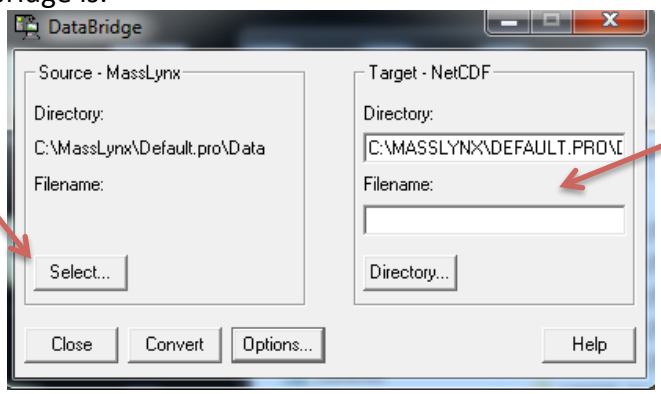
Waters .RAW data (TOF)

For conversion of data acquired on the TOF, Databridge (the Waters conversion tool found with MassLynx) should be used.
Caution: using Proteowizzard to convert Waters data will lead to inaccurate masses in the .mzXML file !



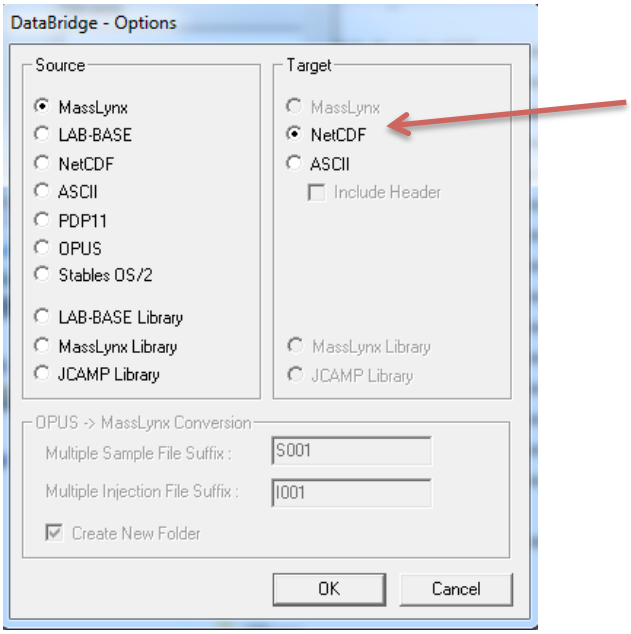
The general window of Databridge is:

.RAW file are selected on the left panel.
.RAW file name should not contain space or special caracters (!, é …) change spaces for _



Choose here where the converted files have to be saved.

By clicking on **Options** this panel is opened. Select NetCDF as target.

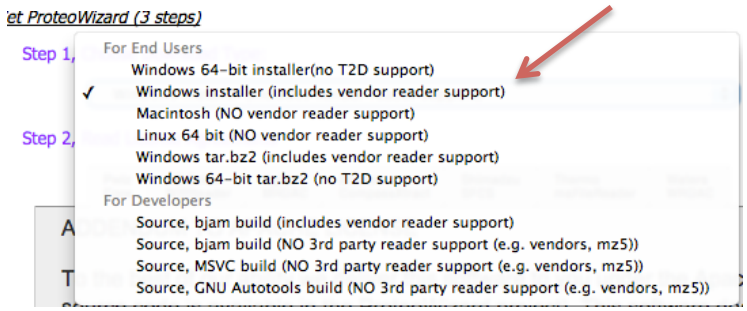


Each FILE_A.raw file is converted into two files FILE_A_01.netcdf and FILE_A_02.netcdf FILE_A_02.netcdf should be a lighter file, it corresponds to the lockmass trace and can be deleted. Actually, MZmine is able to read the format of files from Thermo and Waters directly, but for the Waters files (TOF) for example, MZmine isn't able to remove the signal of the lockmass. It's better to work with a standard format (like mxml or netcdf) that allows more possibilities.
It is possible to select multiple files at once but for each file a message windows will prompt and ask to click for OK to continue to the next one. So if a large number of files has to be converted, a solution is to place something heavy on the enter key of the keyboard and go for a coffee!

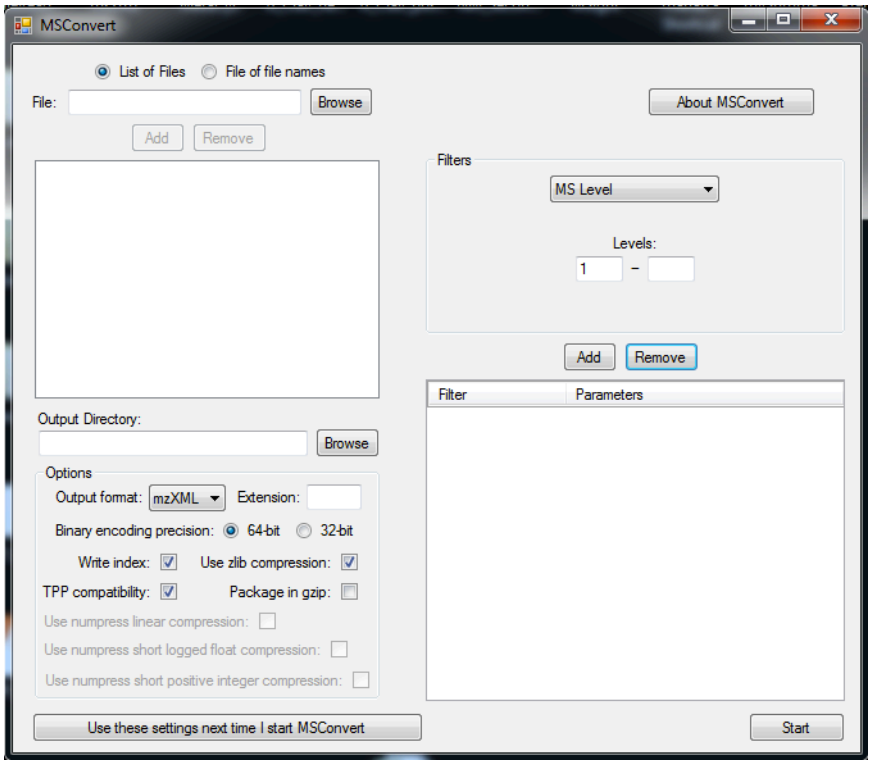
Thermo RAW files (Orbitrap)

With the Orbitrap, the normal spectrum (MS1) and the fragmentation (MS2 or MS/MS) can be acquired.
Thermo RAW files can be converted to .mzXML using Proteowizzard.
Proteowizzard can be downloaded here:
http://proteowizard.sourceforge.net/downloads.shtml

Be sure to select the **Windows installer (includes vendor support)** when you download it.



There are various settings in the MSconvert module depending on the filters to apply during conversion and the output format.
To convert the whole MS data to .mzXML the following settings can be applied.



By this way, the normal spectrum and the fragmentation will be converted. If the fragmentation (MS/MS) isn't needed, in the levels boxes, write 1 -1. And if only the fragementation is needed, write 2-2.

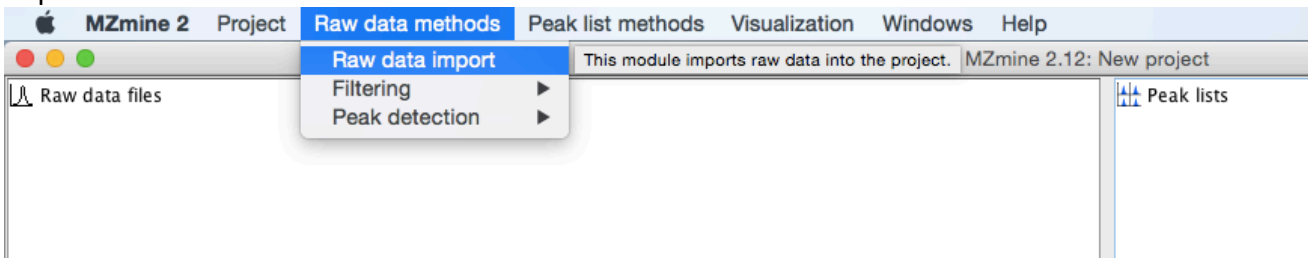
Data processing on MZmine

This chapters contains:

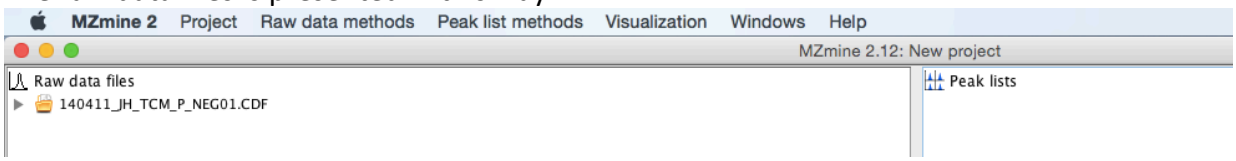
- Raw data import and filter
- Peak detection
- Isotopic peak grouper
- Identification

Raw data import

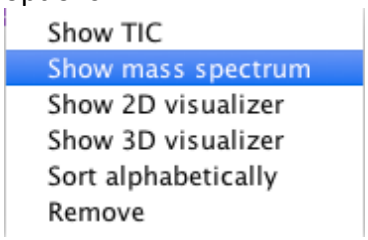
Import of the data from a .CDF file



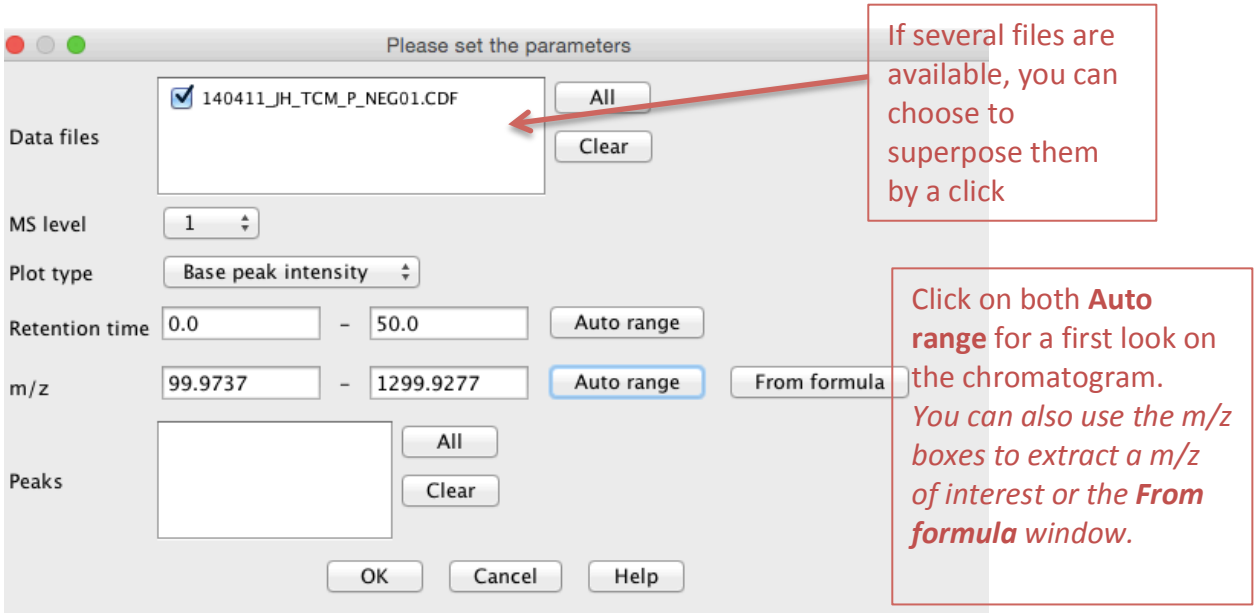
The raw data files is presented in this way:



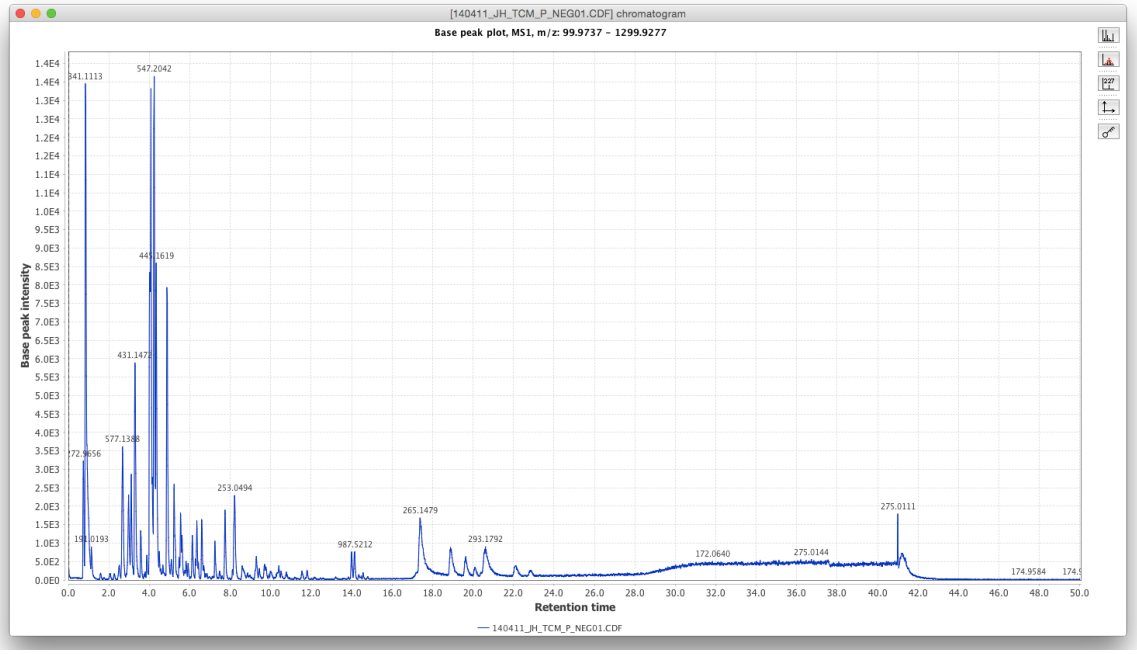
Once the data is imported, right click on the yellow data icon to reveal several display options.



Show TIC offers the option of Base peak or TIC and allows you to set various ranges.
A double click on the raw data icon opens the TIC options windows, too.

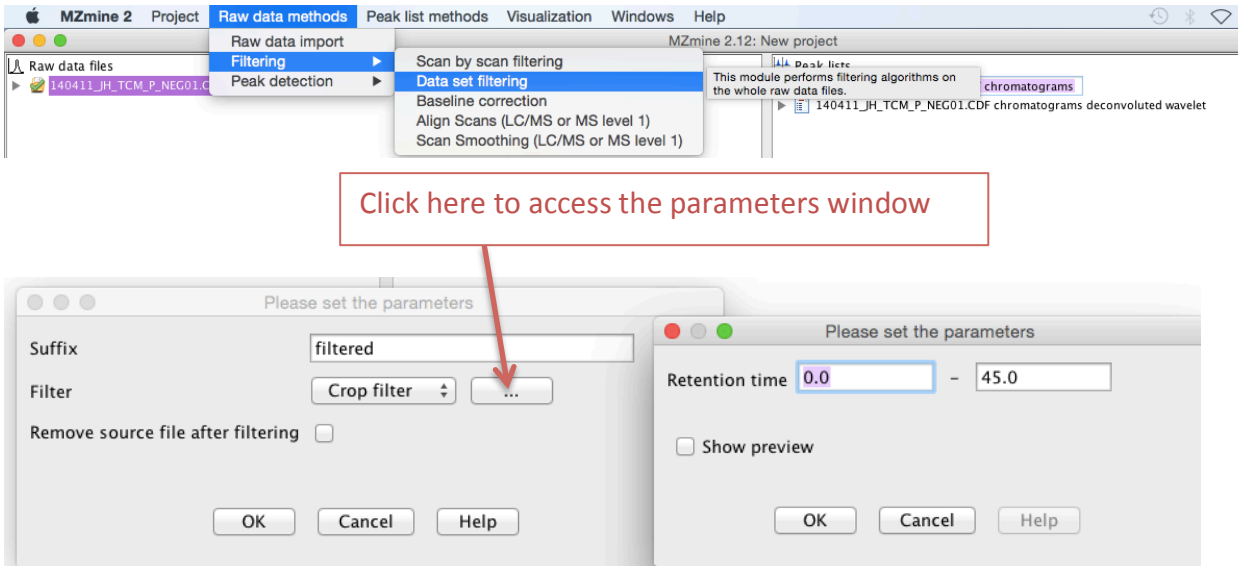


The chromatogram is presented as hereunder. The plot is zoomable by clicking and dragging to the right. Double clicking a peak opens its mass spectrum. Clicking and dragging upward or to the left results in zooming back out to maximum zoom. Clicking and dragging downwards or to the right zooms in.



Take a note of the height of the baseline (or noise level) at different time and the height of the smallest peaks. You will need these values later!
Visualization of MS2 data is possible in the same window.

The possibility to remove the last part of the chromatogram is given with the **filtering** function:



The visualization of the filtering is possible when you click “show preview”.

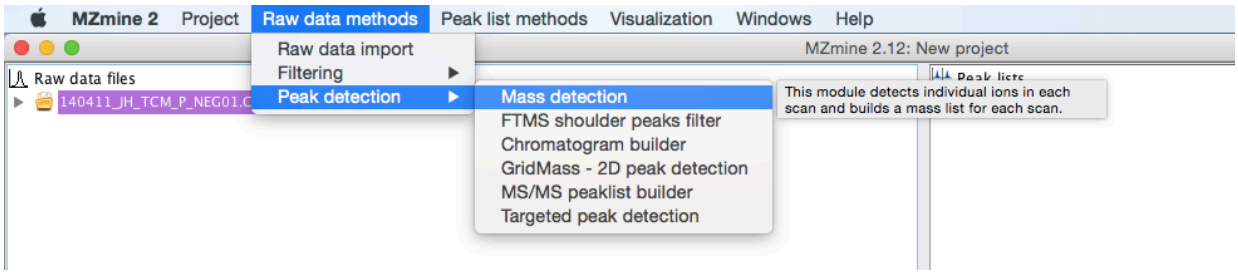
Peak detection

Peak detection is a three steps process:

1. Mass detection
2. Chromatogram building
3. Peak deconvolution

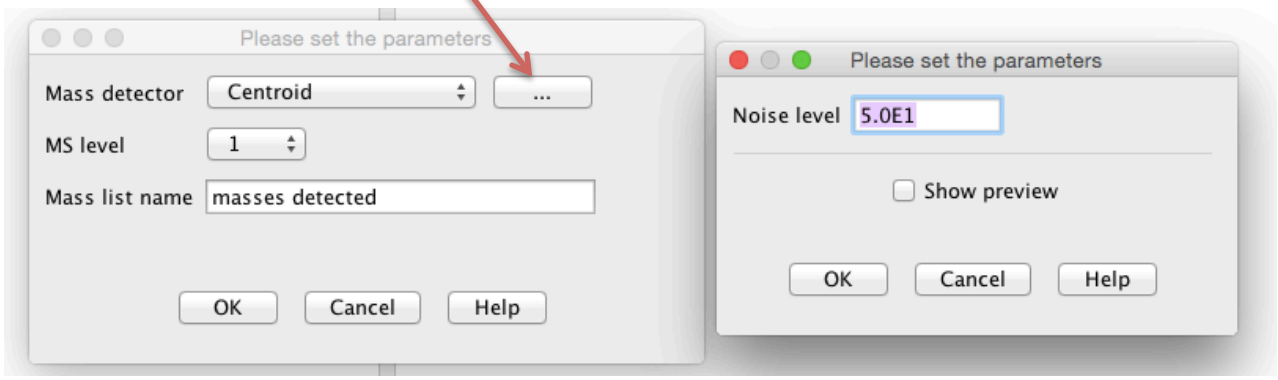
1. Mass detection

Click on Raw data methods/Peak Detection/Mass detection



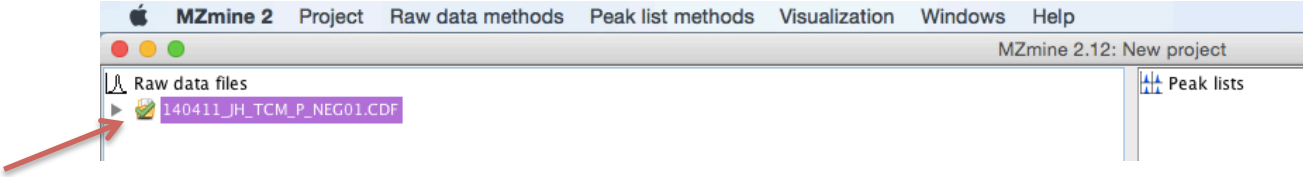
Comment: the FTMS shoulder peaks filter isn't necessary with our kind of analysis.

On the window, click on the … to open the parameters windows. Set the noise level with the value you noted on the chromatogram. You can click on "show preview" to see the spectrum.



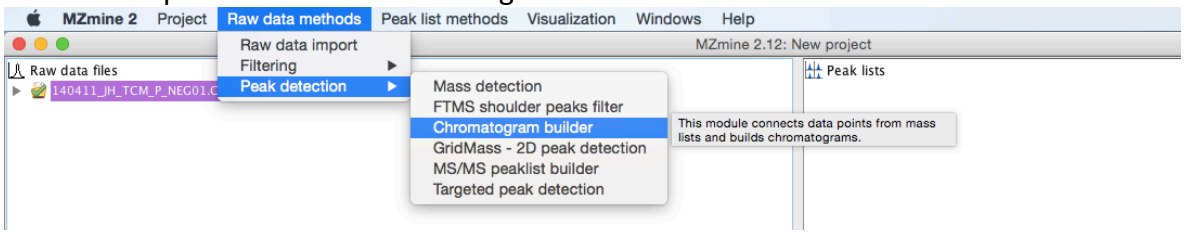
Change MS level to "2" to detect MS2 data peaks (be sure to adapt the noise level as it can be lower than with MS1 data.
Comment: with the TOF and Orbitrap, the data are generally acquired in centroid mode.

When the masses are detected, the icon will show a green tick mark:

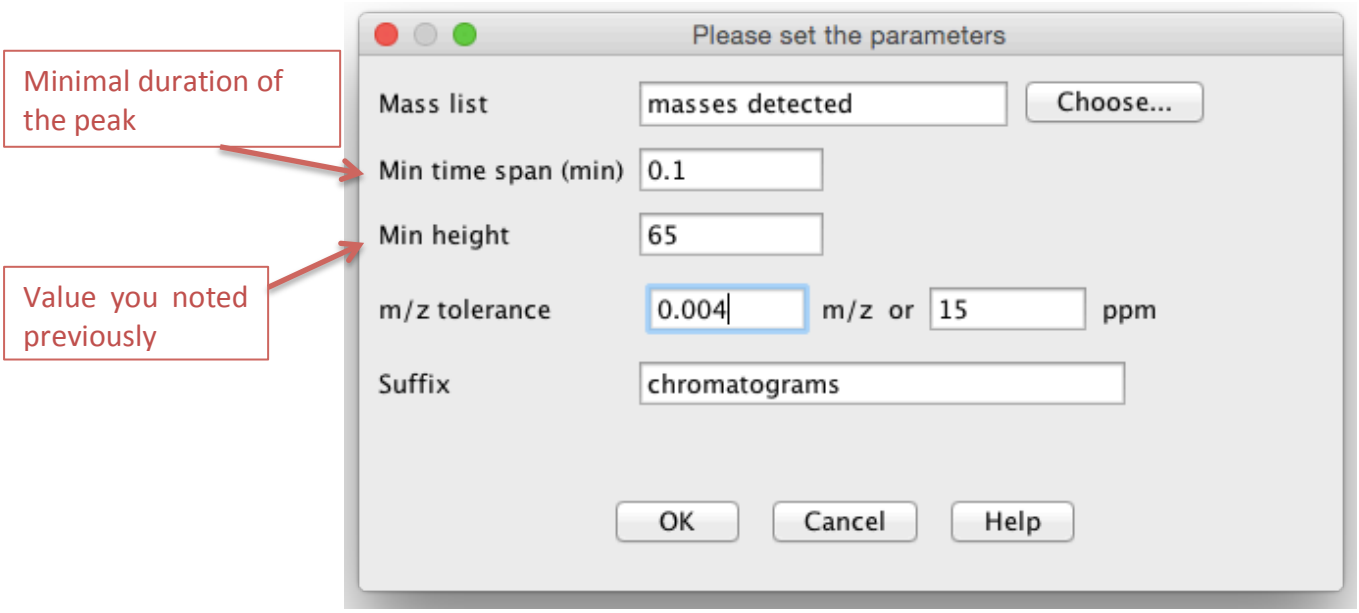


2. Chromatogram building

The next step is to build the chromatogram:



This window presents these parameters:



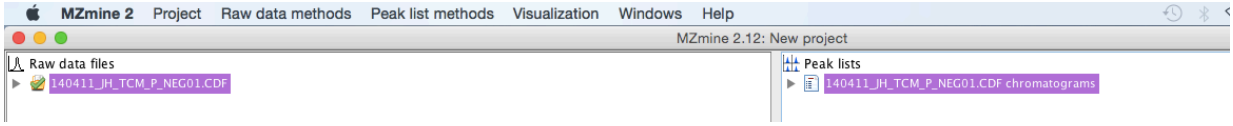
With UHPLC, the min time span has to be below 0.1 min, as the chromatographic peak are really thin.

Comment: about the m/z tolerance and ppm: the box with m/z is the absolute difference (given normally in Da *or in* amu *or* u *(unified atomic mass unit). The ppm is the relative tolerance. MZmine calculates the range of tolerance with the maximum of the absolute and relative tolerances.*
For information:

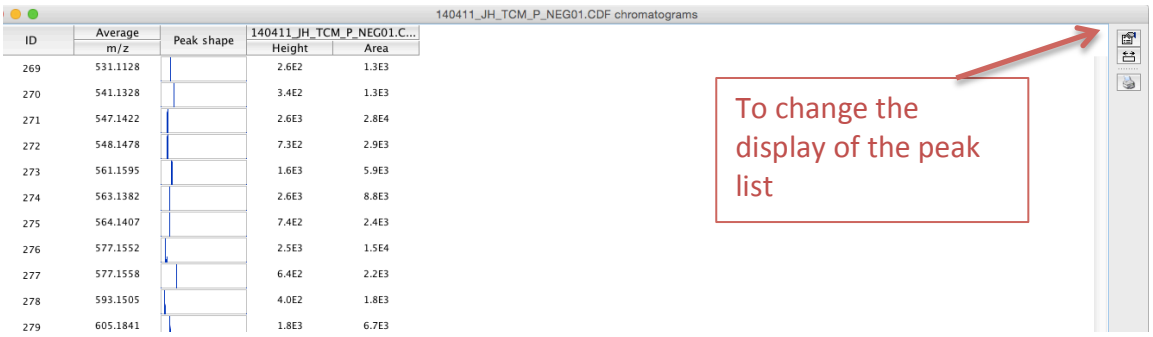
$$ppm = \frac{Observed\ mass - Calculated\ mass}{Calculated\ mass} * 10^6$$

With the TOF, a range between 10 to 15 ppm is acceptable (from 0.003 to 0.004 m/z).
As the Orbitrap has a better resolution, the range can be decreased to 5 ppm (below 0.0015 m/z).

A peak list is done by this step of chromatogram building:

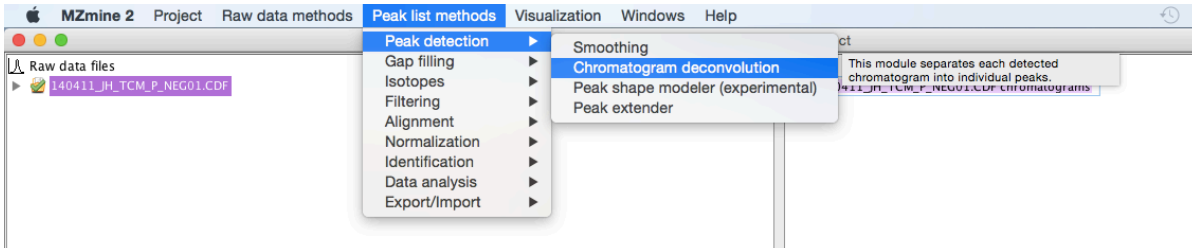


Double click on the peak list to open it:

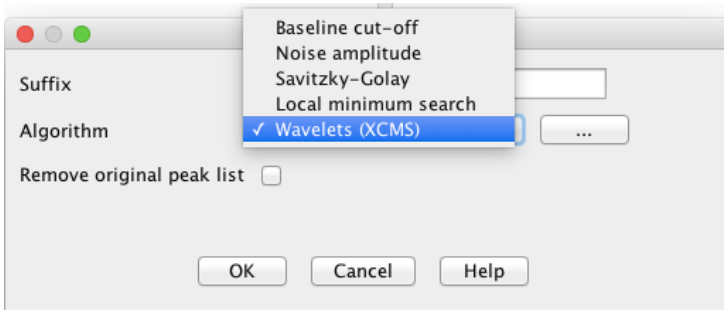


3. Deconvolution

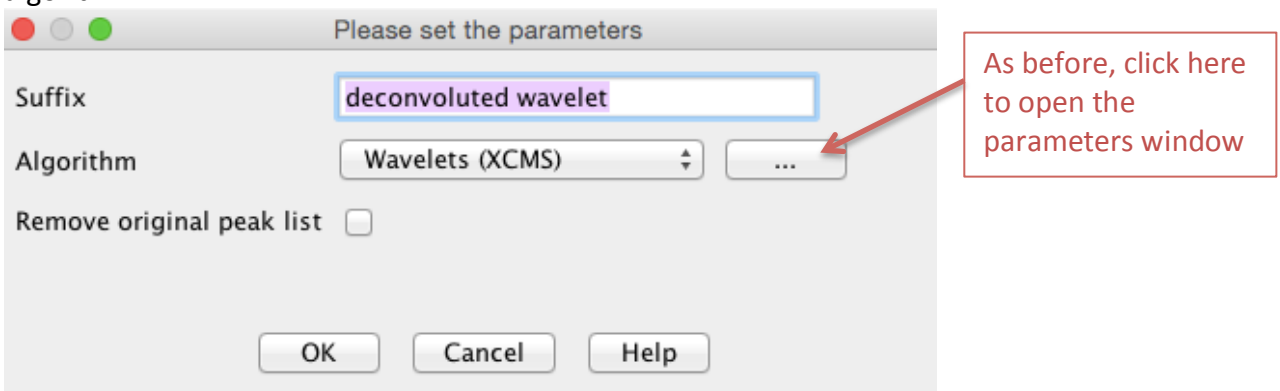
The third step is the deconvolution. Make sure you highlight the chromatograms list in the right pane (left in the old version). The deconvolution step separates every detected mass, that can occurs at different times into one individual peak:



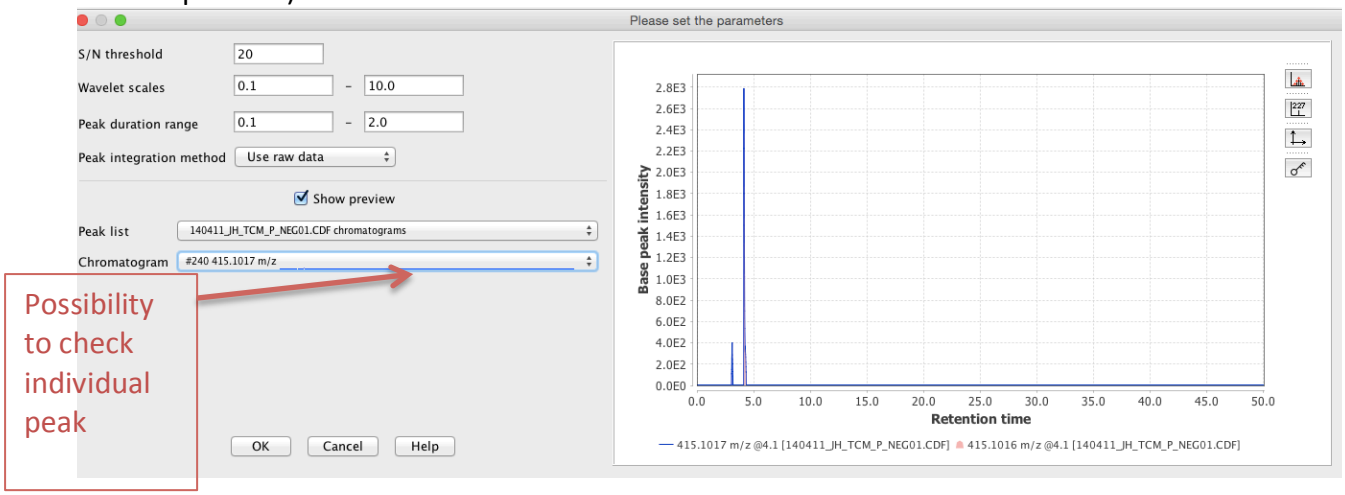
Different algorithms are proposed:



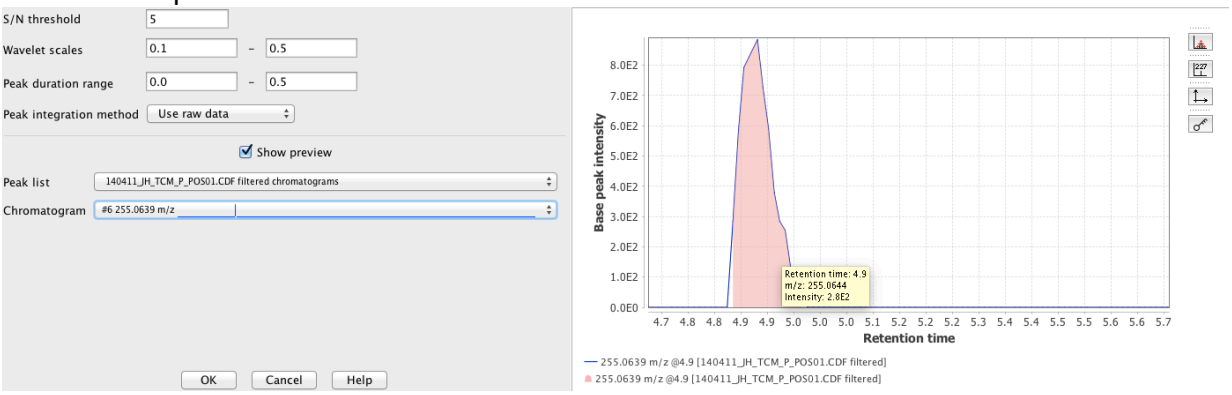
On the suffix box, write the name of the algorithm you choose, mainly if you try different algorithm:



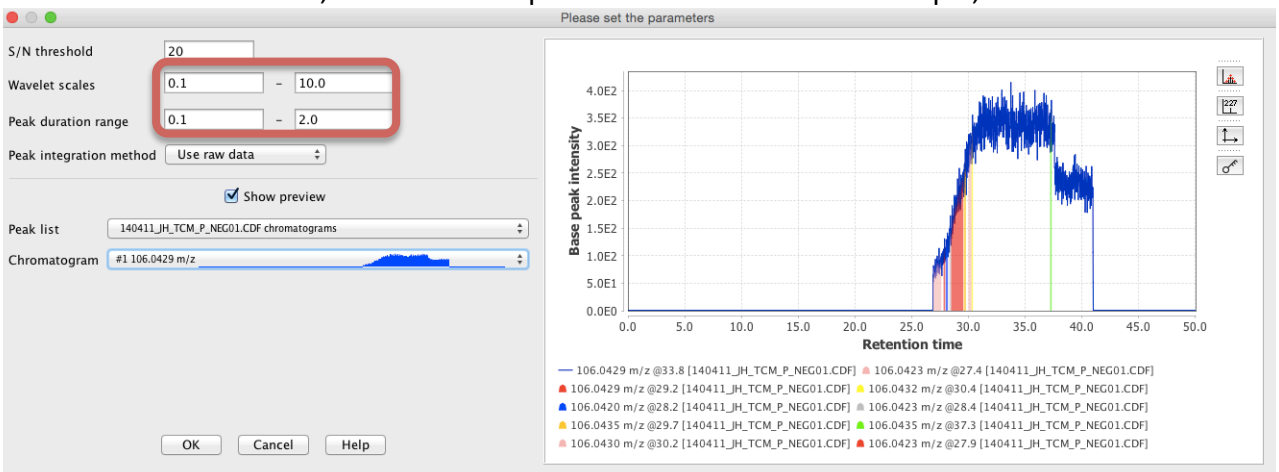
The most standard algorithm is the **Wavelets (XCMS).** To use it, you have to install the software R (see the first part of this tutorial). The wavelets settings are the following (tick the Show preview):



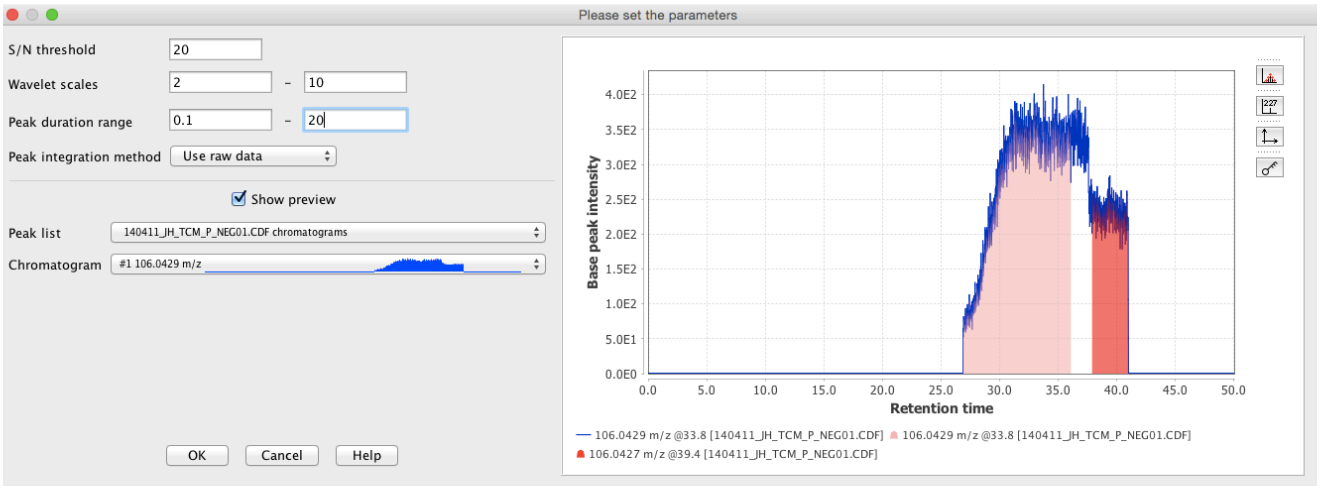
The **Help** explains the different parameters. Make them vary to observe the differences. To start, focus on the peaks with highest intensities (write their ID number from the peak list "Chromatogram"). Check if the deconvolution is adapted, as it's the case above. The blue line represents the signal and the pink color represents the observed peak. You can still zoom on the peak:



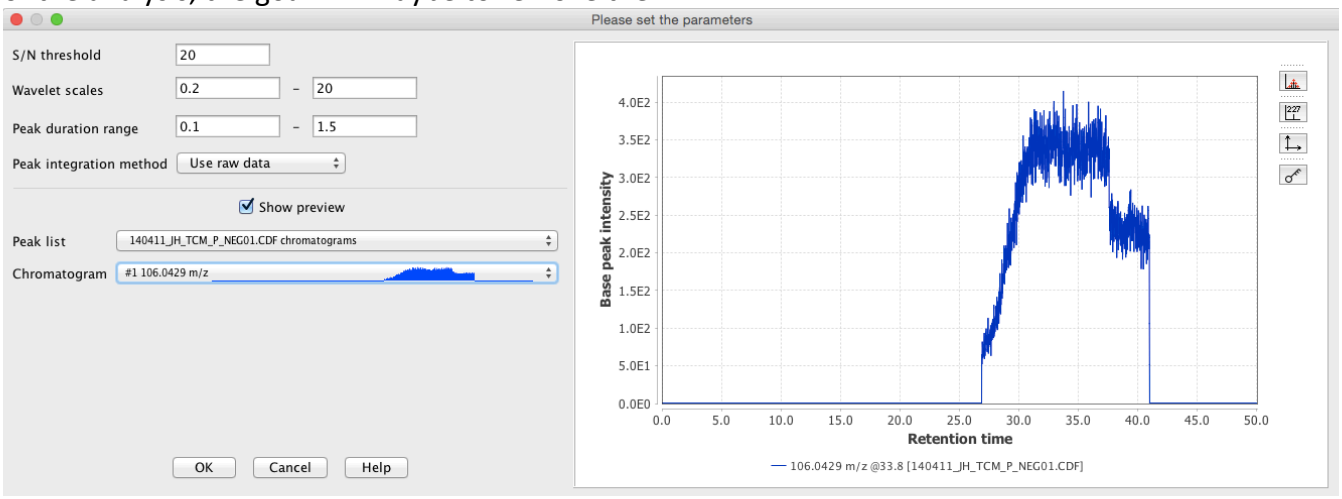
In the case hereunder, with the same parameters as in the first example, we obtain this:



These peaks have all the same masses and are present at the end of the gradient.
If the wavelet scales and the duration range are increased, we can obtain this:

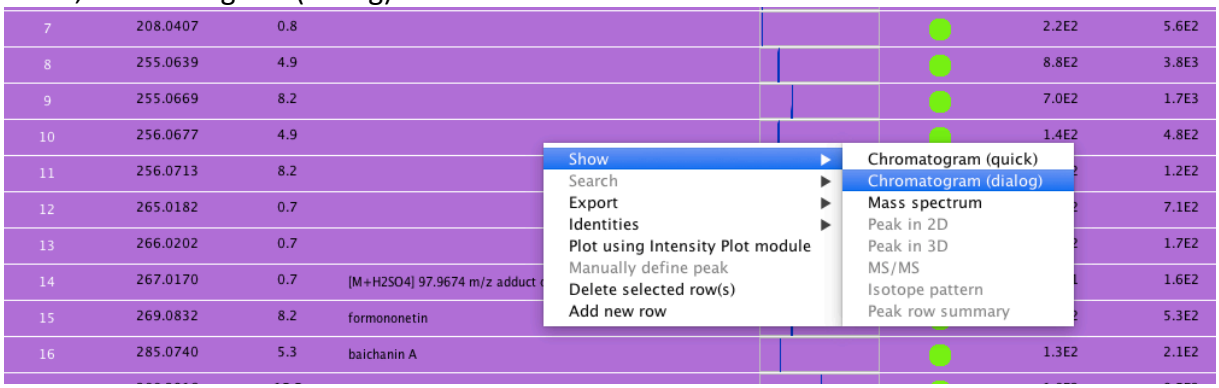


But these peaks are at the end of the gradient and are probably in the blank too. Depending of the analysis, the goal will maybe to remove them:

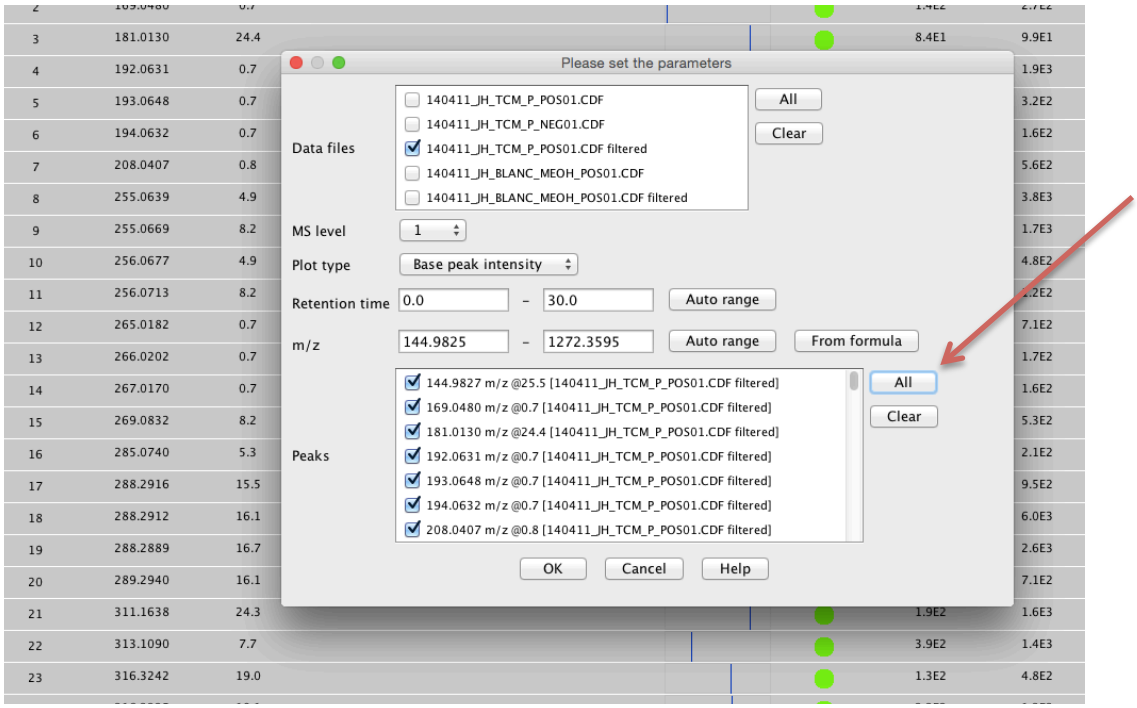


Then, the peaks of interest have to be checked, to see if the new parameters recognise them.

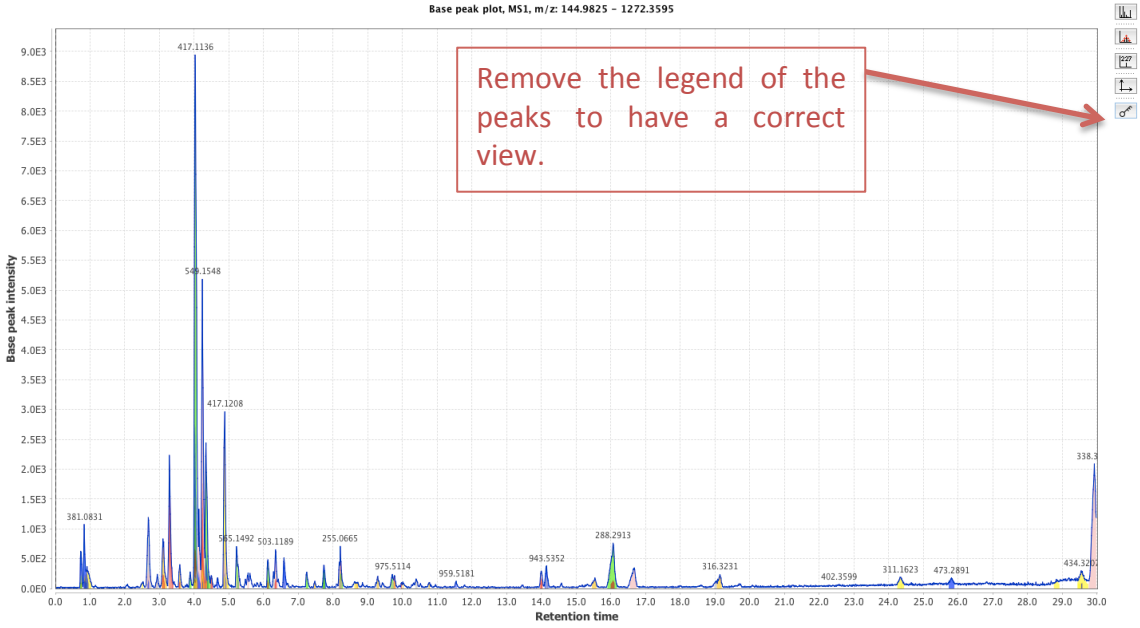
After the construction of the peak list, have a check on the whole deconvolution: open the created peak list by double clicking. Then select all the peaks (ctrl A). Click right and select Show, Chromatogram (dialog):



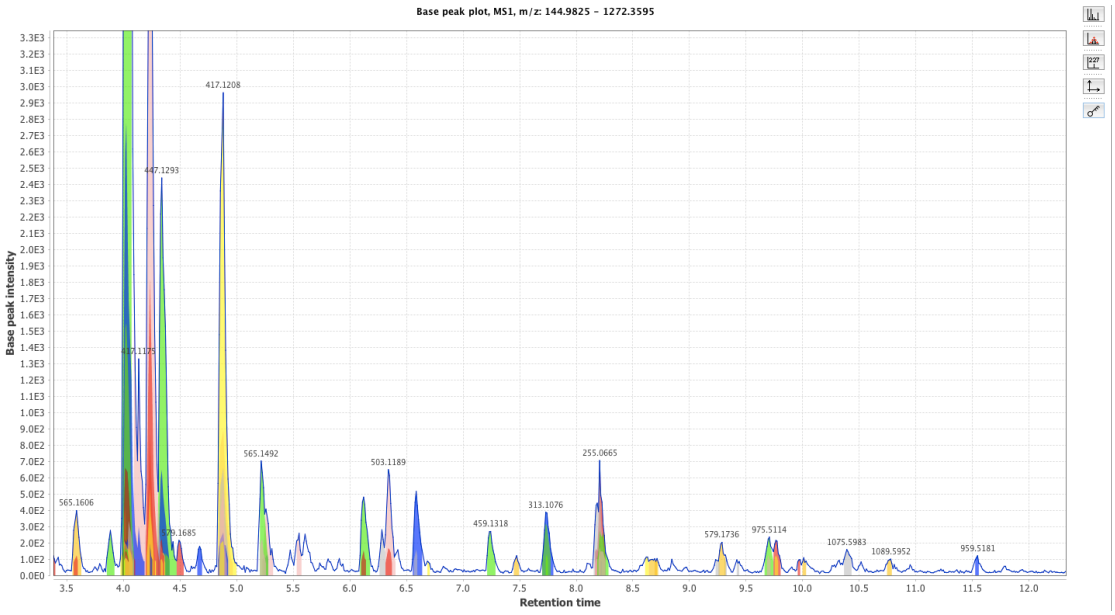
In the dialog boxe, select **ALL** to have all the peaks shown in a chromatogram:



Then, you can observe the coloured peaks in the chromatogram:

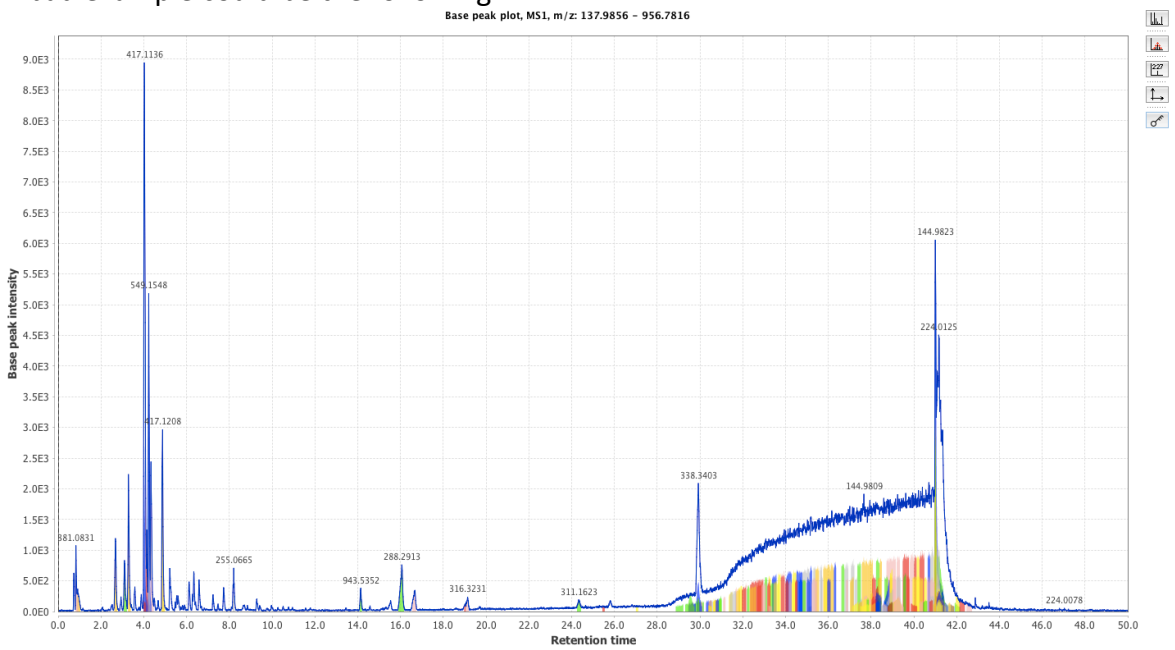


With the Zoom, inspect the chromatogam:

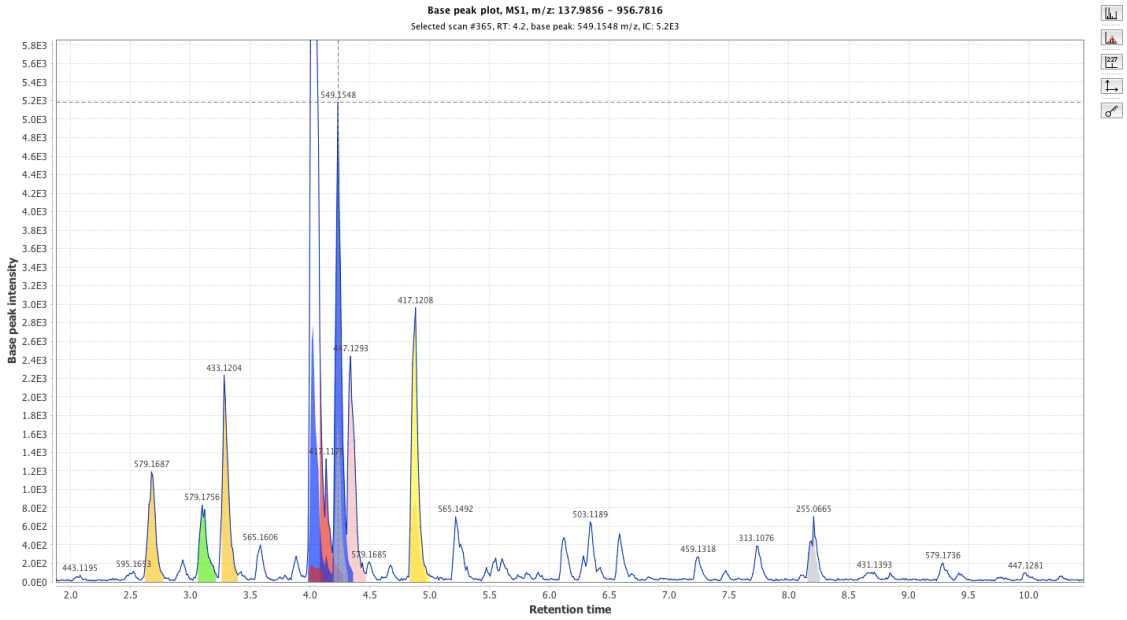


In this example, a few signal around 5.5 min aren't considered as peaks. Setting the limits depends of your analysis and of your goal.

A bad example could be the following:



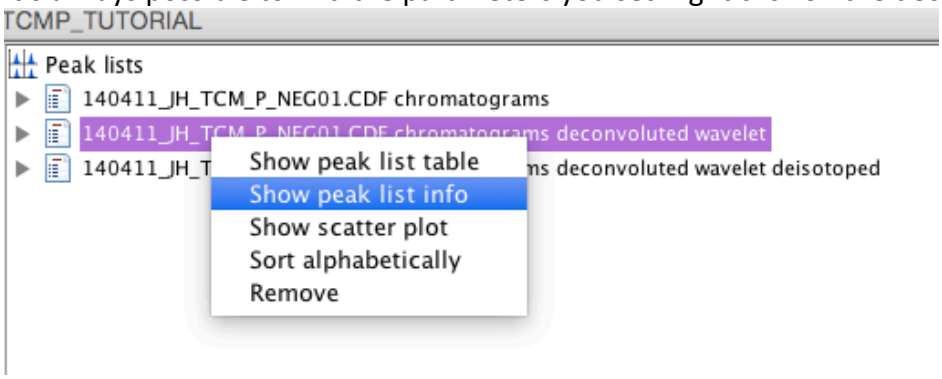
In this case, we observe first that the noise at the end of the gradient is considered as many peaks. You could have removed the end of the analysis by the filtering function (see above). Another problem is visible in this zoom:



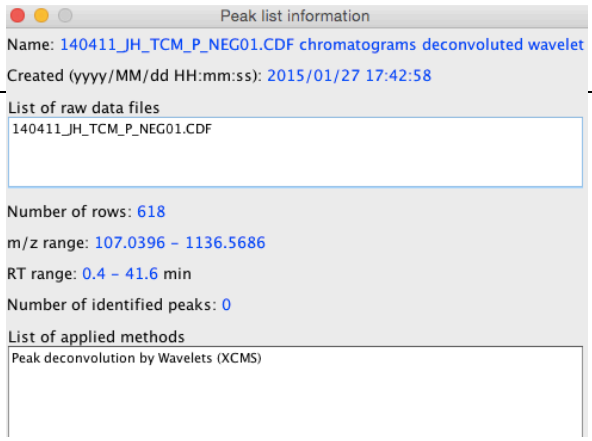
Here, the problem comes from the **Min time span** set in the chromatogram builder which was set at 0.1 min. For UHPLC, it's too long. This comment calls your attention to the fact that the problem can result from a previous step, before the deconvolution algorithm.

The peak picking requires to optimise carefully all the different parameters. We have to keep in mind that the ionisation with ESI depends of the molecule and that the intensity of the signals isn't always correlated with the concentration. It means that a minor signal in MS could be due to a molecule that doesn't ionisate well and that could be in an more important concentration.

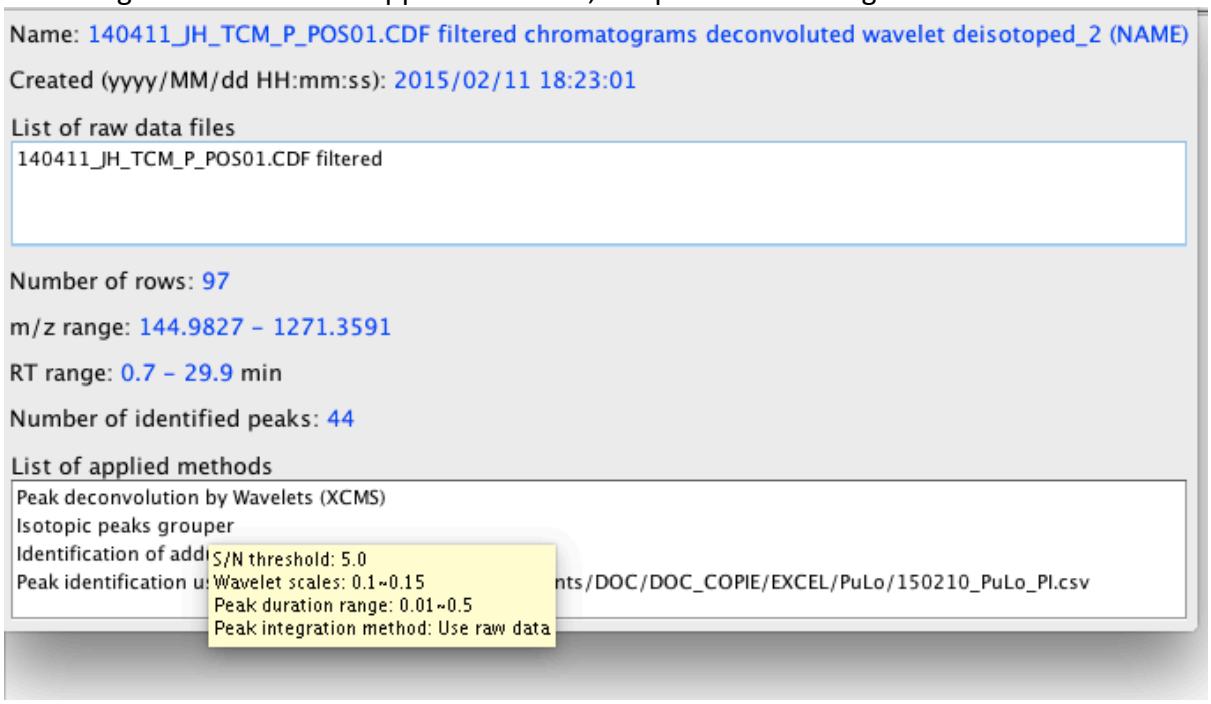
It's always possible to find the parameters you set: right click on the deconvulated peak list:



This window is opened:

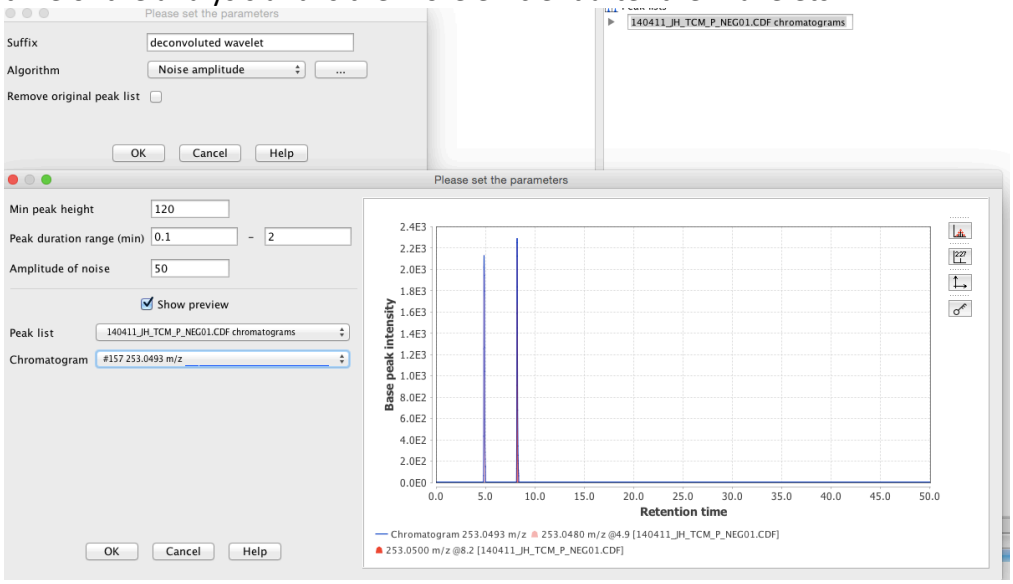


In moving the mouse of the applied methods, the parameters are given:



If the software R doesn't work, MZmine offers the following algorithms: **Baseline Cut-off, Noise Amplitude, Savitsky-Golay** and **Local minimum search**.

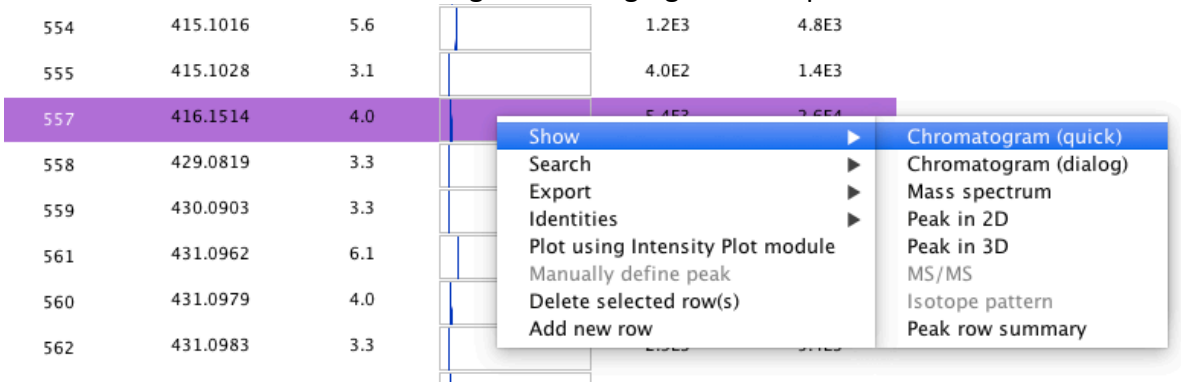
Baseline Cut-off, Local minimum search and Savitsky-Golay aren't adapted with the data from the TOF and Orbitrap. The **Noise amplitude** can be useful if the noise varies during the time of the analysis and is the more efficient after the Wavelets.



Fill in the boxes with the appropriate values, with the same method as described for the wavelets. Peak picking is a compromise and requires a lot of experimentation and patience for optimal results.

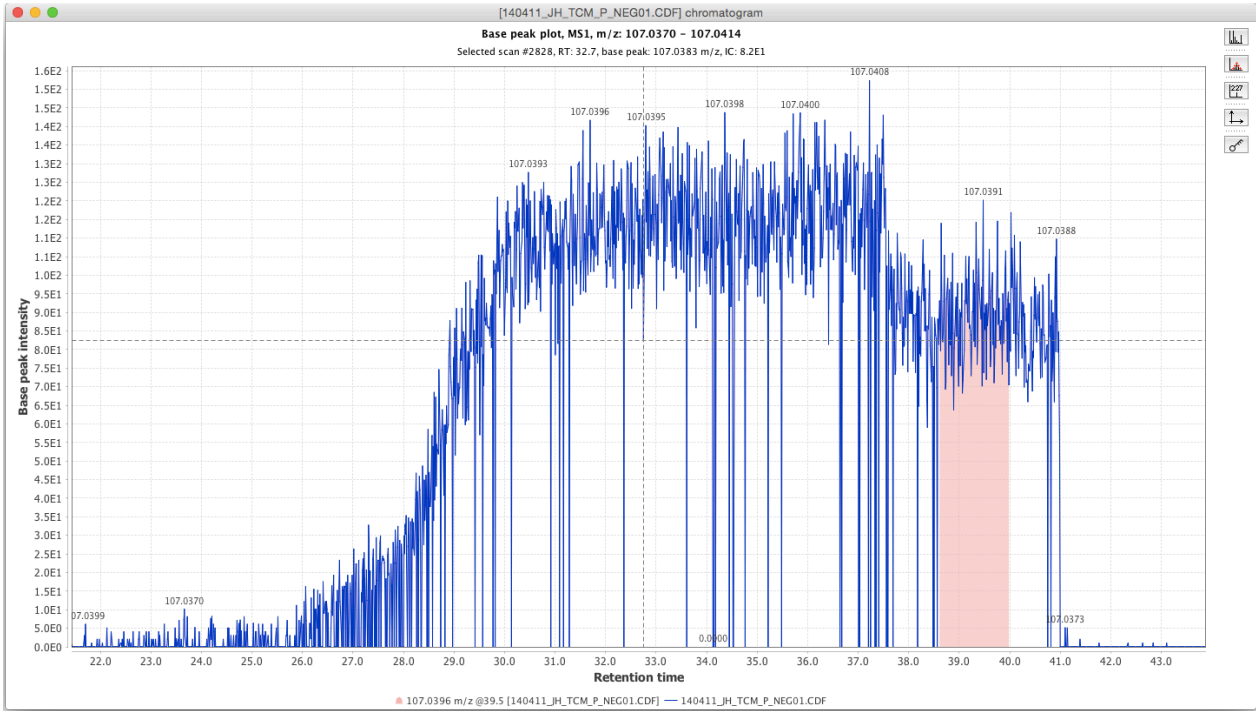
An additional algorithm for peak-picking called Grid-Mass and based on an alternative approache using 2D ms/rt maps is now available in MzMine. The paper describing the peak picking algorithm and how to assess the optimal parameters is available here : http://onlinelibrary.wiley.com/doi/10.1002/jms.3512/abstract. The evaluation of the algorithm was performed on different datasets and in conclusion it seemed to performed equally or better than XCMS wavelets

After the peak deconvolution step, MZmine produces a resolved peak list with one peak per row. We can visualize the chromatogram clicking right on the peak of interest:



Hereunder is an example of peak picking: in blue, the normal chromatogram, before peak picking, in pink the result of peak picking.

Hereunder is an example that shows that the peak picking wasn't really efficient:



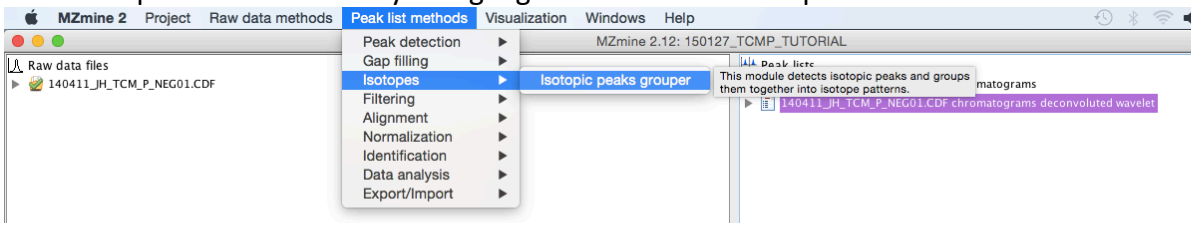
We can observe that all the masses of the massif are the same, but the deconvolution parameters did that only a bit of the massif is considered as a peak.
To resolve this kind of problem, different possibilities are realizable:

- Filter the end of the gradient at the beginning of the process (Raw data/Filter/Data set filtering, see above).
- Compare the peaks lists of the analysis of interest with the peak list of a blank, treated in a similar way (a step of alignment is necessary, see the tutorial). It's always possible to delete a peak. This step of comparison with a blank should be done anyway.

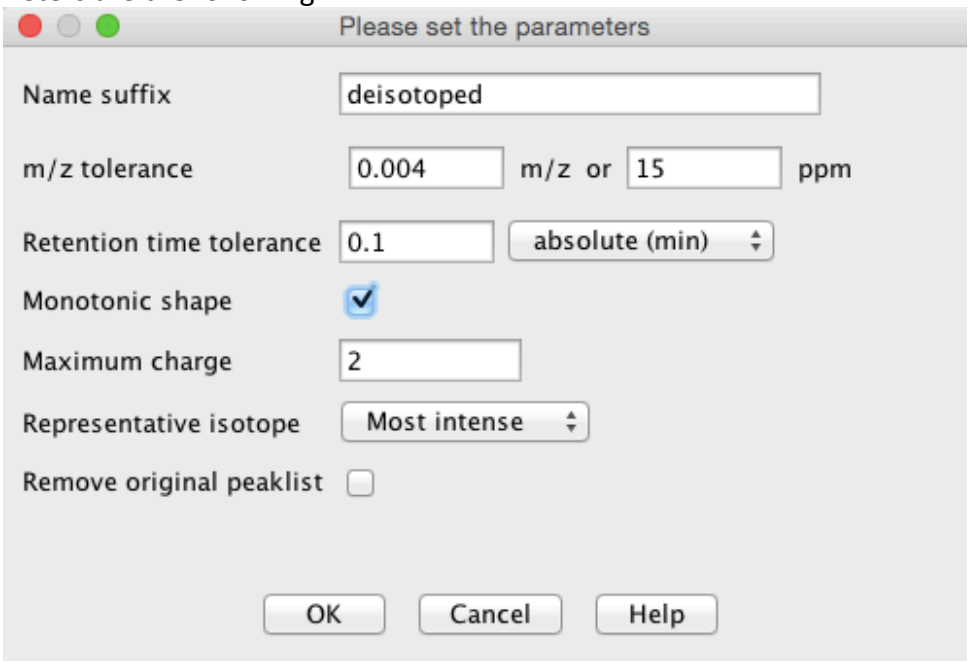
We can also visualize the peaks using the 3D visualizer plot on the raw data, if you have installed the Java 3D module. This is a useful check of the accuracy of peak picking. There is also a 2D "gel view" of the data.

Isotopic peak grouper

The three steps of peak detection results on a peak list of m/z, where the isotopes are separated. An isotopic peaks grouper is necessary to connect the peaks emanating of the same component. Make sure you highlight the deconvoluted peak list.



The parameters are the following:



For the steps of comparisons of peaks lists, refer to the official manual and tutorial of MZmine, in particular for:

- Peak alignment
- Gap filling
- Export
- Batch analysis

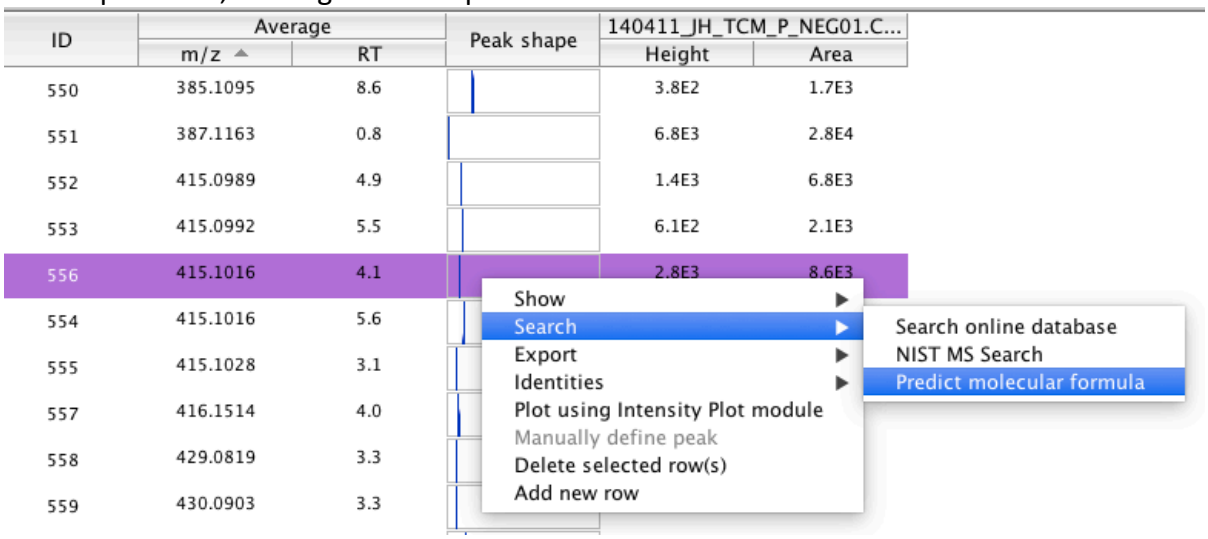
Alignment and Gap-filling steps are particularly important steps in a metabolomics analysis (when you will be comparing various LC-MS profiles one against another).
In the case of MS2 data treatment, Peak Extender module (Peak List methods/Peak detection/Peak extender) will be used to rebuild chromatographic data from MS2 peaklist.

Identification

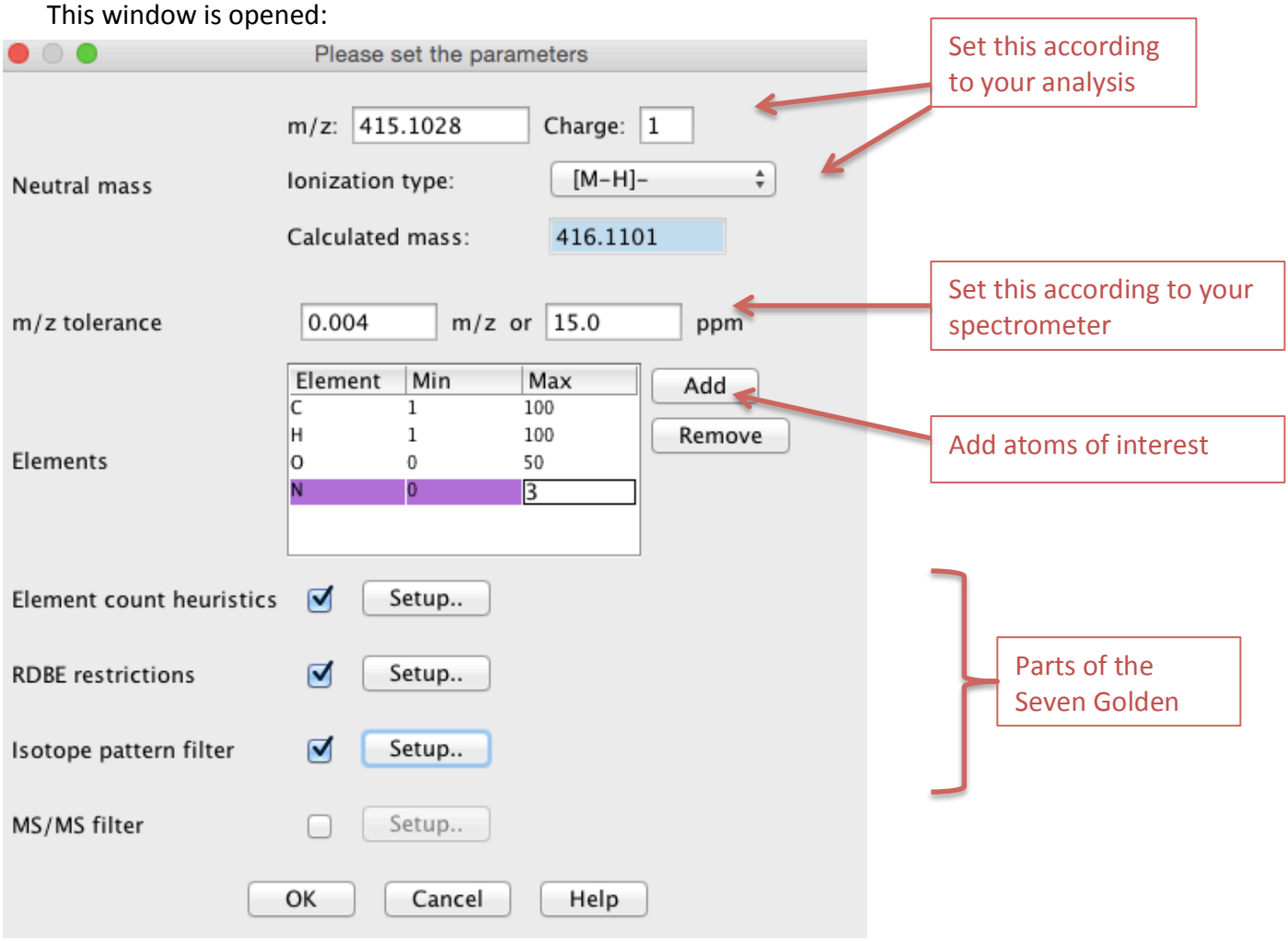
MZmine offers modules for compound identification and molecular formula prediction:
1. Prediction of molecular formula
2. Comparison with database

1. Prediction of molecular formula

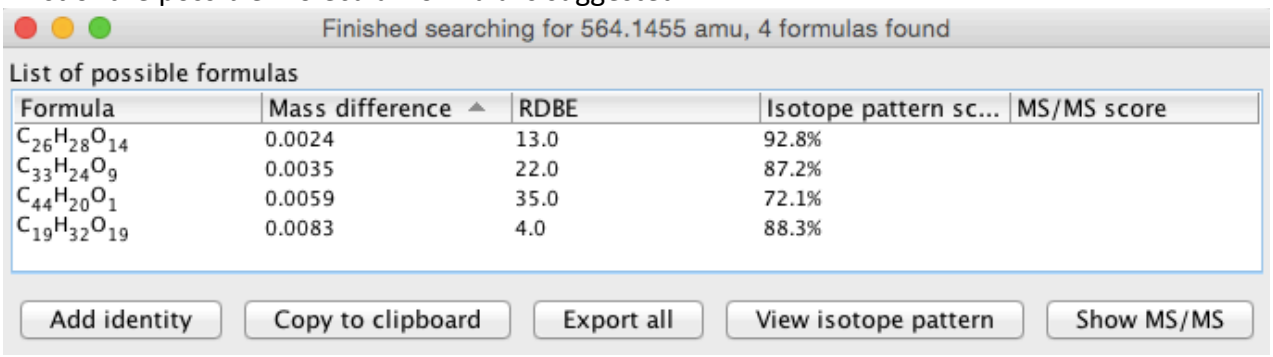
On the peaks list, click right on one peak:



This window is opened:

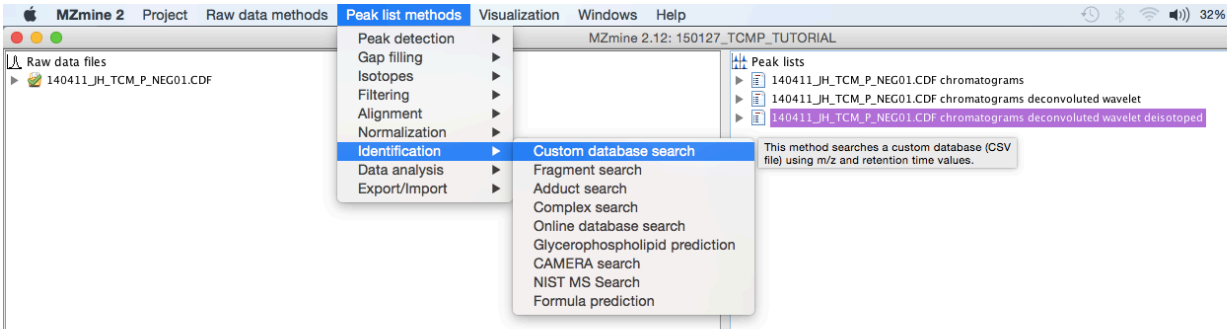


A list of the possible molecular formula is suggested:



2. Identification with database

MZmine offers several modules of identification:

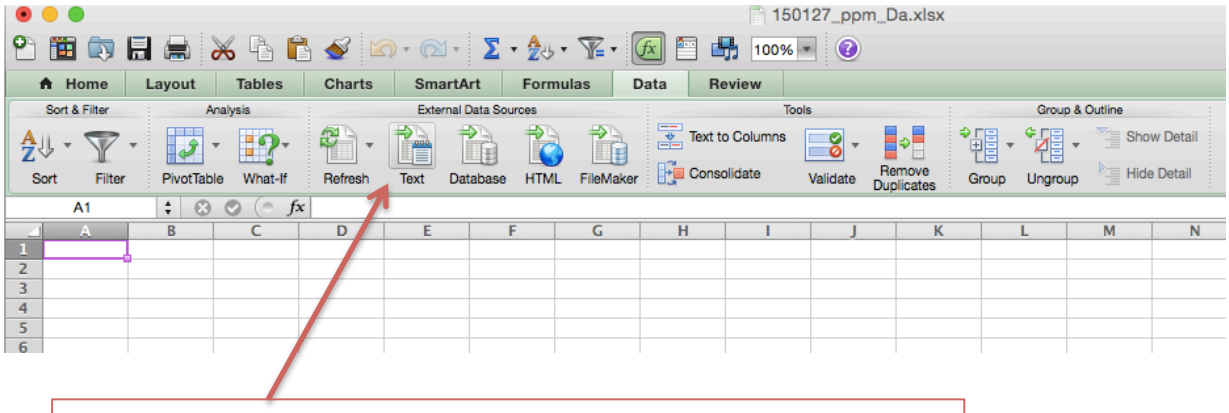


The more useful are **Custom database search** and **Adduct search**. To search with the **Online database search** is an option, but is should be done with caution on the entire peak list. As it is said in the MZmine tutorial, this method often returns many compounds, drugs and pharmaceuticals which are irrelevant to a plant based studies and it is time-consuming. For this reason it is recommended to search in online database only individual peaks using the peak list, in a similar way than with the module of prediction of molecular formula.

Custom database search

From the Dictionary of Natural Product (DNP), it is possible to export the hits for a given species or genus or family into a .CSV file (Comma Separated Values).
For the module of identification of MZmine, we need from the DNP the exact mass, the molecular formula, the chemical name of the compound and an identification code. We recommend using the CRC code that is the only code that the DNP has for all the molecules. The CAS number isn't always given in the DNP.

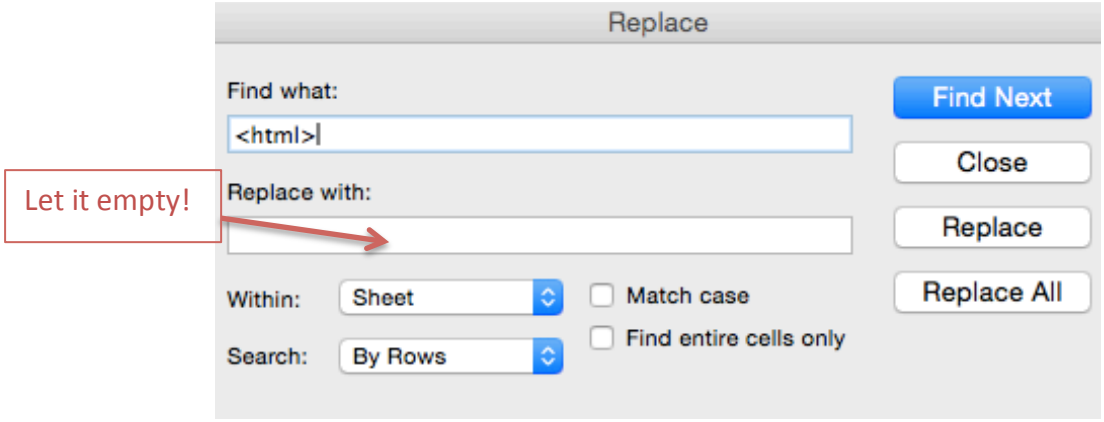
The DNP exports it with a lot of unnecessary signs of layout, so we need to adapt it on Excel. The first step is to convert the .CSV file into a normal excel file. One of the possibilities to do it quickly is to use the **Data** menu on Excel:



The .CSV file is converted into this kind of nice file:

	A	B
1	Chemical Name	Molecular Formula
2	<html>Arnebiabinone</html>	<html>C₃₀H₂₆O₈</htm
3	<html>Arnebifuranone</html>	<html>C₁₈H₂₀O₅</htm
4	<html>Arnebinol</html>	<html>C₁₆H₂₀O₂</htm
5	<html>Arnebinone</html>	<html>C₁₈H₂₂O₄</htm
6	<html>Deacetylshikonofuran A; 11-O-(2-Methylbutanoyl)</html>	<html>C₂₁H₂₆O₅</htm
7	<html>Deacetylshikonofuran A; 11-O-(3-Methylbutanoyl)</html>	<html>C₂₁H₂₆O₅</htm
8	<html>5,8-Dihydroxy-2-(1-hydroxy-4-methyl-3-pentenyl)-1,4-naphthalenedione; (R)-form</	<html>C₁₆H₁₆O₅</htm
9	<html>5,8-Dihydroxy-2-(1-hydroxy-4-methyl-3-pentenyl)-1,4-naphthalenedione; (R)-form, <	<html>C₁₈H₁₈O₆</htm
10	<html>5,8-Dihydroxy-2-(1-hydroxy-4-methyl-3-pentenyl)-1,4-naphthalenedione; (R)-form, <	<html>C₁₉H₂₀O₇</htm
11	<html>5,8-Dihydroxy-2-(1-hydroxy-4-methyl-3-pentenyl)-1,4-naphthalenedione; (S)-form</	<html>C₁₆H₁₆O₅</htm
12	<html>5,8-Dihydroxy-2-(1-hydroxy-4-methyl-3-pentenyl)-1,4-naphthalenedione; (S)-form, <	<html>C₁₈H₁₈O₆</htm
13	<html>5,8-Dihydroxy-2-(1-hydroxy-4-methyl-3-pentenyl)-1,4-naphthalenedione; (S)-form, <	<html>C₂₁H₂₂O₆</htm
14	<html>5,8-Dihydroxy-2-(1-hydroxy-4-methyl-3-pentenyl)-1,4-naphthalenedione; (S)-form, <	<html>C₂₃H₂₆O₆</htm
15	<html>5,8-Dihydroxy-2-(1-hydroxy-4-methyl-3-pentenyl)-1,4-naphthalenedione; (S)-form, <	<html>C₂₃H₂₆O₈</htm
16	<html>5,8-Dihydroxy-2-(1-hydroxy-4-methyl-3-pentenyl)-1,4-naphthalenedione; (S)-form, <	<html>C₂₁H₂₄O₆</htm
17	<html>5,8-Dihydroxy-2-(1-hydroxy-4-methyl-3-pentenyl)-1,4-naphthalenedione; (S)-form, <	<html>C₂₀H₂₂O₆</htm

Use the module Replace to remove all the layout signs:

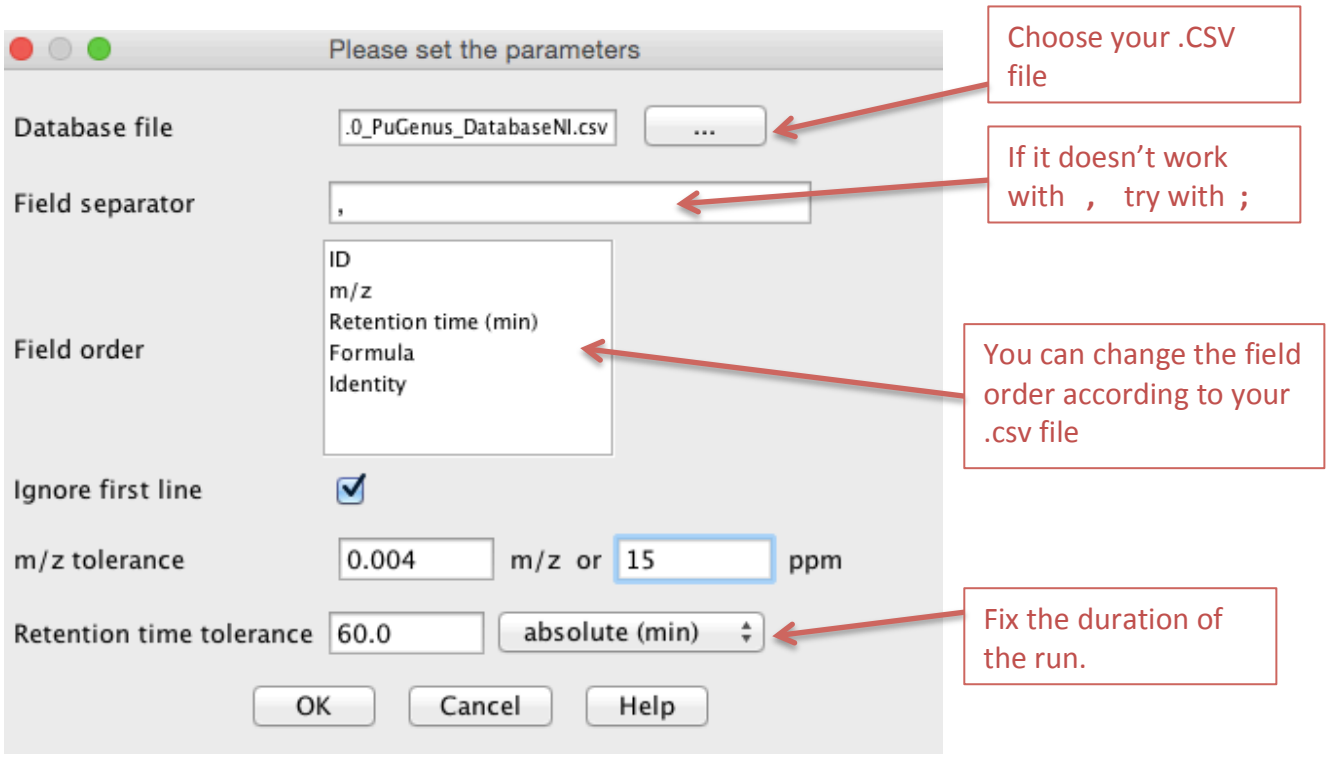


To adapt the excel file to the MZMine module of identification, do these columns (one sheet for each mode):

1. ID: CRC code or what you want
2. m/z: add or remove the exact mass of the proton (1.007825) for negative and positive analysis
3. Molecular formula
4. Chemical name
5. Retention time: fix it at 0 as we usually don't know the retention time.

The last step on Excel is to save it into a CSV file.

The DNP is one source of information. Be careful that the major compound of a species is maybe only given into the genus of the plant. Depending on the plant, to compile the molecules that are given in the articles can be useful.
On MZmine, click on **Custom database search** as shown previously. This window is opened:

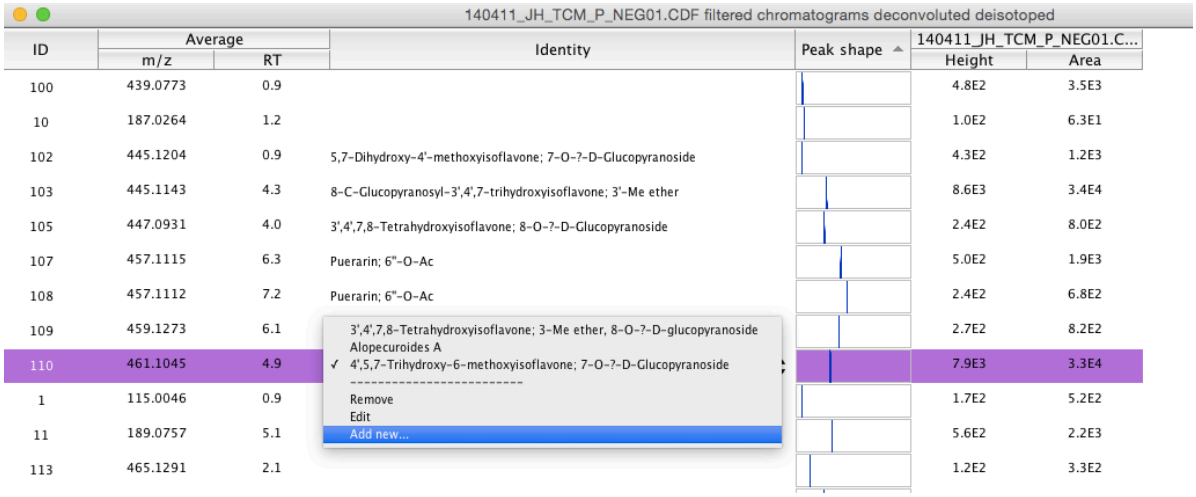


The results of the search are written in the peak lists on the identification column.

140411_JH_TCM_P_NEG01.CDF filtered chromatograms deconvoluted deisotoped

ID	Average m/z	Average RT	Identity	Peak shape	140411_JH_TCM_P_NEG01.C... Height	140411_JH_TCM_P_NEG01.C... Area
100	439.0773	0.9			4.8E2	3.5E3
10	187.0264	1.2			1.0E2	6.3E1
102	445.1204	0.9	5,7-Dihydroxy-4'-methoxyisoflavone; 7-O-?-D-Glucopyranoside		4.3E2	1.2E3
103	445.1143	4.3	8-C-Glucopyranosyl-3',4',7-trihydroxyisoflavone; 3'-Me ether		8.6E3	3.4E4
105	447.0931	4.0	3',4',7,8-Tetrahydroxyisoflavone; 8-O-?-D-Glucopyranoside		2.4E2	8.0E2
107	457.1115	6.3	Puerarin; 6"-O-Ac		5.0E2	1.9E3
108	457.1112	7.2	Puerarin; 6"-O-Ac		2.4E2	6.8E2
109	459.1273	6.1	Puerol A; 2"-O-?-D-Glucopyranoside		2.7E2	8.2E2
110	461.1045	4.9	4',5,7-Trihydroxy-6-methoxyisoflavone; 7-O-?-D-Glucopyranoside		7.9E3	3.3E4
1	115.0046	0.9			1.7E2	5.2E2

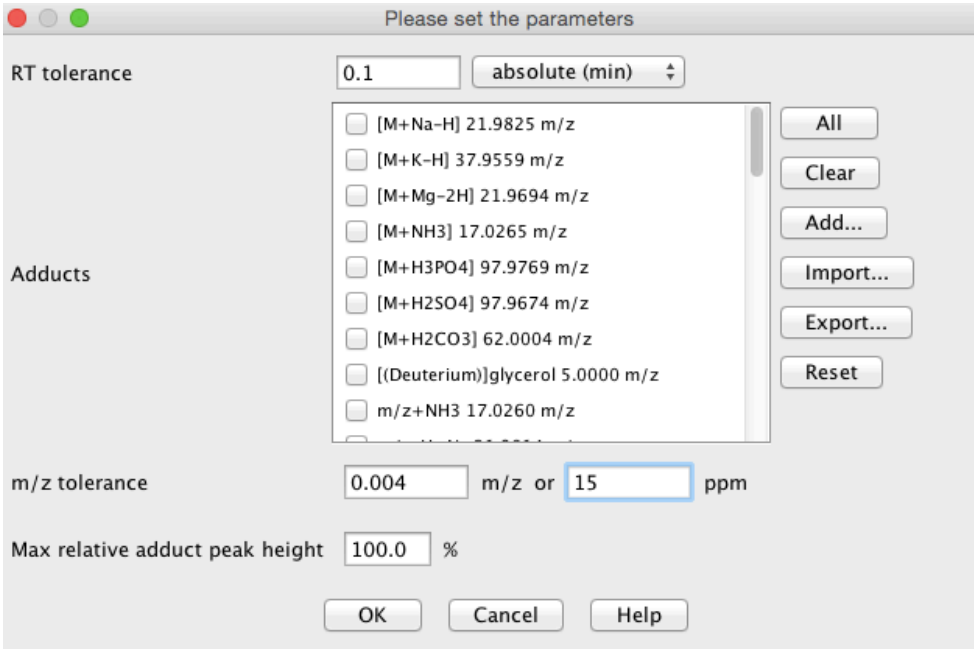
Check on the found identity of each peak if several possibility are given, as a m/z can coincide with different compounds:



You can edit each possibility to find the information contained in the .csv file.
A crosscheck of the hit with the **Predict molecular formula** is recommended.

3. Adduct search

On the menu of identification, click on Adduct search:



By default, MZmine doesn't have a lot of adducts. As for the custom database search, it is possible to import from a .CSV file a list of adducts and then to choose them according of the mode (positive or negative) of your analysis. Be careful that MZmine searches for adduct of the m/z and not of the entire formula, as it is usually given in the article.
To have a list of adducts, see this website where an excel sheet contains the main observed adduct:
http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/MS-Adduct-Calculator/

Don't forget that MZmine compares the mass difference between m/z feature and not from the entire formula.