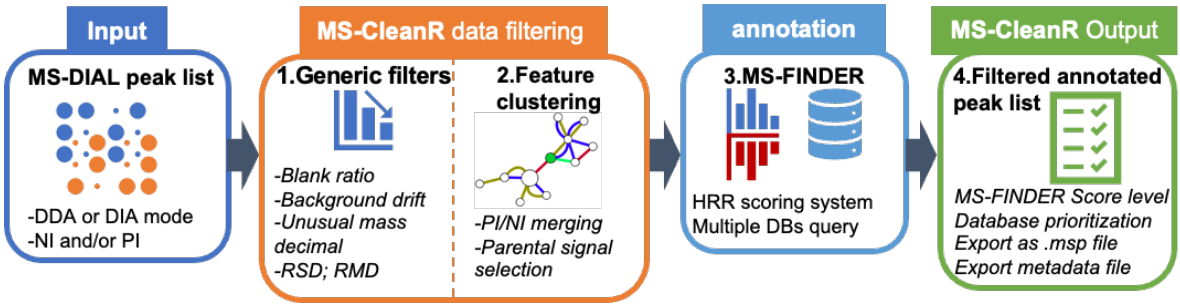
guillaume.marti@univ-tlse3.fr @GMarti_AgromiX

An integrated metabolic workflow

MS-DIAL/CleanR/FINDER Tutorial

Untargeted metabolomics using liquid chromatography-mass spectrometry (LC-MS) is currently the gold-standard technique to assess the chemical complexity of biological samples. However, this approach still has many limitations; notably, the difficulty of accurately estimating the number of unique metabolites profiled among the thousands of MS ion signals arising from chromatograms. Here, we describe the MS-CleanR workflow, based on the MS-DIAL/MS-FINDER suite, which tackles feature degeneracy and improves annotation rates.

This tutorial aims at guiding the user to process complex LC-MS data through the use of MS-DIAL for spectral deconvolution and data alignment, MS-CleanR for feature filtration and MS-Finder for peak annotation.

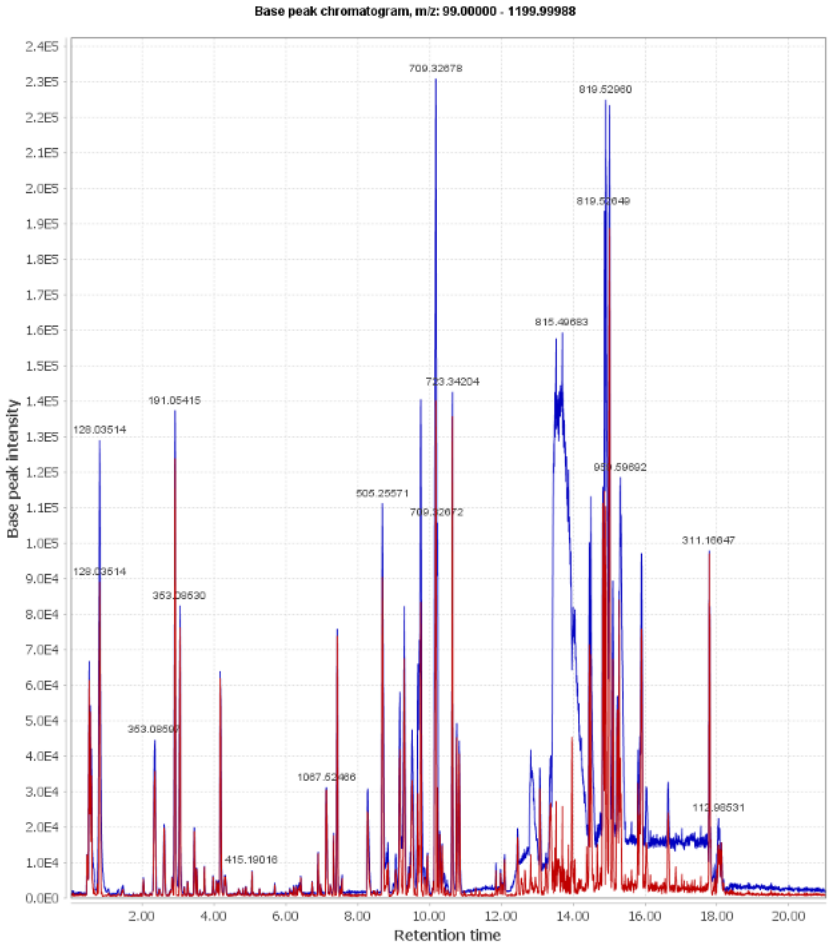


LC-MS metabolomic pipeline

Preliminary steps: inspect your data !!!

A thorough chromatogram inspection is mandatory before any data processing to, for instance:

- Evaluate noise threshold level.
- Superpose samples like QCs, blanks or biological replicates to assess any RT or m/z drifts or remove weird acquisitions.
- Define informative rich RT range.



LC-MS chromatogram superposition using MZ-Mine

1. MS-Dial data processing

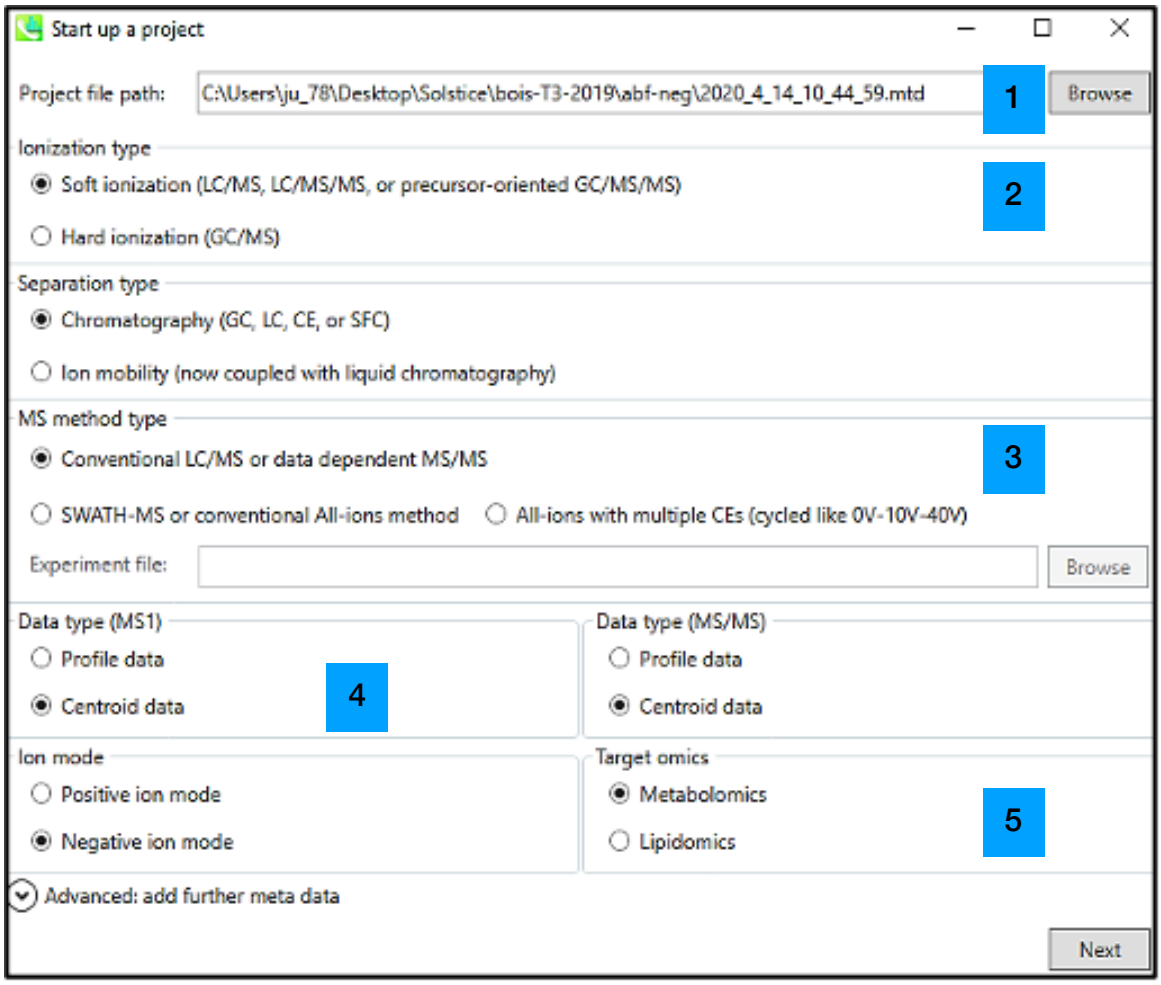
MS-Dial is a highly a versatile LC/GC-MS data processing software. Since version 4.60, the MS raw data of SCIEX (.wiff, .wiff2), Thermo (.raw). Agilent (.d), and Waters (.raw) can be imported directly without any data conversion. Otherwise, use Abf converter.

Software download:

http://prime.psc.riken.jp/compms/msdial/main.html
http://prime.psc.riken.jp/compms/msfinder/main.html
https://mzmine.github.io/download.html
https://www.reifycs.com/AbfConverter/

Mandatory: The regional parameters of the windows computer has to be set in US or UK (decimal separator as « . »).

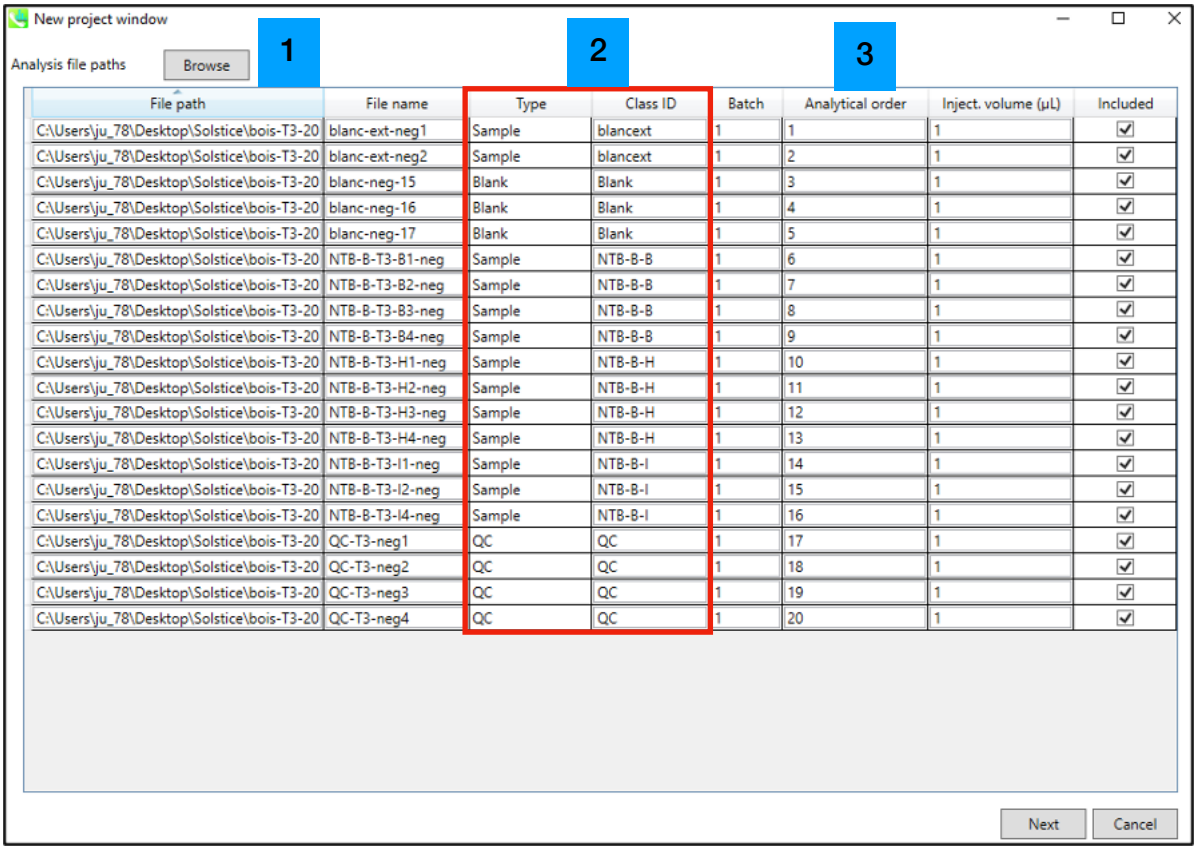
Launch MS-DIAL and select « File—>new project »



1. *Select the directory containing raw data or converted .abf files. Note that pos or neg mode must be processed separately*
2. *Soft ionization is for ESI/APCI mode, hard ionization for EI*
3. *Conventional is for DDA mode, SWATH-MS is for DIA mode (including AIF, MSE…). In this case, a tab delimited « Experiment file » with MS1 scan range and precursor window must be called (See compMS website for example file)*
4. *Select data type and ionization mode depending of your acquisition parameters*
5. *« Metabolomics » is to use an in house compound DB, « Lipidomics » leverage on internal MS-Dial lipidBlast DB. Select « Metabolomics » if no DB is used during this process.*

The following window appears and click on "**Browse**" to import the .abf or raw files you want to process. During data importation, it is important to note the type (Blank, QC or Sample) and class of every sample in **Class ID column** (blank, sample class, QC).

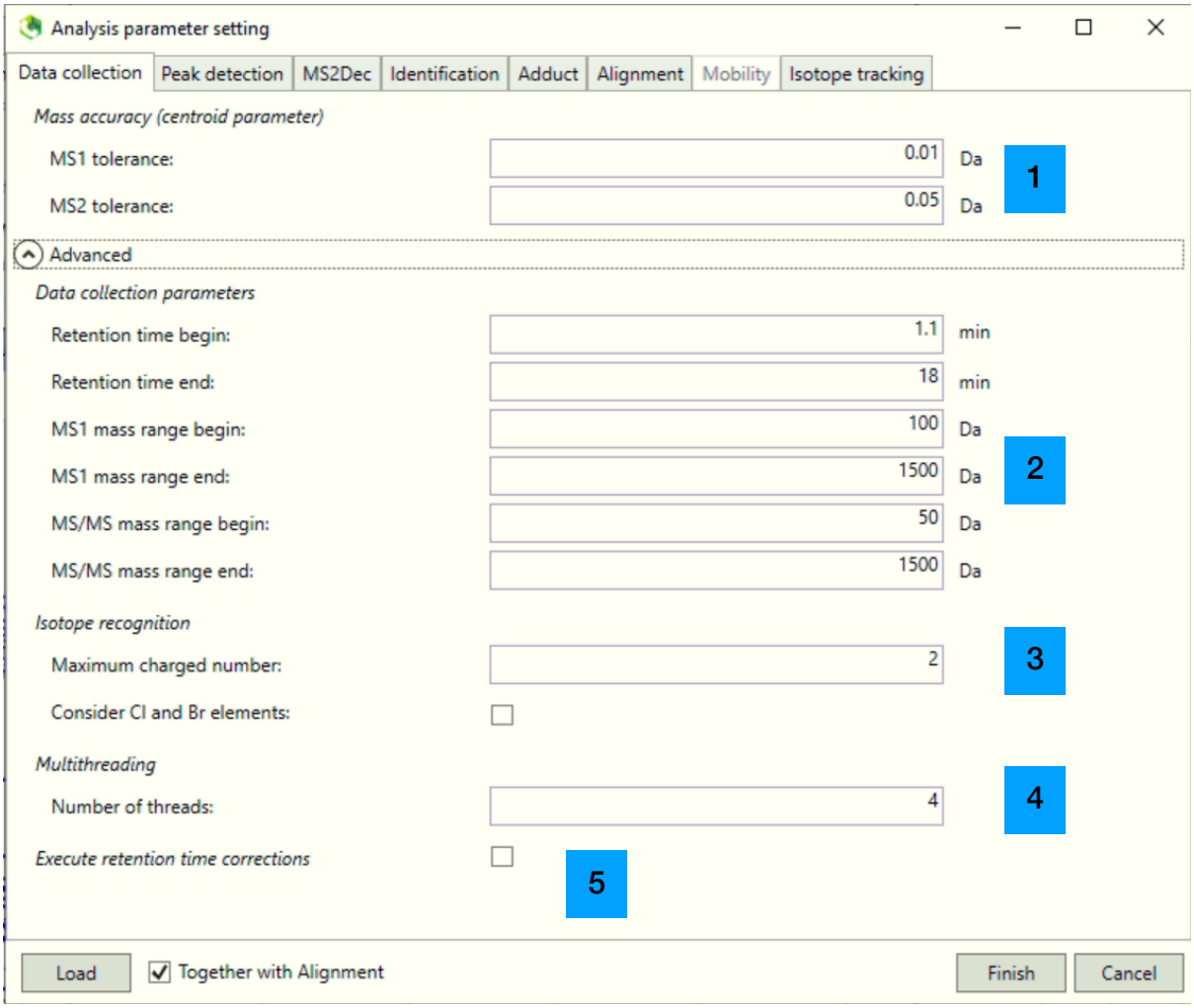
Optionally: If both ionization modes have been acquired, be careful to have the **same number of samples** between pos and neg modes and in the **same order.** Note that no less than three samples per class are needed to use Pearson correlation clustering within MS-CleanR.



File path	File name	Type	Class ID	Batch	Analytical order	Inject. volume (µL)	Included
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	blanc-ext-neg1	Sample	blancext	1	1	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	blanc-ext-neg2	Sample	blancext	1	2	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	blanc-neg-15	Blank	Blank	1	3	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	blanc-neg-16	Blank	Blank	1	4	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	blanc-neg-17	Blank	Blank	1	5	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	NTB-B-T3-B1-neg	Sample	NTB-B-B	1	6	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	NTB-B-T3-B2-neg	Sample	NTB-B-B	1	7	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	NTB-B-T3-B3-neg	Sample	NTB-B-B	1	8	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	NTB-B-T3-B4-neg	Sample	NTB-B-B	1	9	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	NTB-B-T3-H1-neg	Sample	NTB-B-H	1	10	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	NTB-B-T3-H2-neg	Sample	NTB-B-H	1	11	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	NTB-B-T3-H3-neg	Sample	NTB-B-H	1	12	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	NTB-B-T3-H4-neg	Sample	NTB-B-H	1	13	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	NTB-B-T3-I1-neg	Sample	NTB-B-I	1	14	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	NTB-B-T3-I2-neg	Sample	NTB-B-I	1	15	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	NTB-B-T3-I4-neg	Sample	NTB-B-I	1	16	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	QC-T3-neg1	QC	QC	1	17	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	QC-T3-neg2	QC	QC	1	18	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	QC-T3-neg3	QC	QC	1	19	1	☑
C:\Users\ju_78\Desktop\Solstice\bois-T3-20	QC-T3-neg4	QC	QC	1	20	1	☑

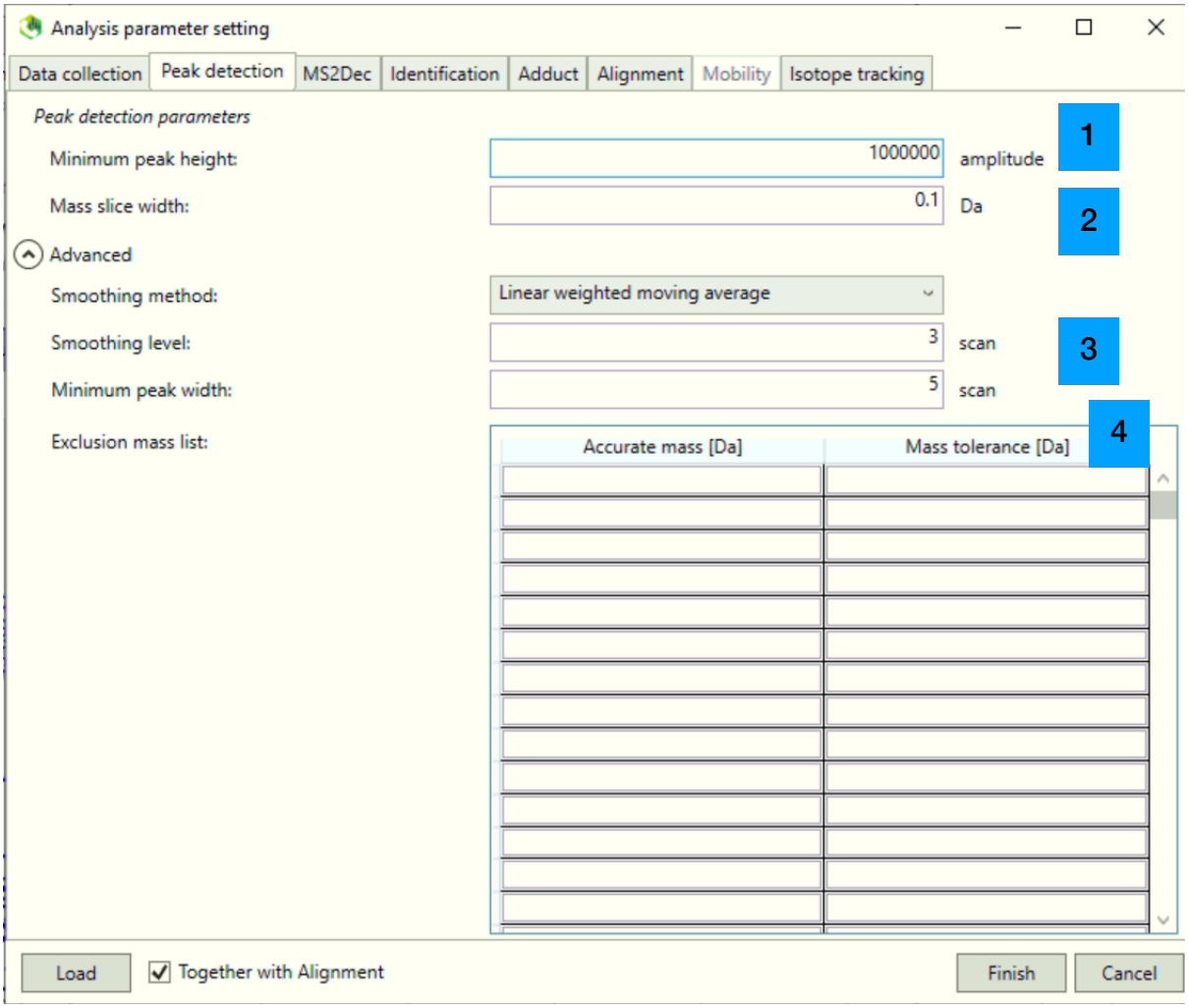
1. *Select the raw or .abf files to be processed. We advise separating pos/neg files in two distinct directories since the data processing has to be done separately for both ionization mode.*
2. *Select sample type and input Class ID for each acquisition. Avoid « space » in class name, or name with single number or letter.*
3. ***Optionally**: input sample analytical order and batch number to perform LOESS based normalization.*

Next: The « data collection tab » provides general settings for data acquisition. A preliminary careful inspection of chromatograms is mandatory to set correctly all parameters.



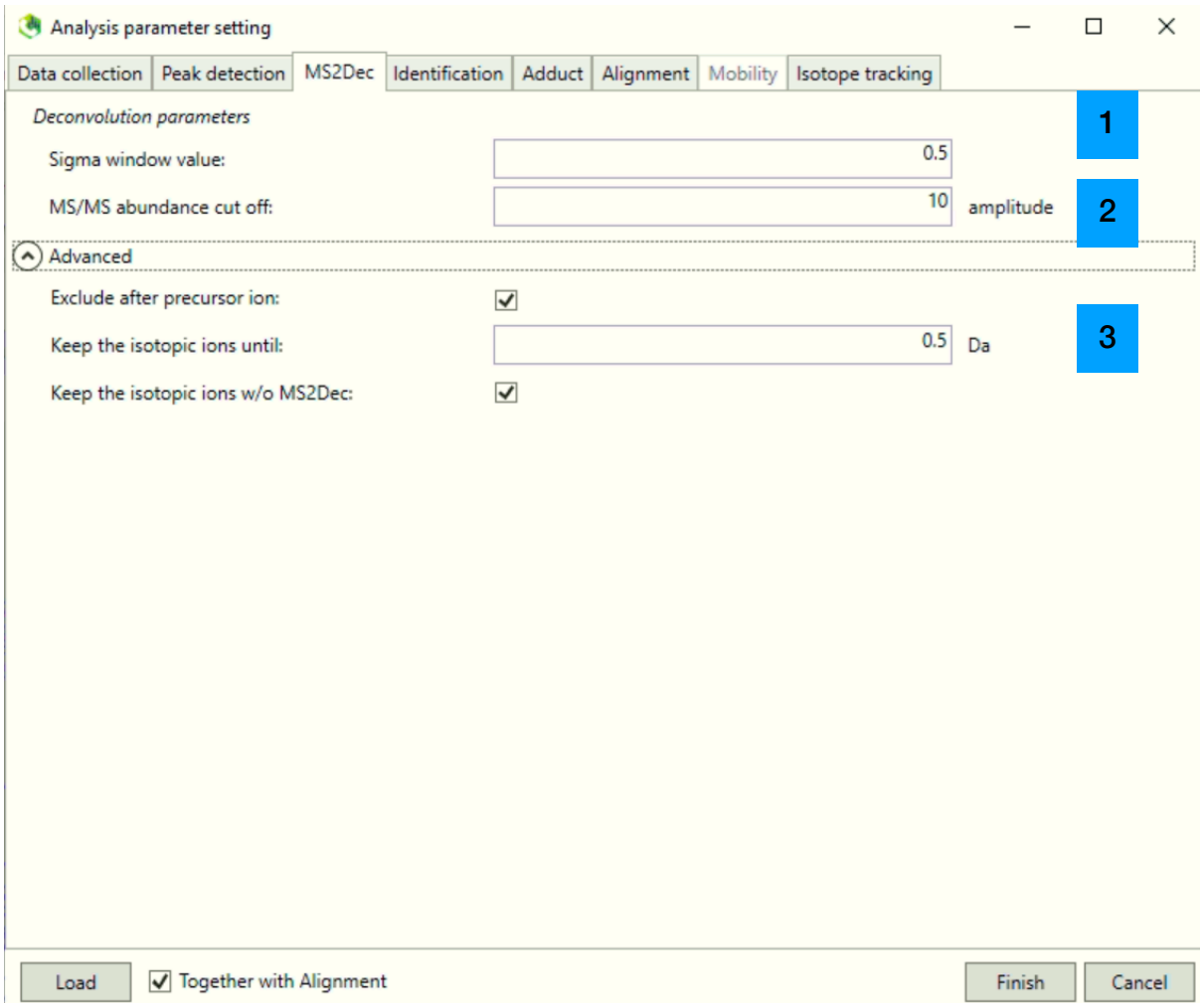
1. *Select MS tolerance according to your instrument. For HRMS, typically MS1 = 0.01 Da and MS2 = 0.05 Da.*
2. *Select RT and m/z range parameters according to the chromatogram inspection of preliminary step and acquisition parameters.*
3. *Maximum « z » number, typically 2 for natural products, more for polypeptides or related compounds. Cl and Br elements can be ticked for marine natural products for instance.*
4. *Number of threads can be set-up to operate the capacity of processor.*
5. ***Optionally**: If « retention time correction » is selected, it will be executed after data processing. The injection of a standard mixture is mandatory in this case. This mixture has to be defined at the beginning of the process (Expected RT, m/z, compound name, RT and m/z windows). Refer to CompMS website for more details.*

Next: The « peak detection tab » provides several parameters which greatly influence final results. The most important one being the « minimum peak height ». The preliminary inspection of chromatograms can help to set it correctly. A good starting point is to tune it at 50% of the Base Peak Intensity chromatogram of a blank sample. Typical value are in the 5e5 to 5e6 range for Orbitrap, and 1e3 to 1e4 range for QTOF mass spectrometers.Set amplitude according to the



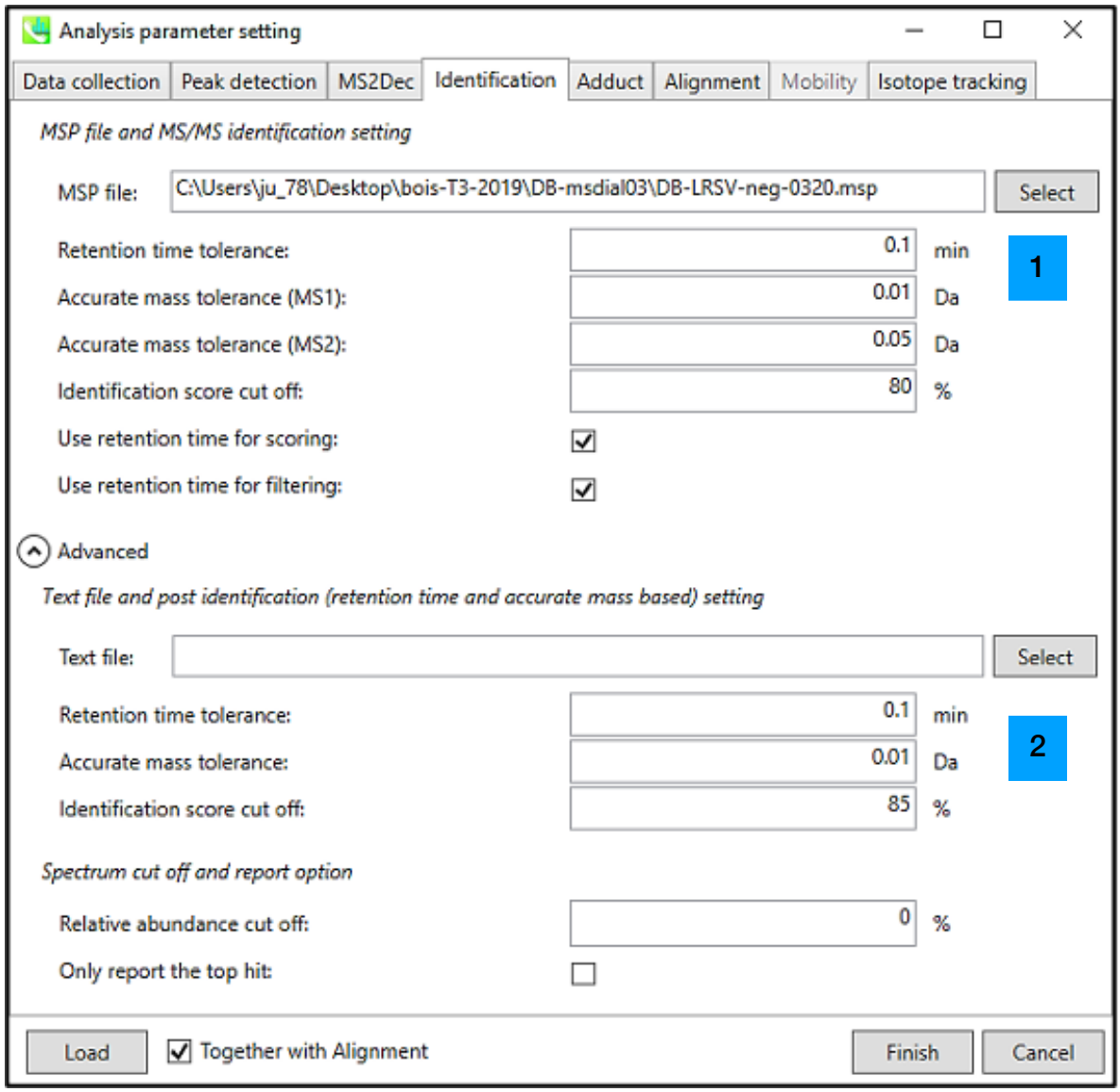
1. *inspection of your baseline. This parameter may increase processing time if tuned very low.*
2. *Mass slice width can be kept as such. Tune it if you observe duplicate peaks after data processing.*
3. *Default parameters are correct for UHPLC with <2 µm particles. Again, inspection of datapoints for tiny peaks are a good prerequisite.*
4. ***Optionally**: A mass list of unwanted m/z can be set here. Each mass will be discarded from the data processing.*

Next: The MS2 Deconvolution tab provide some parameters for MS2 detection.



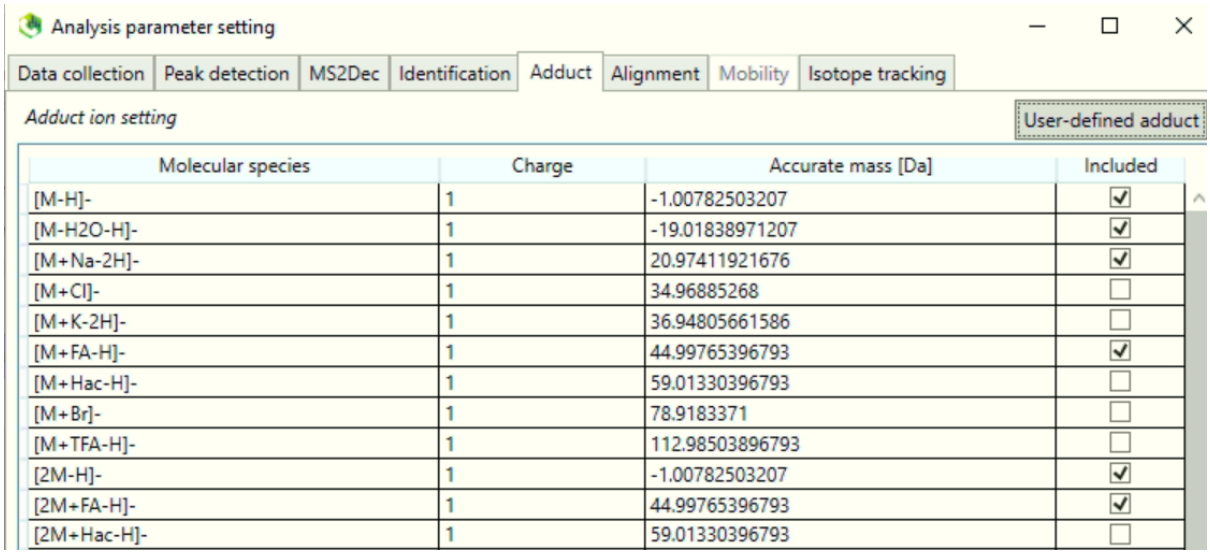
1. *The sigma window value can be kept at 0.5; under 0.3, it will capture more noise, upper than 0.7, it will reduce peak resolution.*
2. *The MS/MS abundance cut-off as to be set according to the spectral chromatogram inspection.*
3. *These parameters are optional. Tick « exclude after precursor ions » for LCMS based lipidomics and metabolomics.*

Next: the « Identification tab » is optional since MS-FINDER provides more features for peak annotation process. However, its interesting to import an in-house database if available, or a set of m/z of interest if any. Refer to CompMS website for example DB files.



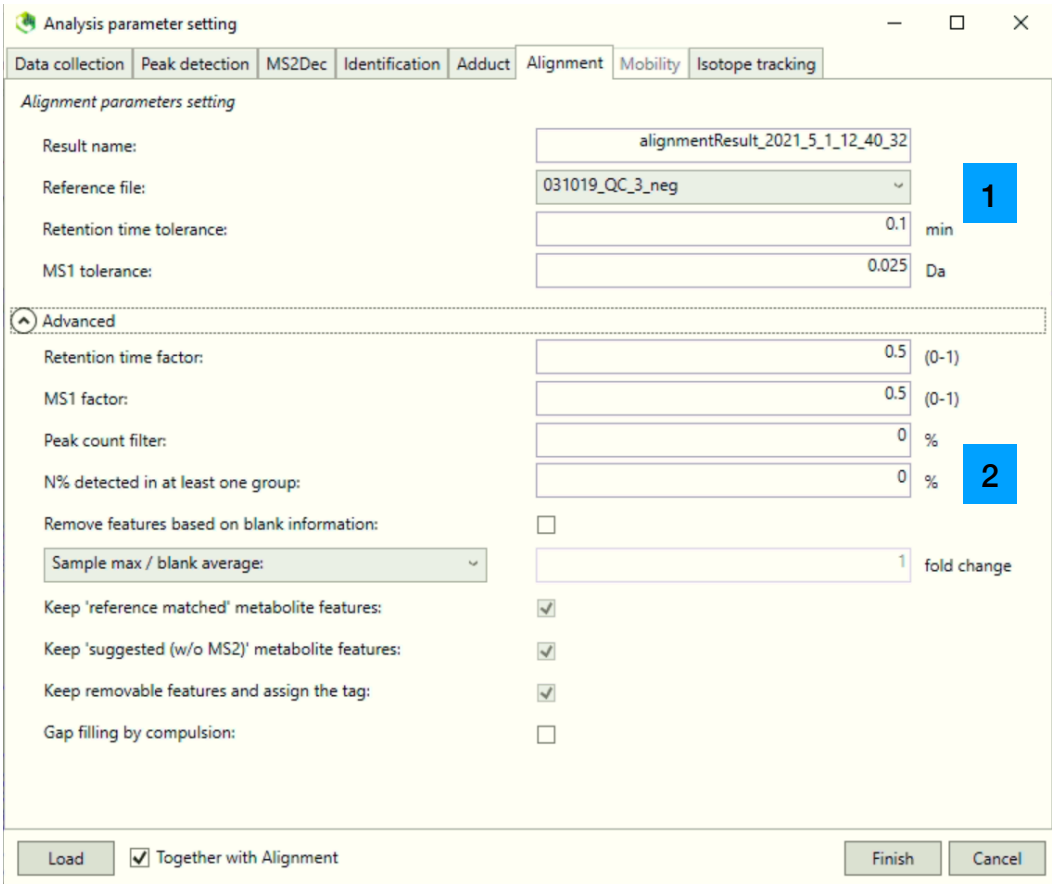
1. ***Optionally***, *a *.msp formatted database may be imported. In this case, annotation are based on MS1, MS2 dot product recognition and RT windows (annotation level 1). Suitable for « in-house » database.*
2. ***Optionally***, *a list of mass, RT and metabolite may be imported. The annotation are based on accurate mass and RT only.*

Next: The « adduct tab » provide a list of adducts typically encountered in neg or pos mode, covering most of LC solvents or buffer used in metabolomics. Other adducts may be added to this list using the « user defined adduct » button. A comprehensive list is not mandatory at this step, since MS-CleanR also has an adduct detection algorithm.



Molecular species	Charge	Accurate mass [Da]	Included
[M-H]-	1	-1.00782503207	☑
[M-H2O-H]-	1	-19.01838971207	☑
[M+Na-2H]-	1	20.97411921676	☑
[M+Cl]-	1	34.96885268	☐
[M+K-2H]-	1	36.94805661586	☐
[M+FA-H]-	1	44.99765396793	☑
[M+Hac-H]-	1	59.01330396793	☐
[M+Br]-	1	78.9183371	☐
[M+TFA-H]-	1	112.98503896793	☐
[2M-H]-	1	-1.00782503207	☑
[2M+FA-H]-	1	44.99765396793	☑
[2M+Hac-H]-	1	59.01330396793	☐

Next: The « alignment » tab provide several options for data alignment across acquisitions. Feature filtering options are not necessary since MS-CleanR provides more tuning for this step.



1. Select a QC for sample alignment. The retention time tolerance has to be adapted according to chromatogram inspection (RT drift). 0.1 minute is suitable for C18-UHPLC. MS1 tolerance is instrument dependent (0.025 for HRMS).
2. RT factor and MS1 factor can be kept to 0.5. Other features filtering options are not necessary in the case of further MS-CleanR processing.

—>**Finish:** Processing time depends on the number of threads, number of samples and baseline threshold. This process is highly ressource consuming. Try several parameters, particularly baseline threshold to obtain satisfactory results.

Crash may be observed if not properly formatted DBs were used. In this case, the « identification » tab can be discarded at this stage.

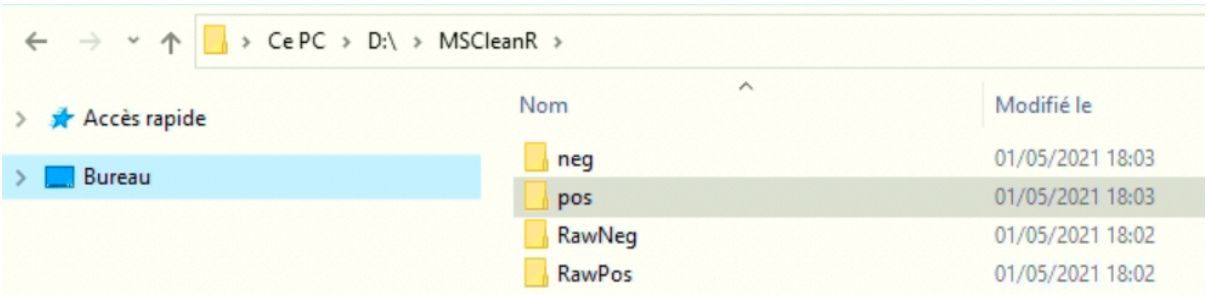
2. Data preparation for MS-CleanR processing workflow

MS-Dial provides a comprehensive GUI to explore your data. Refer to the official tutorial on CompMS website for more information.

MS-CleanR leverage on MS-Dial outputs and provides some tools to filter features, optionally combine pos and neg ionization mode and finally merge annotation results from MS-Finder to the feature data integration.

a. Prepare folders to export data tables as follows:

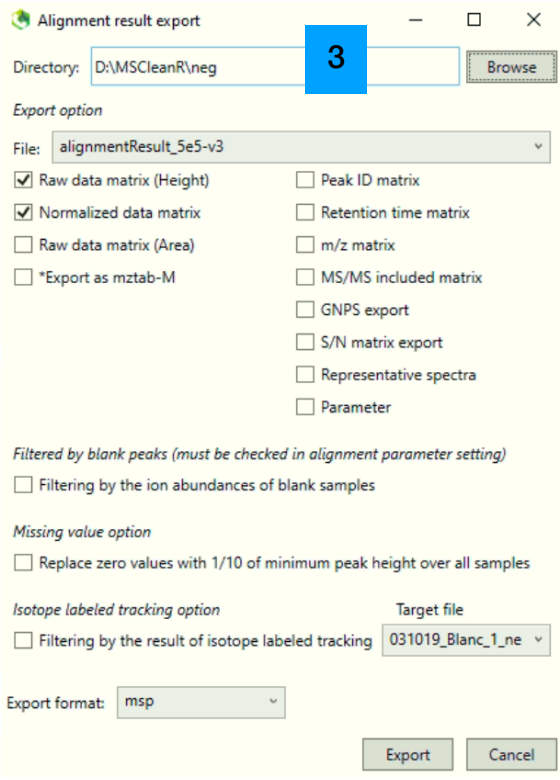
- One main directory for the whole data processing
- Within the main folder, one directory called « neg » for negative ionization mode processing
- One directory called « pos » for positive ionization mode processing
- In each « pos » and « neg » folders, a new subfolder called « peaks » must be created.



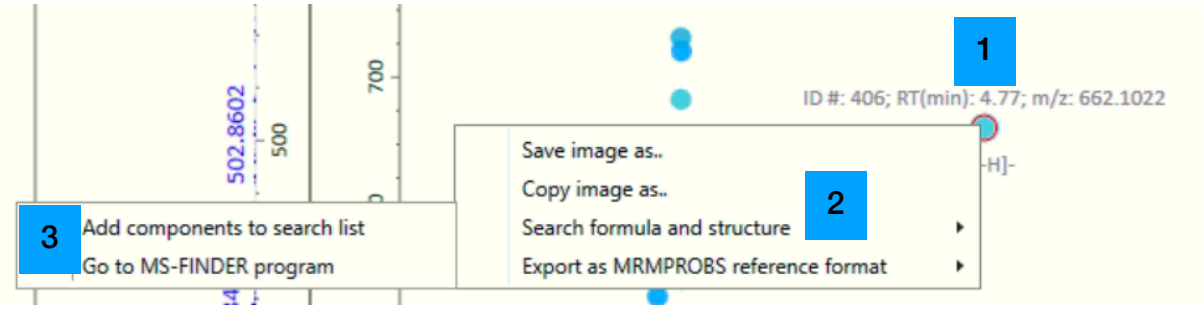
« pos » and « neg » are for MS-CleanR processing, « RawNeg » and « RawPos » host raw data files used for MS-Dial processing

b. The first step is to normalize data and export data tables of features alignment.

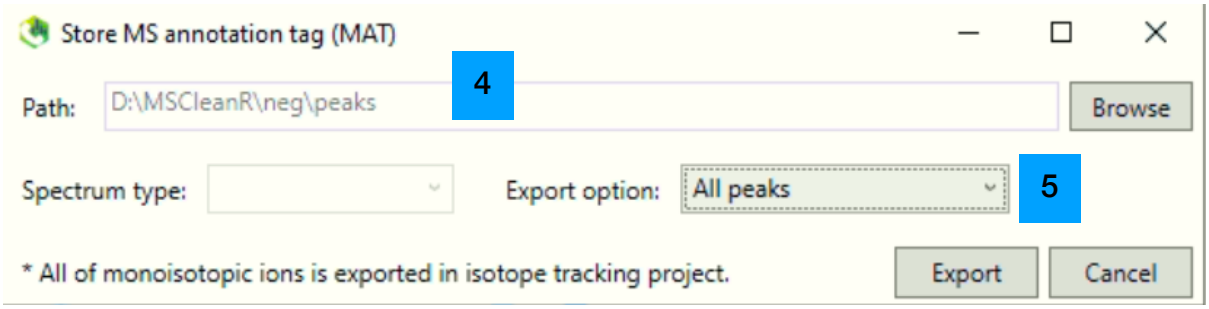
1. *Select data visualization—>feature normalization*
2. *Select proper normalization, by default, normalization by TIC.*
3. *Go to « export » menu and export « Raw data matrix (Height) and « Normalized data matrix » in « pos » or « neg » directory depending of the ionization mode being treated.*



c. The second step is to export .MAT files of all detected features in the « peaks » subfolder previously created in « neg » or « pos » directory.



1. *Left-click on any dots from the main 2D map panel*
2. *Select « search formula and structure »*
3. *Select « Add to component search list »*

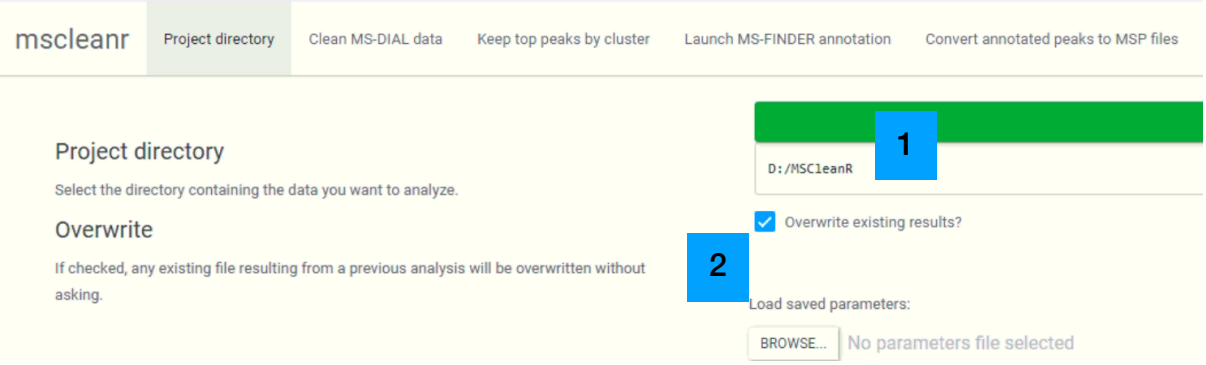


4. Browse file to the « peaks » directory in either « pos » or « neg ».
*5. **Mandatory**: In export options select « All peaks »*

3. MS-CleanR workflow

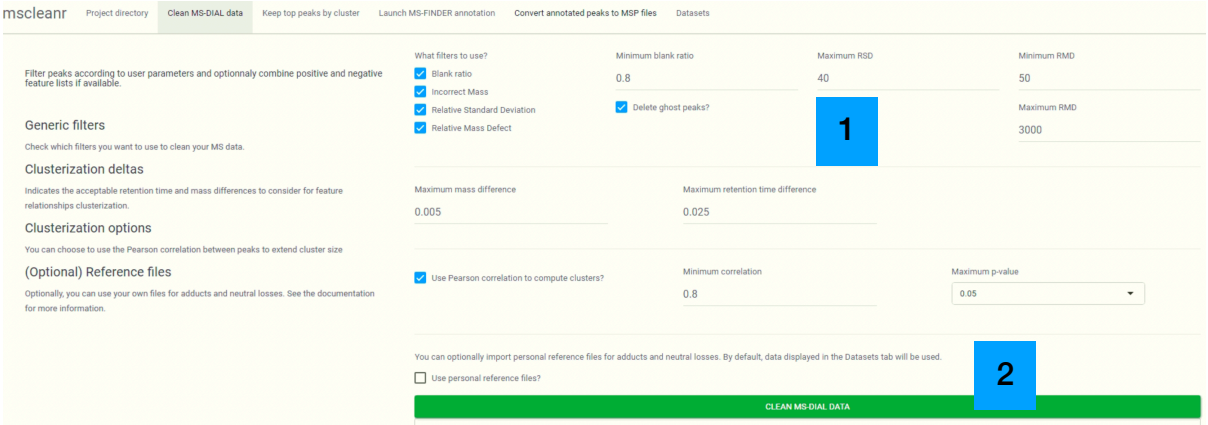
MS-CleanR is an R package with a shiny interface as GUI. We advise using Rstudio to manage installation. Copy paste the following commands for proper installation in R:

```
>setRepositories()
1 2
>install.packages("devtools")
>install.packages("vroom")
>install.packages("igraph")
> devtools::install_github("eMetaboHUB/MS-CleanR")
>library(mscleanr)
>runGUI()
```



1. *Click on the green banner « project directory » to select the folder containing « pos » and « neg » subfolders with aligned data tables and « peaks » subfolders.*
2. *Optionally: previous results may be overwritten. At the end of the processing workflow, parameters can be saved and recall using « load saved parameters ».*

Next: The « clean MS-DIAL data » tab aims to apply several feature filtering and clustering process within a given RT window. Several parameters are optional and can be tuned. Overall, the number of filters selected may result in more or fewer features count in the processed dataset.



1. *Select feature filtering process to apply. Optionally, features may be grouped by Pearson correlation in a given RT window. This process applies a second layer to clusters extracted from MS-DIAL Peak Character Estimation Algorithm (MS-DIAL-PCE).*
2. *Click on the green banner to launch the process (1 to 5 minutes, depending of features numbers to handle)*

Command	Description
Blank ratio	Subtract blank peaks to samples based on the indicated "**Minimum blank ratio**" by default at 0.8. This operation is done on the **Height files** between Blanks and QCs. **Tips**: Carefully inspect your raw data between blanks and QCs to set a proper value.
Incorrect Mass	Delete all peaks with a mass defect in X.8 and X.9 which appear to be artifacts. **Tips**: If working on C,H,N,O compounds only, this option can be used.
Relative standard Deviation (RSD)	Filter based on the **Maximum RSD** value set at 30 by default. The RSD is calculated on each defined class. If RSD of one feature is under the defined value for all class, it is removed from the peak list. **Tips**: To set a proper value, we advise superposing QC chromatograms and inspect major and minor peaks.
Relative Mass Defect (RMD)	RMD is calculated in ppm as ((mass defect/measured monoisotopic mass) × 10e6) **Tips**: Analysis of natural products from the DNP shows that 95 % of RMD are comprised between 50 and 3000 (values by default).
Delete ghost peaks	Delete variables with *m/z* values corresponding to mass drift of blank peaks.
Maximum mass difference	m/z value tolerance set by default to 0.005 (in Da) for Pearson correlation and pos/neg merging. This parameter depends of MS instrument used. **Tips**: A value tow times above the threshold used in MS-DIAL is a good starting point
Maximum retention time difference	RT value tolerance set by default to 0.025 (absolute value) for Pearson and pos/neg merging. This parameter will depend of chromatographic condition. **Tips**: A value tow times above the threshold used in MS-DIAL is a good starting point
Use Pearson correlation to compute clusters?	Extend MS-DIAL-PCE clusters with Pearson correlation. **Minimum correlation** and **maximum p-value** are respectively set by default to 0.8 and $p \leq 0.05$ **Tips**: This function increase cluster size and consequently reduce the number of features in the final annotated peak list.

Details of each option for feature filtering available in MS-CleanR

During this step:

- Clusters are formed based on MS-DIAL-PCE algorithm, Pearson correlation, links such as adducts, neutral losses, dimers, …;
- Adducts are corrected based on previous found links;
- If both modes were acquired, Pos and Neg clusters are concatenated if relational links are found (adducts mass difference)
- Once the cleaning is done, one new folder is created named "intermediary_data" in the working directory. Some details are displayed at the bottom of the index "Clean MS-DIAL data" and written in several files to inspect the results.
-

Files	Description
Adducts_massdiff_filtered	Reference file for mass difference between regular adducts
Adducts_massdiff_total	Reference file for mass difference between all possible adducts
Adducts_detected_by_MSDIAL	Reference file for adduct ponderation of regular adducts found by MS-DIAL
Adducts_filtered.graphml	A graph to display feature clusters based on adducts links
Adducts_final_selection	Final adducts resulting from MSdial and modified after pos/ neg concatenation
Adducts_initial.graphml	A graph to display feature clusters based on MS-DIAL data
Annotated_MS-peaks-MSDial	List of annotated peaks based on the database (msp file) imported in MS-DIAL
Deleted_blank_ghosts	List of peaks deleted with “delete ghost peaks”
Deleted_blanks	List of peaks deleted with the filter “blank ratio”
Deleted_mz	List of peaks deleted with the filter “incorrect mass”
Deleted_rmd	List of peaks deleted with the filter “RMD”
Deleted_rsd	List of peaks deleted with the filter “RSD”
Links_clusters_final	List of correlation (adduct, neutral loss, msdial) between peaks in neg and pos
Links_post_selection	Feature links after adduct prioritization process
Links_pre_selection	Feature links with all adducts possibilities
MS_peaks-clusters.graphml	A graph of final clusters (MS-DIAL + Pearson)
MS_peaks-clusters_final	List of final clusters (MS-DIAL + Pearson) in both pos and neg ionization
MS_peaks-clusters_msdial	List of MS-DIAL clusters in both pos and neg ionization
samples	List of samples with indication of sample name, class, file type, script class and column name

Details of files written in « intermediary data » folder during the filtering process

Possible errors during this step:

```
Error: undefined column selected
```

mismatch between *class names* in “pos” and “neg” mode

```
Error: Can't process data without blanks.
```

Blank samples not defined in “File type”

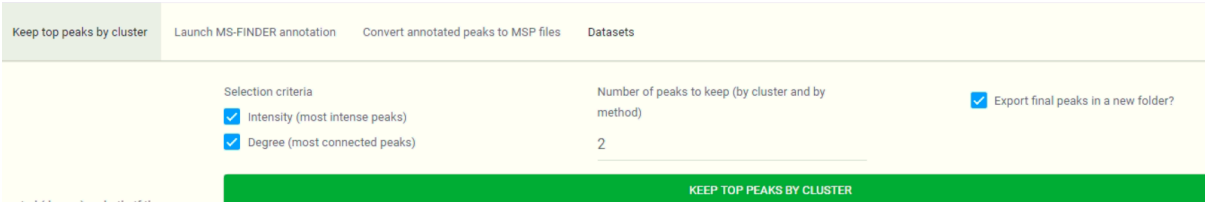
- *At least 3 blanks and 3 QCs samples are needed for Blank ratio analysis. These samples must be identified as such in the MS-Dial sample list.*
- *Avoid spaces in samples or classes names and replace it by "-"; "." or "_"*
- *Avoid class names with only one letter*
- *MSCleanR handle LCMS acquired in DIA or DDA mode.* ***All features without MS/MS will be discarded during the first step. If no MS/MS are detected, an error will arise during this process.***

Next: In the third tab “Keep top peaks by cluster” you can select the number of features you want to keep in each cluster.

This step is based on the hypothesis that in one cluster, only one unique metabolite is present. The other variables used to come from feature degeneration. Generally, this metabolite appears to be the **most intense** and/or **the most connected within the** graph (adducts, neutral loss, dimers…).

You can then choose to select as many peaks as you want and either the most intense(s) by clicking “**Intensity**”, the most connected by clicking “**Degree**” or both.

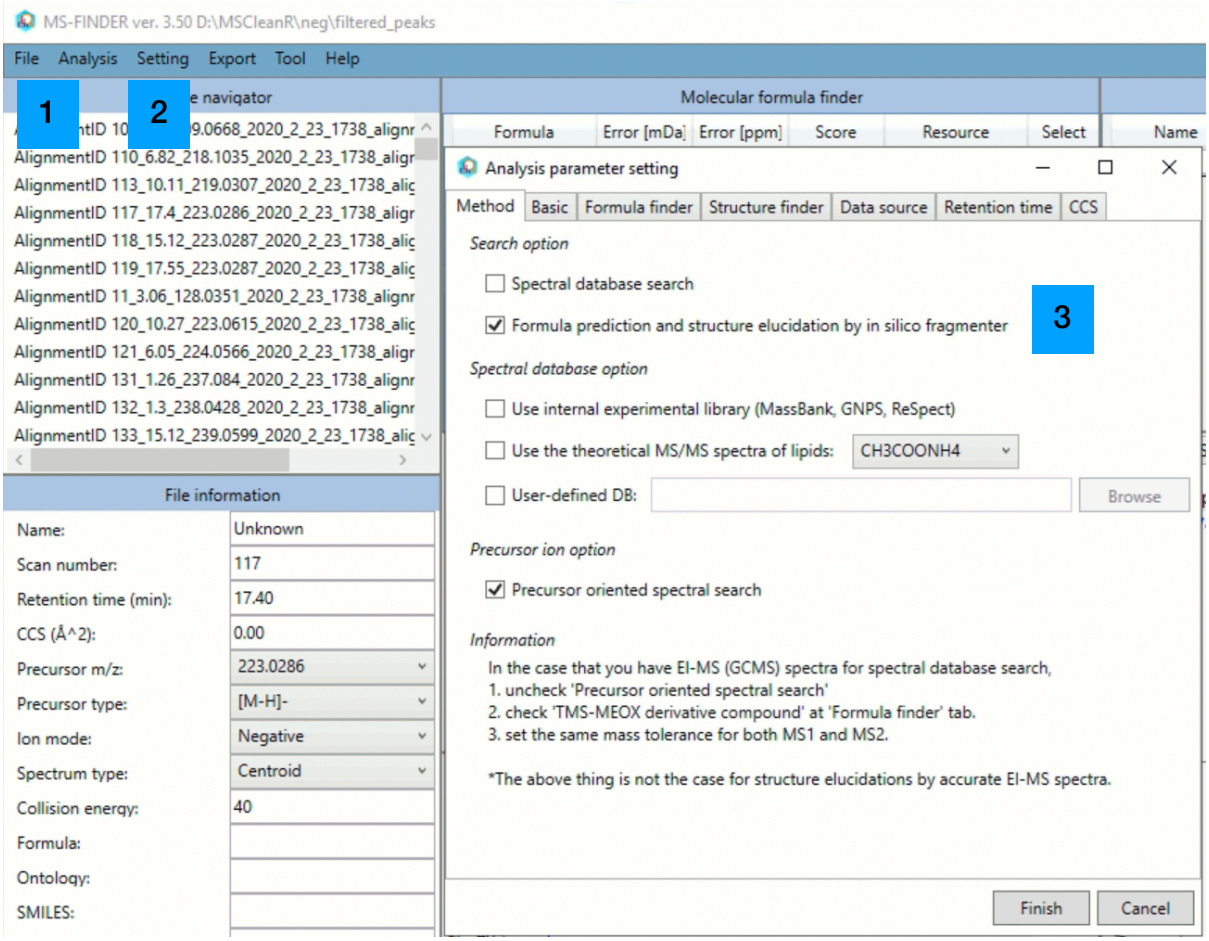
***Tips**: We advise to select both criteria and keep 2 top peaks by cluster for further MS-finder request.*



At this step, a new folder is created in both « pos » and/or « neg » folders named “**filtered peaks**”. All .MAT files corresponding to selected peaks are copied from “peaks” folder and pasted in « filtered peaks ». At this step, the annotation process can start using MS-FINDER.

4. MS-FINDER annotation

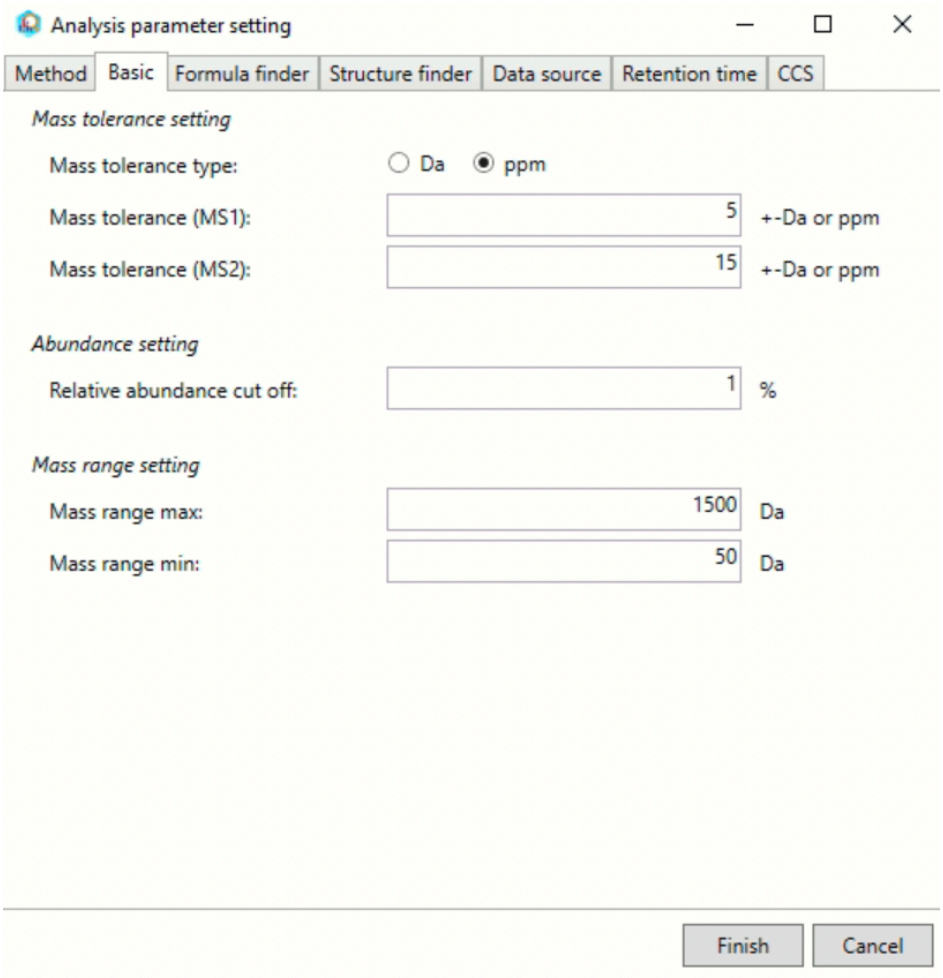
MS-FINDER provides an all-in-one solution for molecular formula calculation and structure elucidation based on spectral database or *in-silico* fragmentation matches. This tutorial focus on the *in-silico* fragmenter based matching.



1. *Import .MAT files from the « filtered_peaks » folder previously created by MS-CleanR (File—>import)*
2. *Go to « setting—>parameter setting »*
3. *For MF and structure prediction based on in-silico fragmenter (using a list of putative compounds in SMILE format). Spectral DB may be requested using « Spectral DB option »*

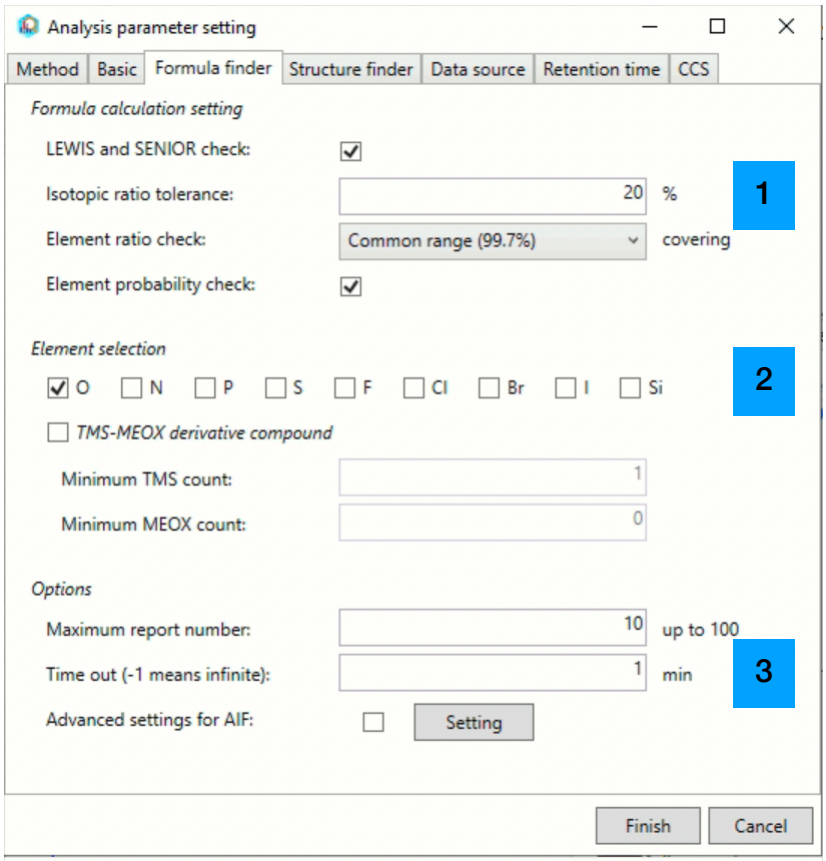
The second tab relies with mass accuracy settings.

Put mass range settings and mass tolerance according to your acquisition parameters.



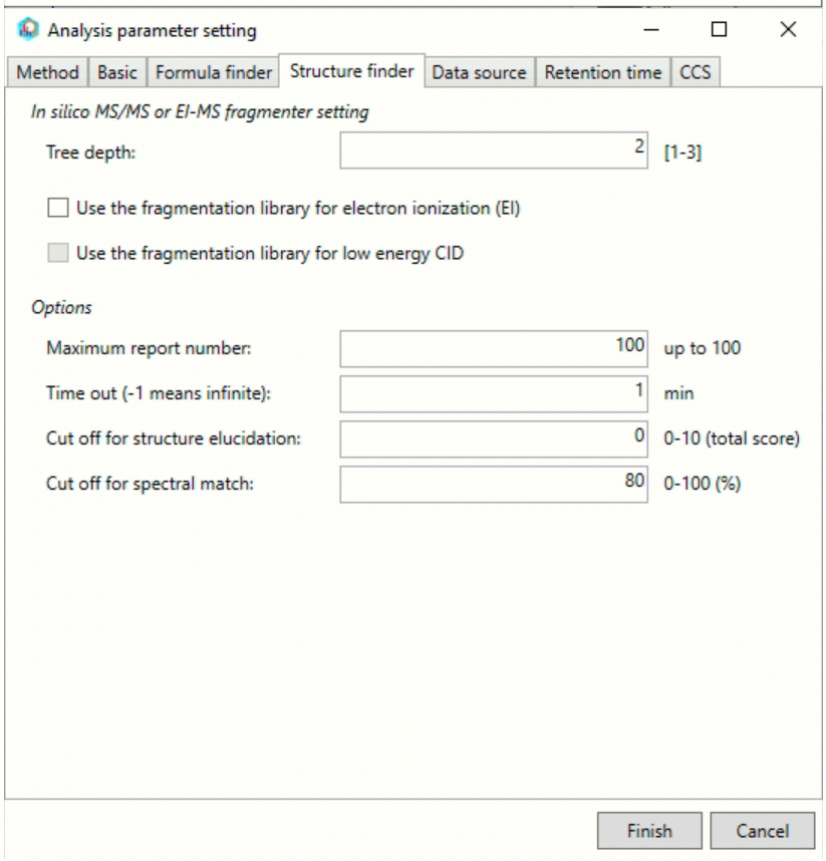
The third tab defines MF calculation.

1. *We advise users to tick every « formula calculation settings » to respect the «seven golden rules ».*
2. *Elements should be chosen according to expected compounds. More elements are selected, more calculation time increase.*
3. *Set the max number of MF to retain for structure elucidation and define a time-out to apply a tile limit for MF calculation of each feature*



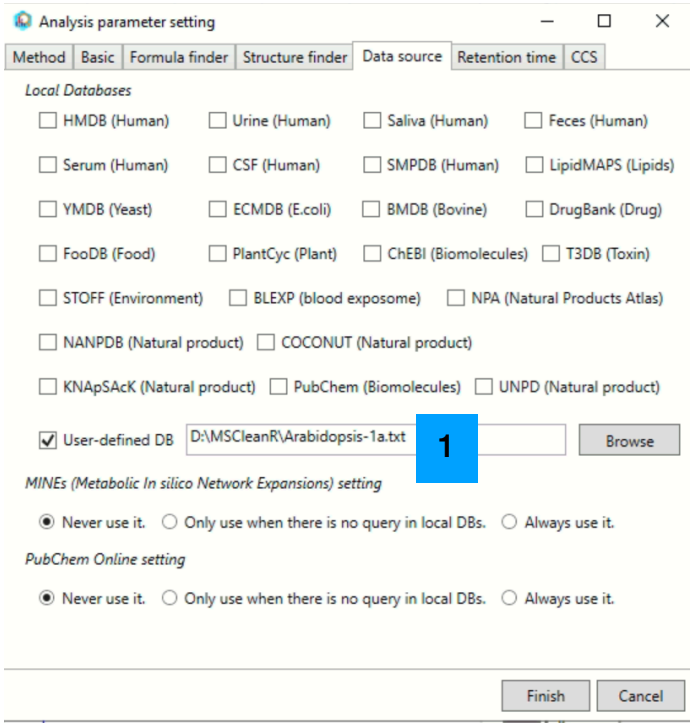
The structure finder tab defines settings for *in-silico* fragmenter:

Tree depth: To restrict in silico cleavages. With 2, you generate fragments until product ions of a product ion.
The available options may constrain the number of results according to the total score (based on RT, MF accuracy and spectral match) or spectral match only.



The « data source » tab allows the selection of several DBs included in MS-Finder.

1. *Another option is to import an in-house DB. This DB encompass structure « name , InChiKey, AccurateMass, Formula, SMILES and Database-ID ». Several other columns may be added and will be reported in the final annotation result.*
2. *MS-CleanR is able to prioritize annotation matches within a « in-house » DB by adding a « Compoun_level » column to the files. See next section for details.*



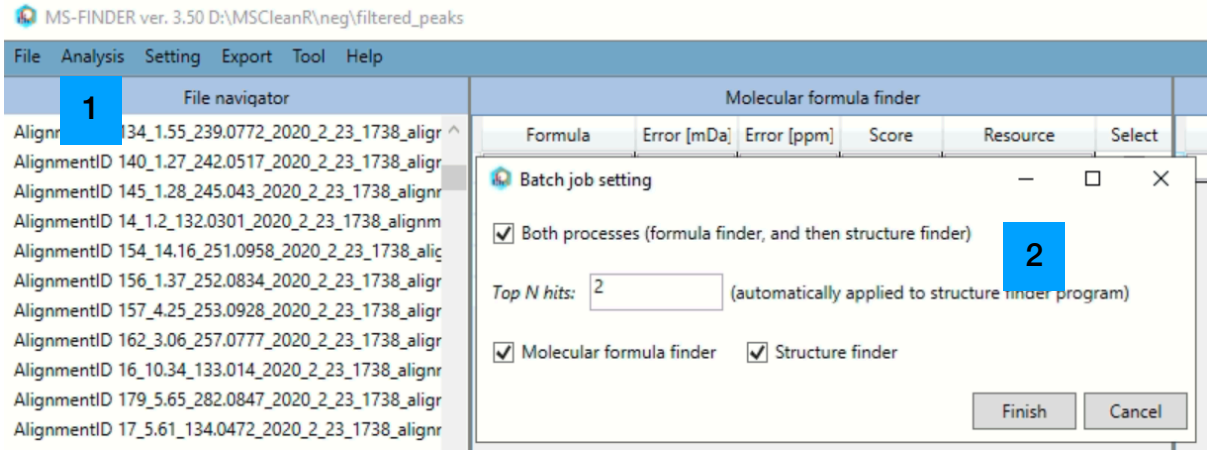
Name	InChiKey	PubChemID	Accuratemass	Formula	Smiles	Internal_id	Compound_level
(2S,3R,4R,5S)-2-[(1R,3aS,3bS,5S,5aS,7R,9	QQPPQQNYVDYNAA-VKTWQGKFSA-N	NA	450.3709	C28H50O4	CC(C)[C@H](C)[C@@H](O)[C@H](O)[C@@H](C	250535	1a
1,3-Propanediamine	XFNJVJPLKCPIBV-UHFFFAOYSA-N	NA	74.084398	C3H10N2	NCCCN	46860	1a
3,4',5,7-Tetrahydroxyflavylium(1+)	XVFMGWDSJLBXDZ-UHFFFAOYSA-O	NA	271.06065	C15H11O5	OC1=CC=C(C=C1)C1=[O+]C2=CC(O)=CC(O)=C2C	37557	1a
5,6-Dihydroxy-1H-indole-2-carboxylic aci	YFTGOBNOJKXZJC-UHFFFAOYSA-N	NA	93.037509	C9H7NO4	OC(=O)C1=CC2=C(N1)C=C(O)C(O)=C2	39975	1a
9-Tetradecenoic acid|9-Tetradecenoic ac	YWWVWXASSLXJHU-AATRIKPKSA-N	NA	226.19328	C14H26O2	CCCC\C=C\CCCCCCCC(O)=O	43951	1a
6-Hydroxy-1H-indole-3-carboxylic acid; C	KVKJZZMEYYZNDI-UHFFFAOYSA-N	NA	339.095419	C15H17NO8	OCC1OC(OC2=CC3=C(C=C2)C(=CN3)C(O)=O)C(C	1009329	1a
5-methylthioadenosine	WUUGFSXJNOTRMR-IOSLPCCCSA-N	NA	297.0896	C11H15N5O3S	CSC[C@H]1O[C@H]([C@H](O)[C@@H]1O)N1C=	43824	1a
{[(1S,2R,3S)-2-[(3Z)-4,8-dimethylnona-3,7	ATZKAUGGNMSCCY-VYCBRMPGSA-K	NA	586.3188	C30H52O7P2	CC(C)=CCC\C(C)=C/CC\C(C)=C\[C@H]1[C@H](C	1189	1a
(2R)-2-amino-3-{[(2E,6Z)-3,7,11-trimethy	SYSLNQMKLROGCL-BCYUYYMPSA-N	NA	325.2076	C18H31NO2S	CC(C)=CCC\C(C)=C/CC\C(C)=C\CSC[C@H](N)C(C	37294	1a
5-carboxy-4,5-dihydro-3H-pyrrol-2-olate	ODHCTXKNWHHXJC-UHFFFAOYSA-M	NA	129.0426	C5H7NO3	OC(=O)C1CCC([O-])=N1	22020	1a
(1R,3aS,3bR,5aS,7R,8S,9aR,9bS,11aR)-1-(	RFTIRMPAYWBDKJ-GSTFJIRZSA-N	NA	334.2508	C21H34O3	CC[C@H]1C(=O)C[C@H]2[C@@H]3CC[C@H]4C	754445	1b

1

2

***Optionally**, an in-house DB of SMILE against retention time may be added in the « Retention time » tab using « Calculated by XlogP equation » parameter.*

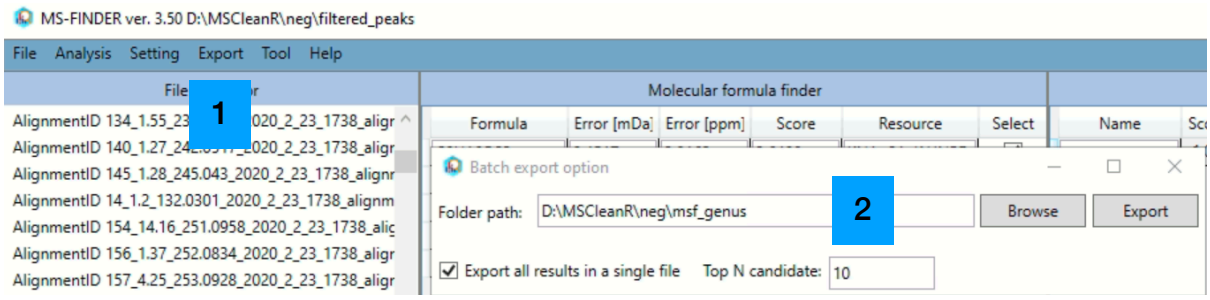
Click on « finish », then:



1. Analysis—>Compound annotation (batch job)

2. If mass accuracy of your instrument is below 5 ppm, we advise selecting only the 2 first MF for structure elucidation process. Calculation time may be long depending of number of features to proceed.

After MF and structure elucidation processing, create a subfolder in either « neg » or « pos » directory called « msf_xxx » (xxx = genus, family, or what ever else).

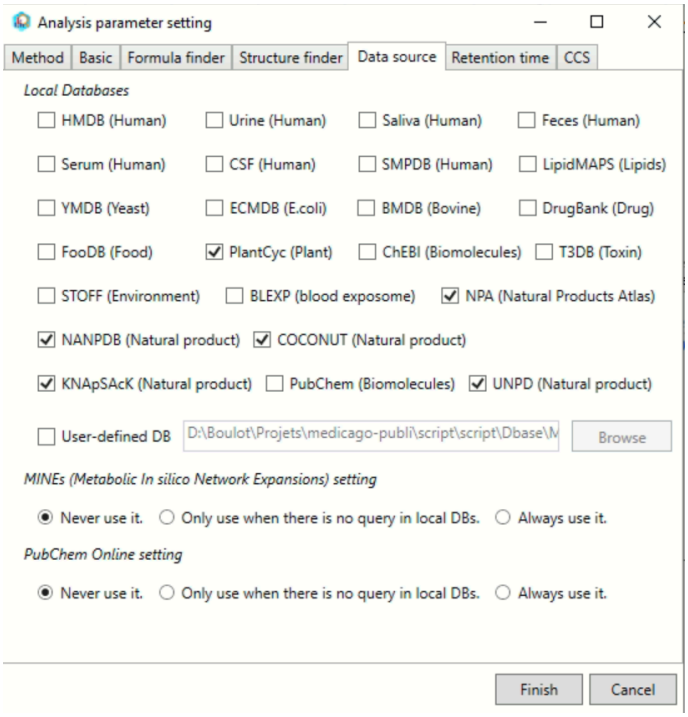


1. *In « export » select « export batch results »*
2. *In folder path, select the folder « msf_xxx » and tick « export all results as single file ». There is no limitation to top N candidates (up to 100).*
3. *Check if « Formula result-xxx.txt » and "Structure result-xxx » have been written to the msf_xxx folder.*

Several request can be performed using many DBs as needed on the same set of features (parameter settings—>Data source->select suitable DB—>Batch job annotation).

The important thing is to store the final « Formula result » and « Structure result » files in separated msf_xxx folders.

If pos and neg modes are being processed, the same « msf_xxx » subfolders must be present in each directory. For instance, a second interrogation may be done using internal MS-FINDER DBs:

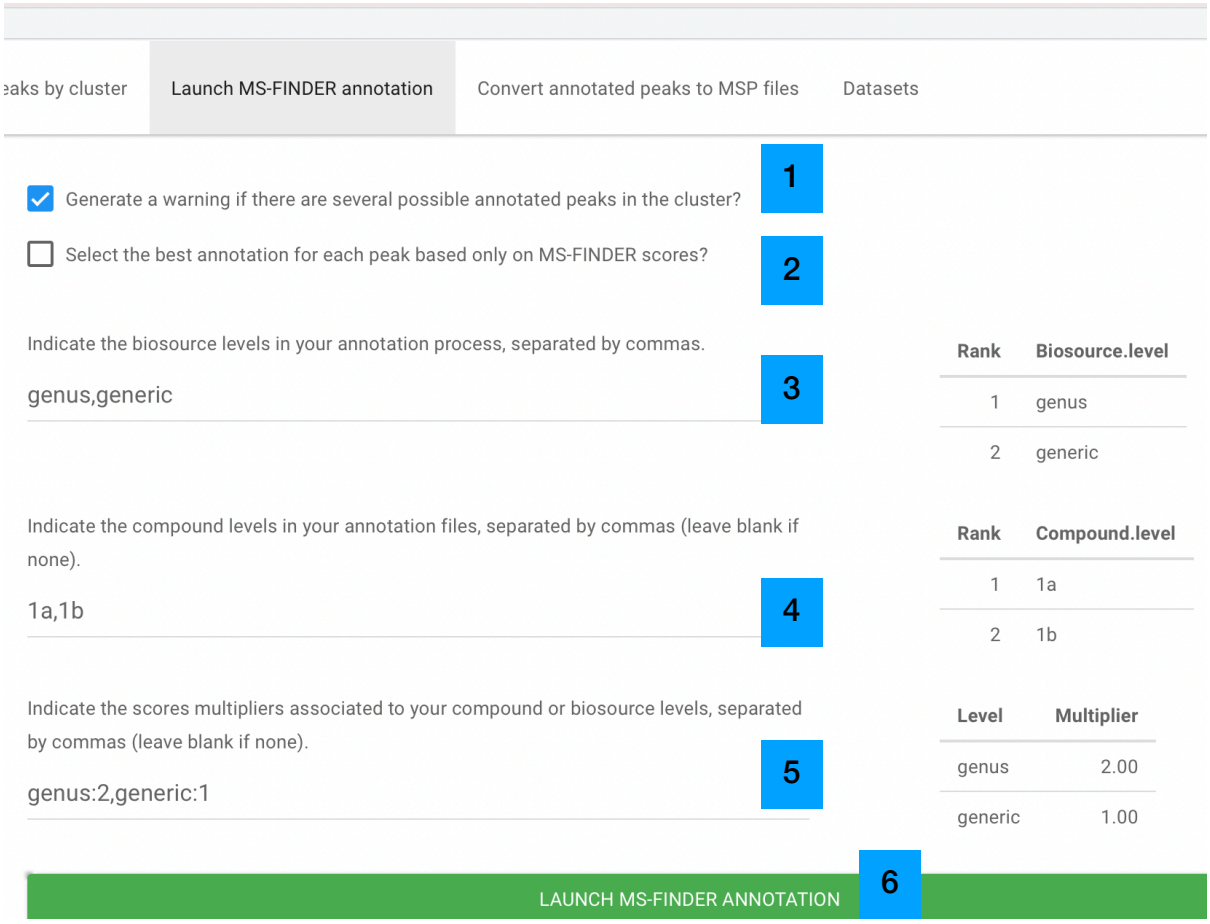


In this case, results are stored in « msf_generic » subfolder. **This folder name is mandatory for proper MS-CleanR annotation.**

5. Merge annotation results in MS-CleanR

As MS-FINDER annotation is complete, come back to the shiny interface of MS-CleanR and switch to « Launch MS-FINDER annotation » tab.

This step will merge feature annotation to the « m/z x RT x area » dataset using either only MS-FINDER score or based on the prioritization of the different DBs used to indicate the more pertinent annotation.



1. *The « generate warning option » add a column in the final result data table to indicate whether annotation need to be checked. This warning arises when several compound matches for a given feature.*
2. *If annotation is based on « best MS-FENDER score », the annotation process take the best score of all matches for each feature without DBs prioritization*
3. *DBs prioritization 1: in case several DBs were used for feature annotation, rank each DB according to their priority during the annotation process. If several matches are found in different DBs for a given feature, it's always the match of the first ranked DB which will be selected.*
4. *Optional: DB prioritization 2: If you add a « Compound_level » column to your internal DB with a compound ranking number, details the rank used here.*

5. *Optional: Give a multiplicative number to the MS-FINDER score according to the DB rank. It has no influence on the annotation results.*
6. *Click on the green banner to launch data merging.*

Two files are created in the "**final-data**" folder:

Annotated MS peaks cleaned = the final peak list with annotation from MS-FINDER

Annotated MS peaks normalized = the final peak list renormalized based on total peak area

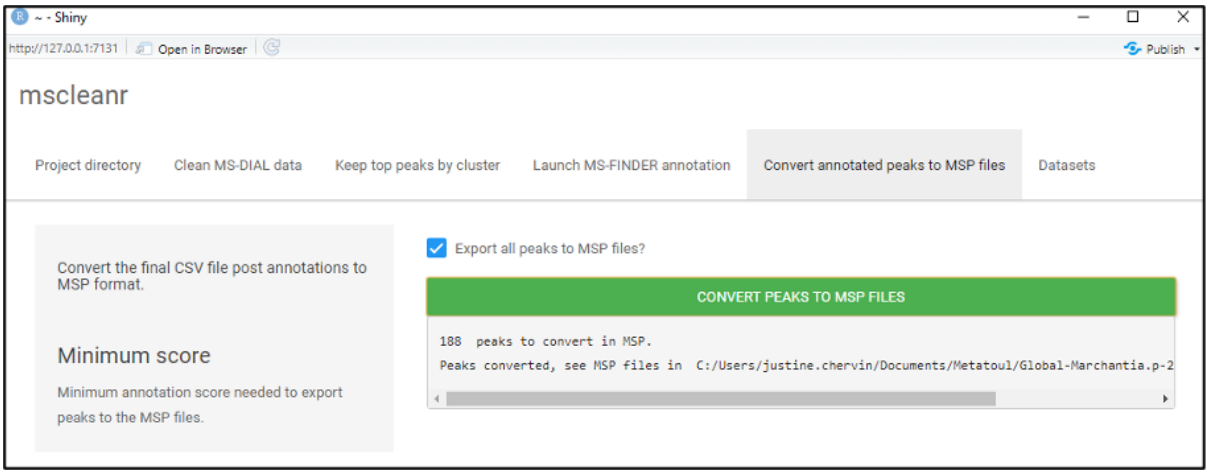
Unknown error can arise during normalization of the "Annotated MS peaks cleaned". In this case, "Annotated MS peaks normalized" will not be created and have to be calculated manually using excel for instance.

The final peak list looks like as follows. Different information are available such as:

- The average m/z value;
- The average RT value;
- The annotation based on MS-FINDER interrogation on the "**Structure**" column with the associated **Total score** of MS-FINDER and **Final score** calculated from the indicated multipliers.
- The source of the annotation in the "**level**" column;
- The ontology of the compound; ...

The features are also identified as:

- Unknown compound = variable with no annotation
- Simple ID = based on a single feature in pos or neg mode
- Double ID =based on same annotation retrieve in pos and neg mode



Optionally: In the fifth tab "Convert annotated peaks to MSP files", you will be able to create two .msp files named « peaks-neg.msp » and « peaks-pos.msp » in the folder « final_data ». All peaks can be converted, or user can choose a scoring threshold based on multiplied MSfinder score. One metadata file per ionization mode is also created containing annotation results and average peak area of each class.

These two files could then be imported in MetGem software or GNPS facility to create mass spectral similarity networks.