HALOSEEKER Documentation

User Guide NOT-SCI-001 version 002 for HaloSeeker 2.0

TABLE OF CONTENTS

1. GENERAL CONSIDERATIONS 5
1.1. Purpose 5
1.2. Citing HaloSeeker 5
1.3. Other references useful to the development of HaloSeeker 5
1.4. Examples of references taking advantage of HaloSeeker 7
1.5. Document version 8
2. GETTING STARTED 9
2.1. Installation 9
2.2. Launching 9
3. PRESENTATION OF THE INTERFACE AND WORKFLOW OVERVIEW 9
3.1. Interface arrangement 9
3.2. Title bar 10
3.3. Menu 11
3.4. Plot interactivity 12
3.5. Tables 13
3.6. Workflow overview 13
4. TAB PROJECTS & FILES 14
4.1. Sub-tab New Project 14
4.1.1. Create a new user 14
4.1.2. Create a new project 14
4.1.3. Import files to a project 15
4.1.3.1. From hard drive 15
4.1.3.2. From other projects 17
4.2. Sub-tab Database tables 18
4.2.1. Project 18
4.2.2. Sample 19
4.2.3. Compound 20
4.2.4. Adduct 21
5. TAB DECONVOLUTION 22
5.1. Sub-tab Process 22
5.1.1. Peak-picking 23
5.1.2. Pairing 24
5.1.3. Fragment/adduct 26
5.1.4. Alignment 26

5.2. Sub-tab Result 27
6. TAB EXPLORE DATA 28
6.1. Sub-tab Interactive plot 28
6.1.1. Overview 28
6.1.2. Axes & Legends 28
6.1.3. Blanks 30
6.1.4. Filters 30
6.1.4.1. m/z and retention time range 30
6.1.4.2. Pairing and ratio annotation status of the paired clusters 30
6.1.4.3. Cluster slope 33
6.1.4.4. Annotation status 33
6.2. Sub-tab TIC/EIC & MS 33
6.3. Sub-tab Enumeration 34
7. SUPERGROUP INVESTIGATION MODAL 35
7.1. Overview 35
7.2. Decomposition parameters 37
7.3. Results 39
7.4. EIC, Slope and Table 39
7.5. SuperGroup table 40
7.6. Hybrid mass spectrum 41
7.7. Formula annotation 42
7.7.1. Reset annotation 42
7.7.2. Discard status 42
7.7.3. Annotate formulas 42
7.7.4. Annotate compound 43
7.7.4.1. Create a new compound 44
8. FREQUENTLY ASKED QUESTIONS (FAQ) 44

1. GENERAL CONSIDERATIONS

1.1. PURPOSE

HaloSeeker is an open-source software in *R*, developed with the Shiny framework for the post-acquisition processing, visualization, analysis and annotation of signals from halogenated (chlorine and/or bromine) ions within high resolution mass spectrometry (HRMS) data sets. HaloSeeker 2.0 is freely available, under the GPLv3 license, on request at contact.haloseeker@oniris-nantes.fr. HaloSeeker 2.0 is packed as an executable file of about 335 MB and runs out of the box without need for computer skills or administrative rights.

The present document aims at describing HaloSeeker 2.0 workflow and functionalities.

1.2. CITING HALOSEEKER

Whenever citing HaloSeeker 2.0, please use the following reference presenting the original version, HaloSeeker 1.0. To date, no peer-reviewed article deals with HaloSeeker 2.0 yet.

- Léon A, Cariou R, Hutinet S, Hurel J, Guitton Y, Tixier C, Munschy C, Dervilly-Pinel G, Antignac J-P, Le Bizec B. HaloSeeker 1.0, a user-friendly software application for highlighting halogenated chemicals in non-targeted high resolution mass spectrometry dataset. *Analytical Chemistry 2019, 91 (5), 3500–3507*. https://pubs.acs.org/doi/10.1021/acs.analchem.8b05103.

1.3. OTHER REFERENCES USEFUL TO THE DEVELOPMENT OF HALOSEEKER

Pairing scrip

The following reference presents the original (prior to HaloSeeker 1.0) pairing script. This Visual Basic for Application script was translated into *R* language, improved (see Léon et al., 2019) and embedded in HaloSeeker (1.0 and 2.0).

- Cariou R, Omer E, Léon A, Dervilly-Pinel G, Le Bizec B. Screening halogenated environmental contaminants in biota based on isotopic pattern and mass defect provided by high resolution mass. *Analytica Chimica Acta 2016, 936, 130–138*. https://dx.doi.org/10.1016/j.aca.2016.06.053.

msConvert package, ProteoWizard

- Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak MY, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P. A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology 2012, 30, 918–920*. https://doi.org/10.1038/nbt.2377.

XCMS package, centWave function

- Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry 2006, 78 (3), 779–787*. https://doi.org/10.1021/ac051437y.
- Tautenhahn R, Böttcher C, Neumann S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics 2008, 9, 504*. https://doi.org/10.1186/1471-2105-9-504.

CAMERA package

- Kuhl C, Tautenhahn R, Böttcher C, Larson T R, Neumann S. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry 2012, 84 (1), 283–289*. https://doi.org/10.1021/ac202450g.

OBI-Warp algorithm

- Prince JT, Marcotte EM. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Analytical Chemistry 2006, 78 (17), 6140–6152*. https://doi.org/10.1021/ac0605344.

H/Cl-scale Mass Defect plots

- Taguchi VY, Nieckarz RJ, Clement RE, Krolik S, Williams D. Dioxin analysis by gas chromatography-Fourier transform ion cyclotron resonance mass spectrometry (GC-FTICRMS). *Journal of American Society for Mass Spectrometry 2010, 21 (11), 1918–1921*. https://doi.org/10.1016/j.jasms.2010.07.010.

Formula decomposition and enviPat

- Böcker S, Lipták Z. A Fast and simple algorithm for the money changing problem. *Algorithmica 2007, 48 (4), 413–432*. https://doi.org/10.1007/s00453-007-0162-8.
- Kind T, Fiehn O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics 2007, 8, 105–115*. https://doi.org/10.1186/1471-2105-8-105.
- Loos M, Gerber C, Corona F, Hollender J, Singer H. Accelerated isotope fine structure calculation using pruned transition trees. *Analytical Chemistry 2015, 87 (11), 5738–5744*. https://doi.org/10.1021/acs.analchem.5b00941.

Identification levels

- Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP, Hollender J. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science and Technology 2014, 48, 2097–2098*. https://doi.org/10.1021/es5002105.

1.4. EXAMPLES OF REFERENCES TAKING ADVANTAGE OF HALOSEEKER

So far, HaloSeeker has been used in a few peer-reviewed articles.

With our research team

- Cariou R, Méndez-Fernandez P, Hutinet S, Guitton Y, Caurant F, Le Bizec B, Spitz J, Vetter W, Dervilly G. Non-targeted LC/ESI-HRMS detection of polyhalogenated compounds in marine mammals stranded on French Atlantic coasts. *ACS ES&T Water*, *In Press (Nov-2020)*. http://dx.doi.org/10.1021/acsestwater.0c00091.
- Pourchet M, Narduzzi L, Jean A, Guiffard I, Bichon E, Cariou R, Guitton Y, Hutinet S, Vlaanderen J, Meijer J, Le Bizec B, Antignac J-P. Non-targeted screening methodology to characterise human internal chemical exposure: application to halogenated compounds in human milk. *Talanta*, *In Press (Dec-2020)*.

From independent research teams

- Chilczuk T, Schäberle TF, Vahdati S, Mettal U, El Omari M, Enke H, Wiese M, König GM, Niedermeyer THJ. Halogenation-guided chemical screening provides insight into tjipanazole biosynthesis by the cyanobacterium *Fischerella ambigua*. *ChemBioChem 2020, 21 (15), 1–9*. https://doi.org/10.1002/cbic.202000025.
- Martin D, Lobo F, Lavison-Bompard G, Guérin T, Parinet J. Effect of home cooking processes on chlordecone content in beef and investigation of its by-products and metabolites by HPLC-HRMS/MS. *Environment International 2020, 144, 106077*. https://doi.org/10.1016/j.envint.2020.106077.

1.5. DOCUMENT VERSION

The present document is the second version of NOT-SCI-001 dedicated to HaloSeeker application. It is expected to evolve frequently according to LABERCA's internal quality management system.

Table 1 reports the successive modifications of both the application and the present document, its user guide. Each new version cancels and replaces the previous one.

Table 1: Record of the successive modifications in HaloSeeker application and its user guide.

HaloSeeker application version	Modifications	Entry into force	NOT-SCI-001 documentation version	Modifications	Entry into force
1.0	• Original version	October 21st, 2018	001	• Original version	October 21st, 2018
2.0	Detailed modifications (added, changed, removed) from version 1.0 to version 2.0 are available in the "changelog.txt" file available at the root directory of the installed software. The main added functionalities are: • Reorganization of the menu • Modification of the SQLite database, which also includes converted data traces • Introduction of fragment/adduct and alignment deconvolution processes, leading to SuperGroup constitutions • Upgrade of the interactive plot, with adjustable axes and blank subtraction options • Reorganization of the annotation modal according to the SuperGroups	December 7th, 2020	002	• Thorough rewriting, according to HaloSeeker 2.0 upgrade	December 7th, 2020

2. GETTING STARTED

2.1. INSTALLATION

HaloSeeker 2.0 is packed as an executable file of about 335 MB. All needed instructions are embedded in the executable file. This software is available on Windows with a 64-bit operating system with at least 4 Go RAM. If you have any trouble installing HaloSeeker, please let us know at contact.haloseeker@oniris-nantes.fr.

To install HaloSeeker 2.0, click on the executable file. This creates a directory, by default in the common desktop. If you choose another directory, we strongly recommend that you have full write/read access.

2.2. LAUNCHING

To launch HaloSeeker 2.0, click on the file *HaloSeeker.bat* in the created directory or on the shortcut if created.

The application contains a portable version of the web browser Chrome (version 78.0.3904.97, Google) to avoid browser compatibility problems.

3. PRESENTATION OF THE INTERFACE AND WORKFLOW OVERVIEW

3.1. INTERFACE ARRANGEMENT

The application arrangement is composed (in most cases) of four parts (**Figure 1**):

- Title bar (1): select your username, project and see associated information.
- Menu (2): access to the different tabs.
- Parameters (3): parameters for the selected tab.
- Charts (4): data visualization.

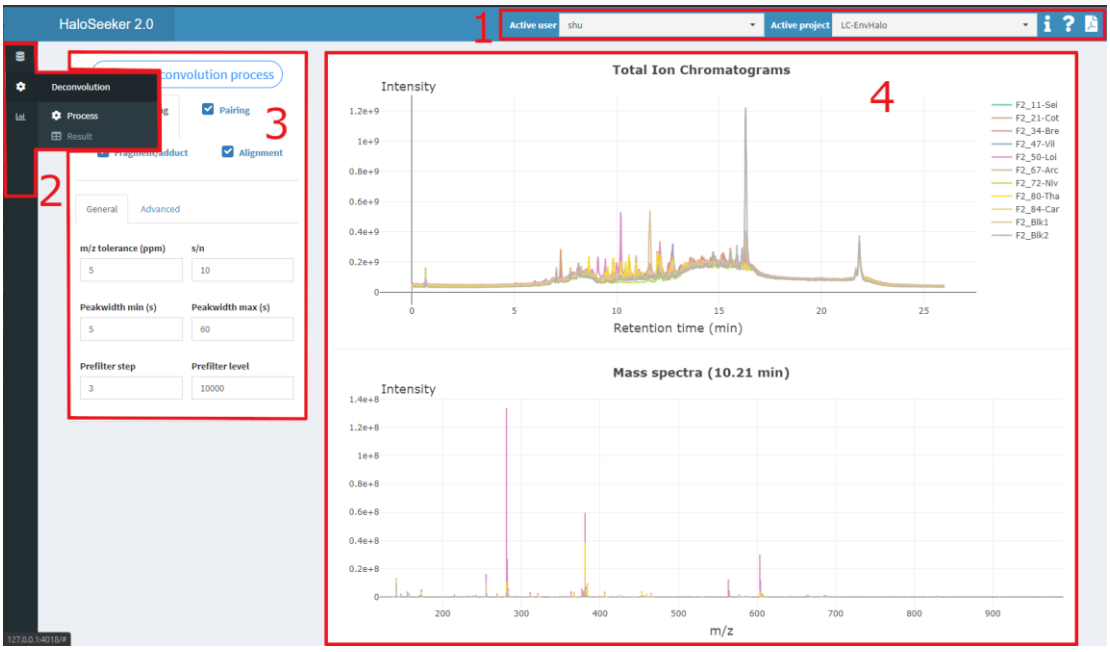


Figure 1 : Arrangement of HaloSeeker 2.0 interface.

3.2. TITLE BAR

The title bar (**Figure 2**) is composed of two picklists used to select the active user and the active project among previously defined users and projects, respectively.

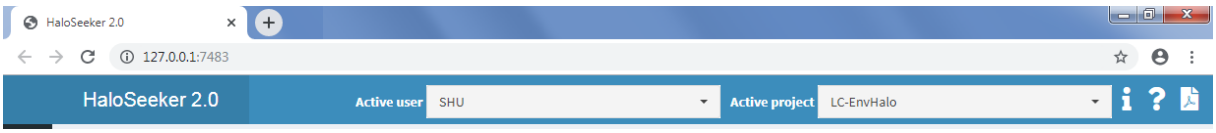


Figure 2 : Title bar.

The three buttons next to the picklists allow to (from left to right):

- Display details about the files of the active project (**Figure 3**). It includes (1) original paths of the HRMS datafiles, (2) file metadata retrieved using the *ThermoRawMetaDump* tool on data from Orbitrap mass analyzers such as polarity, instrument model, resolution, AGC target, maximum IT or scan range. If the data were processed using HaloSeeker 2.0, used parameters are also available for (3) peak picking, (4) halogen pairing, (5) alignment and (6) fragment/adduct deconvolutions.
- Trigger an interactive tutorial.
- Open the user manual in a new Chrome tab.

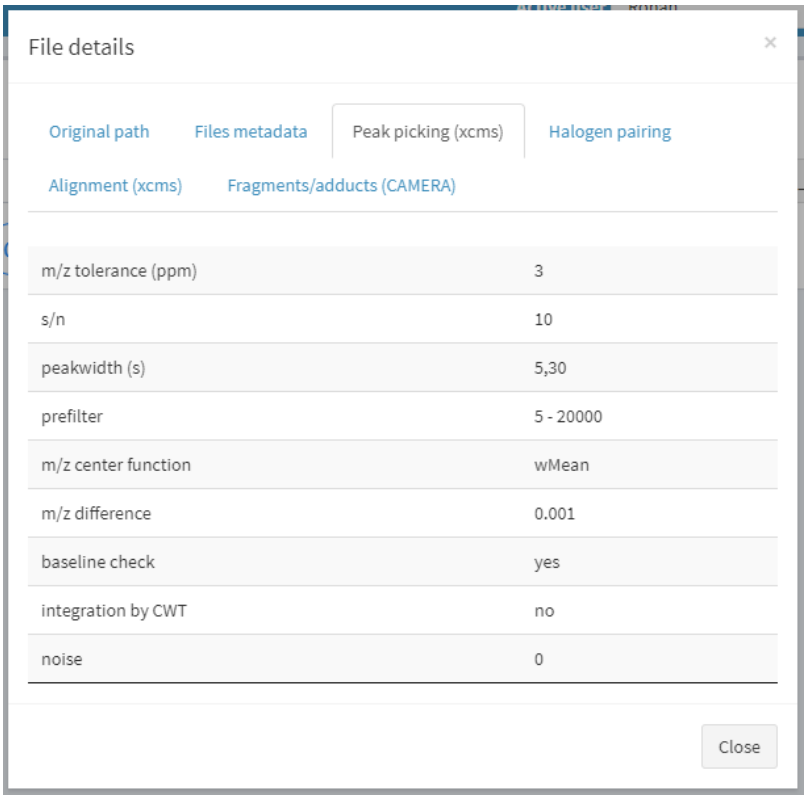


Figure 3 : File details of the active project, accessible using the button of the Title bar.

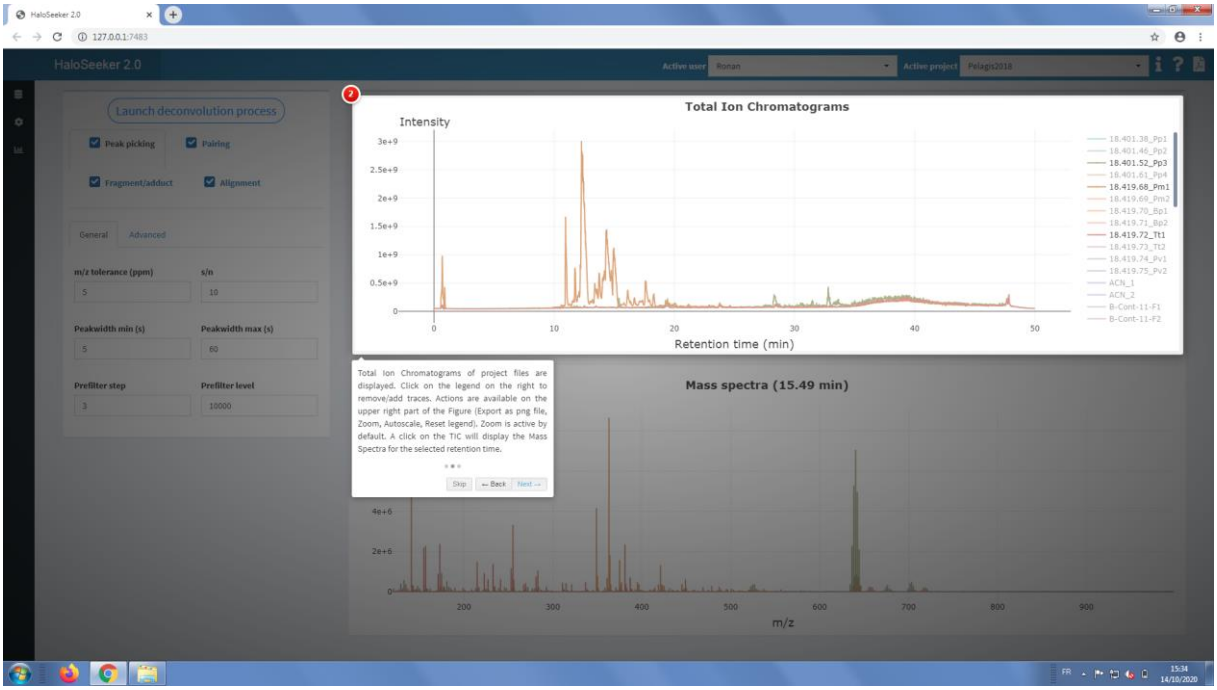


Figure 4 : Interactive tutorial, accessible using the ? button of the Title bar.

3.3. MENU

The menu is composed of three parts (**Figure 5**):

- Projects & Files: Create a new project and manage database tables
- Deconvolution: Process data and see a resulting overview.
- Explore data: Visualize data with the interactive plot, ion chromatograms and see the enumeration overview of annotated features.

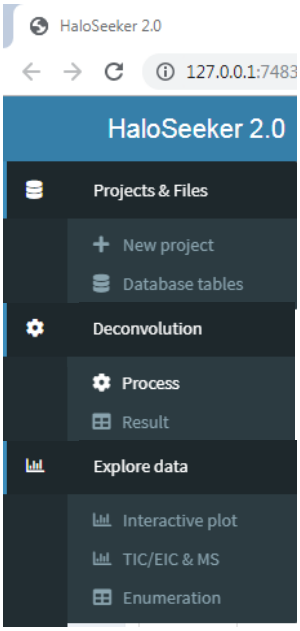


Figure 5 : Arrangement of the menu

3.4. PLOT INTERACTIVITY

Zoom-in is possible using the left button of the mouse. Axes can be reset when double-clicking on the left button of the mouse.

With the mouse over the legend, all traces turn grey except for the one pointed out, in order to highlight it.

Also, the arrow tail of any annotation can be moved to avoid overlapping.

In addition, all plots in the application exhibit a toolbar on their upper right corner (**Figure 6**), which appears when moving the mouse over the plot. Icons are available, as hereafter described from the left to the right.

- Download plot as *png* file.
- Zoom-in.
- Autoscale.
- Reset legend. Restores the legend at its initial state. A click on an item on the legend hides it from the plot, a double click isolates it, other items being hidden.

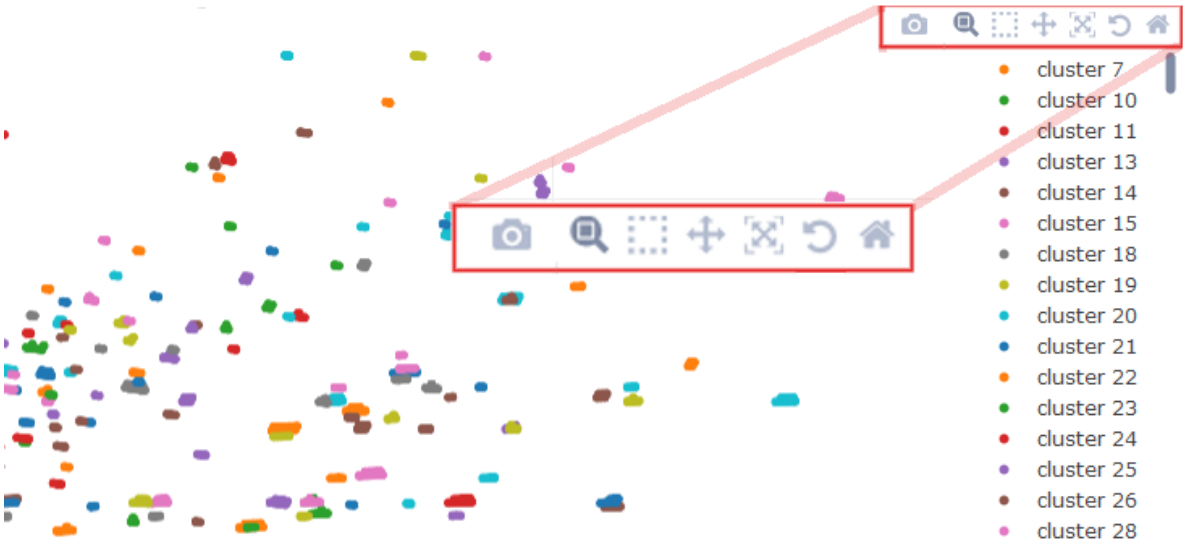
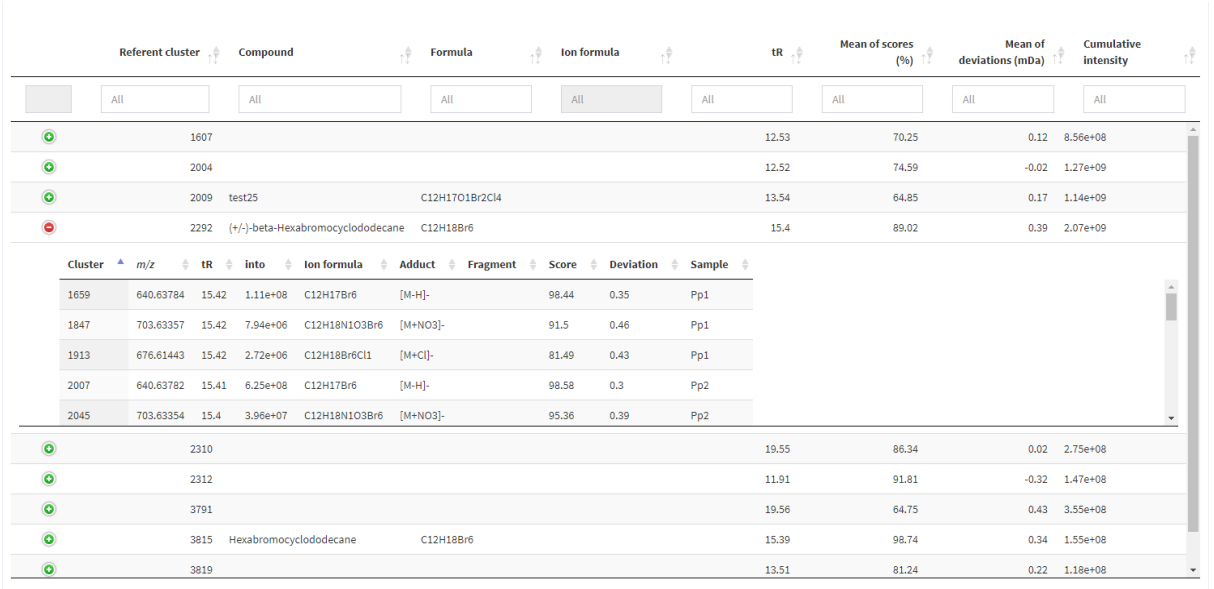


Figure 6 : Interactive plot toolbars.

3.5. TABLES

Most of the tables in the application involve filters on the top for each column and allow multi-column ordering by shift click (**Figure 7**). The table in the *Enumeration* tab has a button on the first column allowing to display another sub-table under the row with detailed information.



	Referent cluster	Compound	Formula	Ion formula	tR	Mean of scores (%)	Mean of deviations (mDa)	Cumulative intensity
	All	All	All	All	All	All	All	All
⊕	1607				12.53	70.25	0.12	8.56e+08
⊕	2004				12.52	74.59	-0.02	1.27e+09
⊕	2009	test25	C12H17O1Br2Cl4		13.54	64.85	0.17	1.14e+09
⊖	2292	(+/-)-beta-Hexabromocyclododecane	C12H18Br6		15.4	89.02	0.39	2.07e+09

Cluster	m/z	tR	into	Ion formula	Adduct	Fragment	Score	Deviation	Sample
1659	640.63784	15.42	1.11e+08	C12H17Br6	[M-H]-		98.44	0.35	Pp1
1847	703.63357	15.42	7.94e+06	C12H18N1O3Br6	[M+NO3]-		91.5	0.46	Pp1
1913	676.61443	15.42	2.72e+06	C12H18Br6Cl1	[M+Cl]-		81.49	0.43	Pp1
2007	640.63782	15.41	6.25e+08	C12H17Br6	[M-H]-		98.58	0.3	Pp2
2045	703.63354	15.4	3.96e+07	C12H18N1O3Br6	[M+NO3]-		95.36	0.39	Pp2

⊕	2310				19.55	86.34	0.02	2.75e+08
⊕	2312				11.91	91.81	-0.32	1.47e+08
⊕	3791				19.56	64.75	0.43	3.55e+08
⊕	3815	Hexabromocyclododecane	C12H18Br6		15.39	98.74	0.34	1.55e+08
⊕	3819				13.51	81.24	0.22	1.18e+08

Figure 7 : Table filters.

3.6. WORKFLOW OVERVIEW

Once a project is created, the workflow (**Figure 8**) first consists in importing LC/HRMS or GC/HRMS data files in the application and converting them in a universal format (*mzXML*).

Secondly, a peak-picking step is applied to detect and integrate all chromatographically resolved signals as a list of so-called "features" (peaks) characterized by a *m*/z, a retention time and an intensity.

The third step corresponds to the halogen pairing. It allows clusterizing features that represent isotopologues, *i.e.* belong to the same halogenated ion (see § 5.1.2).

Then, the clusters obtained are paired according to correlation analysis based on coelution (Gaussian chromatographic peaks) for deconvoluting adduct and fragment clusters supposedly belonging to the same compound.

Further deconvolution is operated by alignment of these fragment/adduct groups (retention time and *m*/z) from one file to another in so-called SuperGroups of clusters, supposedly belonging to the same compound all along the sequence. It corrects for slight retention time drifts within the sequence.

The core of the software consists in investigating and annotating those SuperGroups with a formula or a compound after user reviewing through an interactive plot and a formula decomposition modal.

Eventually, all processed information and annotations can be exported as an Excel sheet.

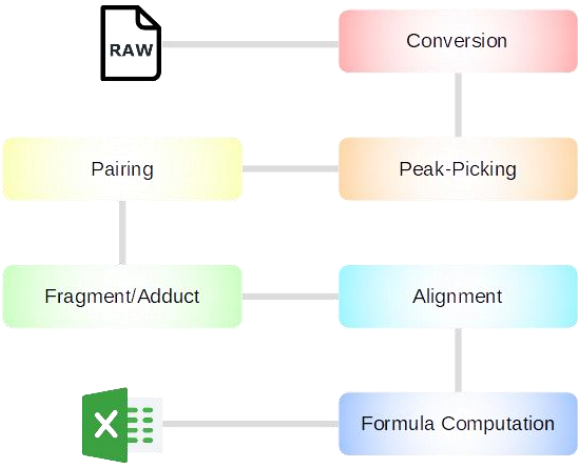


Figure 8 : Workflow overview.

4. TAB PROJECTS & FILES

4.1. SUB-TAB NEW PROJECT

4.1.1. Create a new user

A new user identification can be created using the dedicated sub-tab "*New Project*" (**Figure 9**). Do not forget to click the "*Create*" button. Once created, it is added to the *ad hoc* Active user list of the Title bar. Thus, do not forget to select it as Active user. The active user is recorded for each annotation.

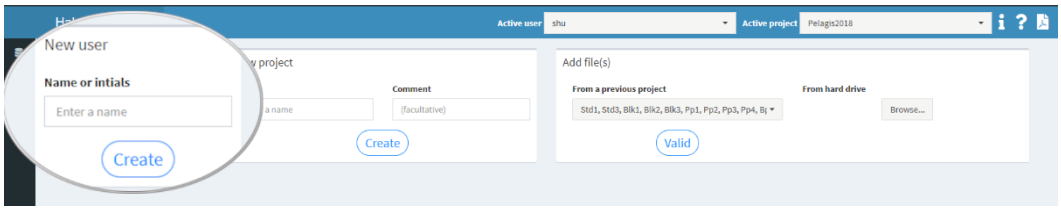


Figure 9 : Creation of a new user.

4.1.2. Create a new project

A new project name can be created using the dedicated sub-tab "*New Project*" (**Figure 10**). A comment can be provided as well. Do not forget to click the "*Create*" button. Once created, it is added to the *ad hoc* Active project list of the Title bar. Thus, do not forget to select it as Active project.

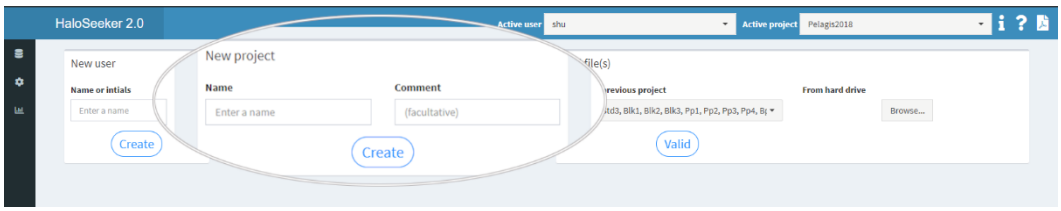


Figure 10 : Creation of a new project.

4.1.3. Import files to a project

File can be imported to the Active project either from the hard drive and/or among other projects (**Figure 11**).

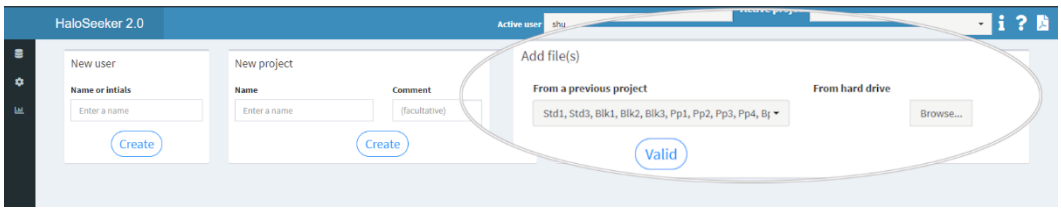


Figure 11 : Import of HRMS files into a project.

4.1.3.1. From hard drive

Click on the "*Browse...*" button. It will display a modal allowing to navigate through the file system (**Figure 12**). Some upper buttons facilitate the navigation by changing the layout or sorting out files by their name, size and creation date. Major manufacturer and open formats are handled in the application (**.raw*, **.RAW*, **.d*, **.YEP*, **.BAF*, **.FID*, **.WIFF*, *.CDF, *.MGF, non-centroided **.mzXML*, **.mzML*, **.MGF*). Waters and Bruker raw directories can also be imported (they are considered as files by HaloSeeker if they have extension-like names). Once files are selected (colored after a click), do not forget to click the "*Select*" button.

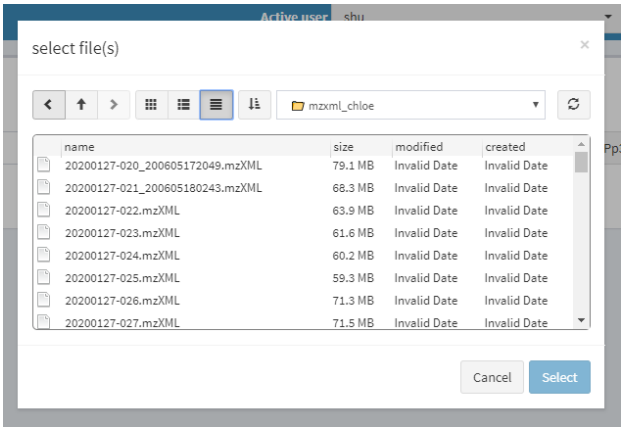


Figure 12 : Selection of HRMS files to be imported into a project from a hard drive.

Once the files are selected, a second modal allows you to choose the same polarity of the scans for all the files (**Figure 13**). It is also convenient to provide a label (sample ID) different from the file name. This label is used for selecting files when reviewing the data and results. Be aware that files can be labelled with the same label, which could be confusing. Otherwise, the default label is <polarity><filename>. Do not forget to click on the "*Valid*" button.

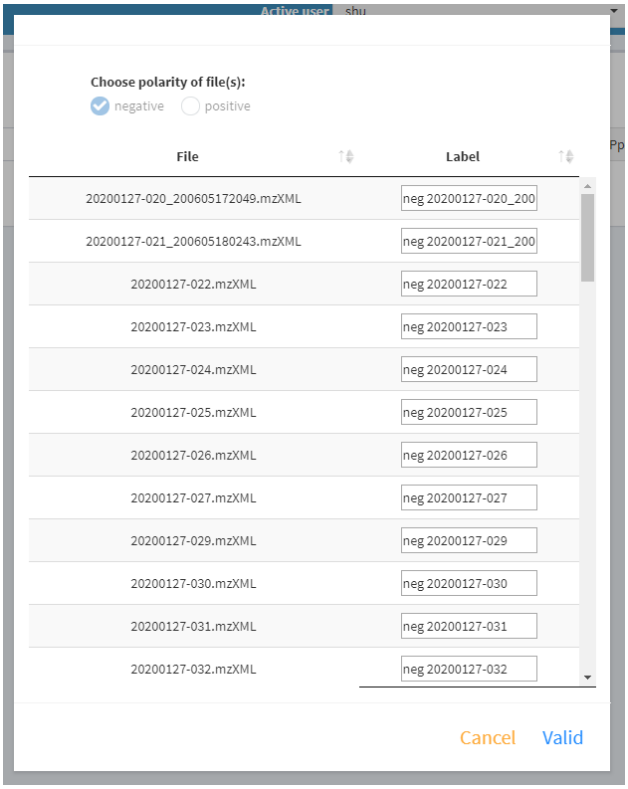


Figure 13 : Choice of polarity and HRMS file labels.

Once validated, the files are automatically converted in an *mzXML* format in the defined polarity, using *msConvert* version 3.0.9810, a software produced by *ProteoWizard* (Chambers et al., 2012). Some Microsoft files are mandatory to make it work. In case you miss them, please refer to the FAQ section.

In case of dual Pos/Neg files, HaloSeeker considers the trace corresponding to the defined polarity. To process both polarities, it is advised to create two distinct projects.

The data are also automatically centroided, in order to "lighten" data. The user is also allowed to import already centroided file(s) in the most common format (**.mzXML*, **.mzML*, **.CDF*). These files are checked for centroided and readable data. Import of non-centroided *mzXML* files does not work.

Why centroid MS data? Because a peak in profile mode contains too much information that the software cannot use for the moment (and it will be too hard to work with). Thus, each peak will be displayed as discrete *m/z* with zero line widths (**Figure 14**). For that, *msConvert* includes most of the manufacturers' algorithms, except for Waters Corporation. In this case, *ProteoWizard* developed their own algorithm and, depending on the instrument, the script might fail. In this case, it is advised that the user converts such files independently and then imports converted files into HaloSeeker. Do not hesitate to contact us for tips.

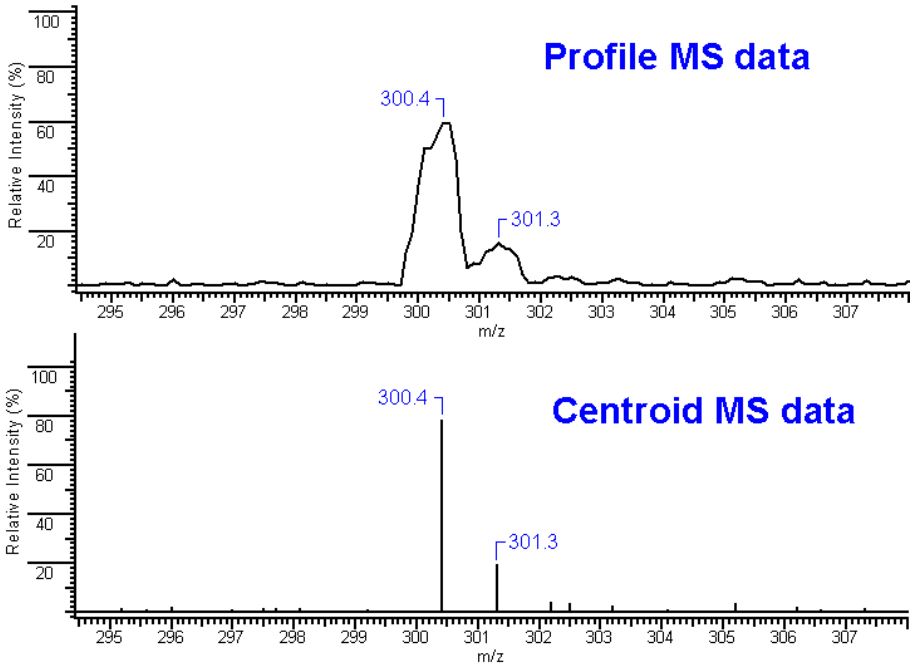


Figure 14 : Illustration of profile *versus* centroid MS data.

4.1.3.2. From other projects

All data files imported within the application can be re-used in multiple projects. Using the picklist showing files sorted by projects (**Figure 15**), it is possible to select previously imported, converted and centroided HRMS data files. Do not forget to click on the "*Valid*" button. Then, the modal allowing to provide a label (sample ID) different from the file name appears (see §4.1.3.1). Do not forget to click on the "*Valid*" button.

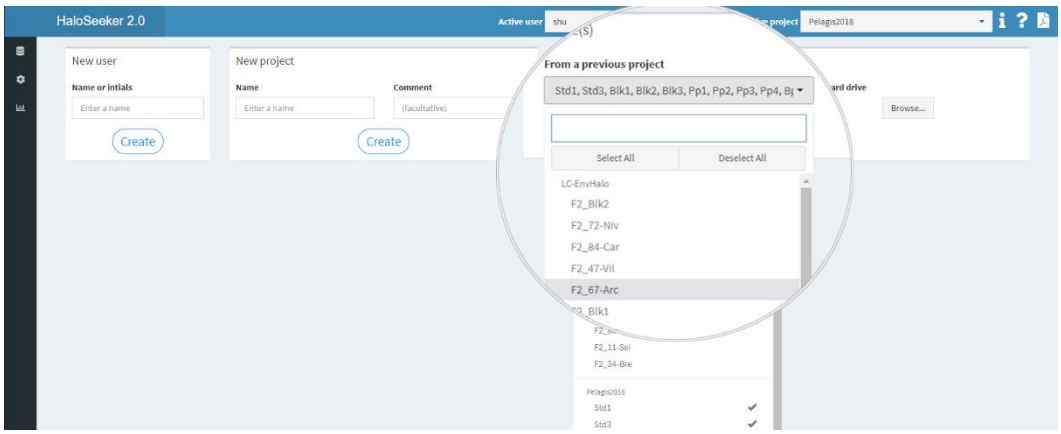


Figure 15 : Import HRMS files from other projects.

4.2. SUB-TAB DATABASE TABLES

The "*Database tables*" sub-tab of the "*Projects & Files*" tab (**Figure 16**) allows to manage entries stored in the HaloSeeker 2.0 database (*SQLite* format). The head area (**1**) allows to select one of the 4 accessible tables (Project, Sample, Compound or Adduct), which are presented on the bottom area (**3**). Two actions are possible. First, deleting entries is made by selecting one or multiple rows in the table (**3**), and then by clicking on the button "*Delete entries*" (**2**). Secondly, creating an entry using the button "*Add new entry*" (**2**) depends on the selected table. Also some tables allow specific actions.

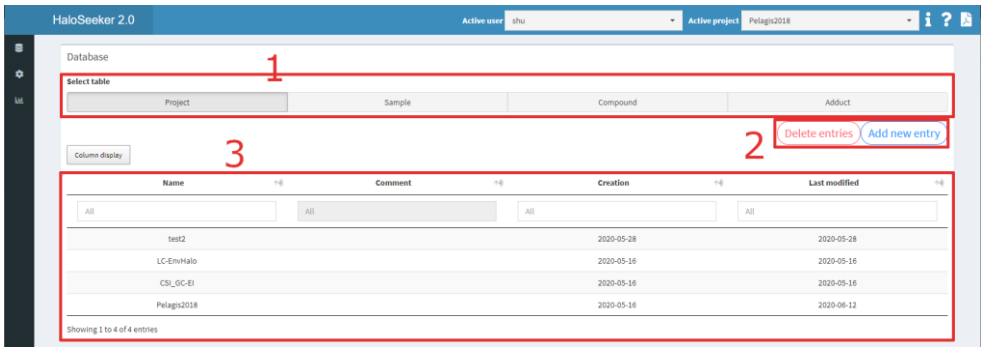


Figure 16 : "*Database tables*" sub-tab.

The *SQLite* database (database.sqlite file) is available at the root directory of the installed software. It is possible (and advised) to archive it elsewhere once a project is over.

4.2.1. Project

Each line represents a project stored in the application database (**Figure 17**).

- Name: name of the project (§ 4.1.2). A double click on a cell allows to rename it.
- Comment: information about the project (optional area, § 4.1.2).
- Creation: creation date of the project.
- Last modified: date of the last modification operated on the project.

Name	Comment	Creation	Last modified
All	All	All	All
test2		2020-05-28	2020-05-28
LC-EnvHalo		2020-05-16	2020-05-16
CSI_GC-EI		2020-05-16	2020-05-16
Pelagis2018		2020-05-16	2020-06-12

Showing 1 to 4 of 4 entries

Figure 17 : Database "*Project*" table.

The button "*Add new entry*" returns directly to the sub-tab "*New project*" of the tab "*Projects & Files*" (§ 4.1.2).

4.2.2. Sample

Each line represents a file imported in a project (**Figure 18**). The same file can be present in more than one row if it associated to more than one project.

- Sample: label of the file imported (**Figure 13**). A double click allows to rename it.
- Project: name of the project associated to the file (§ 4.1.2).
- Polarity: scan polarity of the file (§ 4.1.3.1).
- Size (Mo): actual size in Mo of the compressed & stored *mzXML* file in the database. If the compressed file is above 2 Go, it is stored outside the SQLite database.

In addition, a range of metadata can be retrieved from some Thermo Scientific raw files using the *ThermoRawMetaDump* tool.

- Instrument model.
- Instrument manufacturer.
- Ion source.
- Analyzer.
- Detector type.
- Resolution.
- AGC target.
- Maximum IT.
- Number of scan range.
- Scan range.

Lastly, the original path of the imported file is given.

Sample	Project	Polarity	Size (Mo)	Instrument model	Instrument manufacturer	Ion source	Analyzer	Detector type	Resolution	AGC target	Maximum IT	Number of scan range	Scan range
				All	All				All		All		
Std1	Pelagis2018	negative	34	Q Exactive	Thermo Scientific	electrospray ionization	quadrupole	inductive detector	140,000	5e5	250 ms	1	120 to 1000 m/z
Std3	Pelagis2018	negative	34	Q Exactive	Thermo Scientific	electrospray ionization	quadrupole	inductive detector	140,000	5e5	250 ms	1	120 to 1000 m/z
Blk1	Pelagis2018	negative	34	Q Exactive	Thermo Scientific	electrospray ionization	quadrupole	inductive detector	140,000	5e5	250 ms	1	120 to 1000 m/z
Blk2	Pelagis2018	negative	34	Q Exactive	Thermo Scientific	electrospray ionization	quadrupole	inductive detector	140,000	5e5	250 ms	1	120 to 1000 m/z
Blk3	Pelagis2018	negative	34	Q Exactive	Thermo Scientific	electrospray ionization	quadrupole	inductive detector	140,000	5e5	250 ms	1	120 to 1000 m/z

Figure 18 : Database "*Sample*" table.

The button "*Add new entry*" returns directly to the sub-tab "*New project*" of the tab "*Projects & Files*" (§ 4.1.2).

4.2.3. Compound

Each line represents a compound of the chemical formula database (**Figure 19**).

- Formula: chemical formula.
- m/z: *m/z* of the base peak (the most intense isotopic peak).
- Compound: name of the compound.

Formula	m/z	Compound
All		All
C10H7N3O3S1	249.02081	((((4-Methylphenyl)sulphonyl)oxy)imino)malononitrile
C13H24N4O1	252.19501	(((2-((2-Aminoethyl)amino)ethyl)amino)ethyl)amino)methyl)phenol
C16H12O3S2	316.02279	(((9-Oxo-9H-thioxanthen-2-yl)methyl)thio)acetic acid
C7H13N1O2S2	207.03877	(((Diethylamino)thioxomethyl)thio)acetic acid
C10H14O4	198.08921	(((Phenylmethoxy)methoxy)methoxy)methanol

Figure 19 : Database "*Compound*" table.

HaloSeeker 2.0 is released with an empty compound database. The user can populate the database compound by compound through the annotation modal (§ 7.7.4) and/or import his/her own database as follows.

The button "*Add new entry*" leads to a modal with three areas (**Figure 20**). (1) The left button downloads an Excel® template with 3 columns to be filled: the chemical formula and the name. It is possible to add two compounds with the same formula but with different names (isomers or same compound). Thus, it is advised to check if the desired compound already exists in the database under another name. (2) Upload the filled templates to import them in the application. (3) The table on the right shows up at the end of the import, each line representing an error with the line number and the reason of the failure.

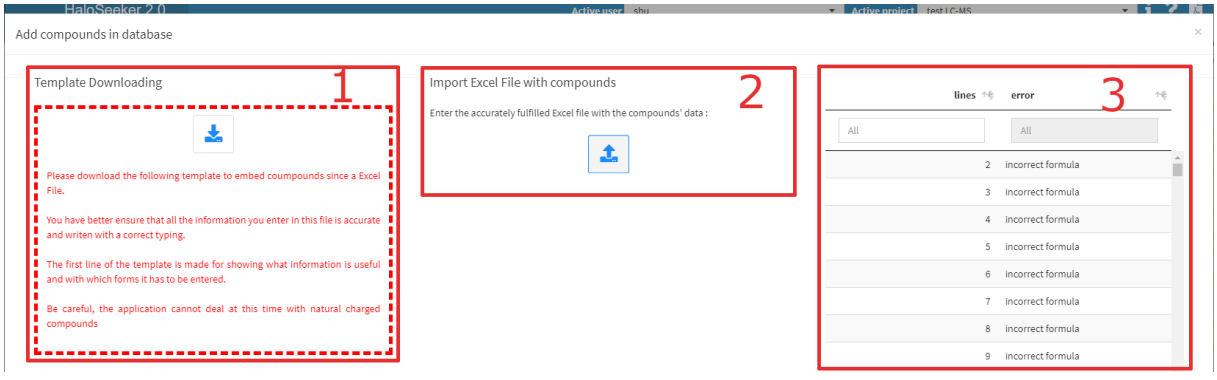


Figure 20 : Importing compounds in HaloSeeker 2.0 database.

4.2.4. Adduct

Each line represents an adduct type. By default, the application already stores all the adduct types of the enviPat 8 and CAMERA *R* package.

- Adduct: name of the adduct. Besides chemical formulas, some abbreviations are used: ACN = acetonitrile; FA = formic acid; Hac = acetic acid; IsoPro = isopropanol; TFA = trifluoroacetic acid.
- Multi: number of repeated units (dimer/trimer specification).
- Charge: charge.
- Mass: mass difference between the adduct ion and the molecular neutral mass.

Adduct	Multi	Charge	Mass
All	All	All	All
[M-H2O-H]-	1	-1	-48.992020312
[M-H-Cl+O]-	1	-1	-19.981214542
[M-Cl+O]-	1	-1	-18.97338951
[M-H]-	1	-1	-1.007276452
[2M-H]-	2	-1	-1.007276452
[3M-H]-	3	-1	-1.007276452
[M]+	1	1	-0.00054858
[M]-	1	-1	0.00054858
[M+H]+	1	1	1.007276452
[2M+H]+	2	1	1.007276452
[3M+H]+	3	1	1.007276452
[M-2H+NH4]-	1	-1	16.019272654
[2M-2H+NH4]-	2	-1	16.019272654
[3M-2H+NH4]-	3	-1	16.019272654

Showing 1 to 65 of 65 entries

Figure 21 : Database "*Adduct*" table.

The button "*Add new entry*" leads to a modal with a form to fill (**Figure 22**). For adding adducts, three arguments are required: name, charge and formula. The formula input must be of the type "*M+Xx-Yy*", using a decomposition generator detailing the atom, its number and a sign for its addition or removal. For example, if the user wants to add the "[M+Na-H_2]-" adduct, it can be done in two steps:

- Enter "1" in the field "*Multi*" (or leave it empty), select the "+" operator and enter "*Na*" in the "*formula*" field. Click on the "+" button to confirm the addition of 1 atom of Na.
- Enter "2" in the field "*Multi*", select the "-" operator and enter "*H*" in the "*formula*" field. Click on the "+" button to confirm the addition of 2 atoms of H.

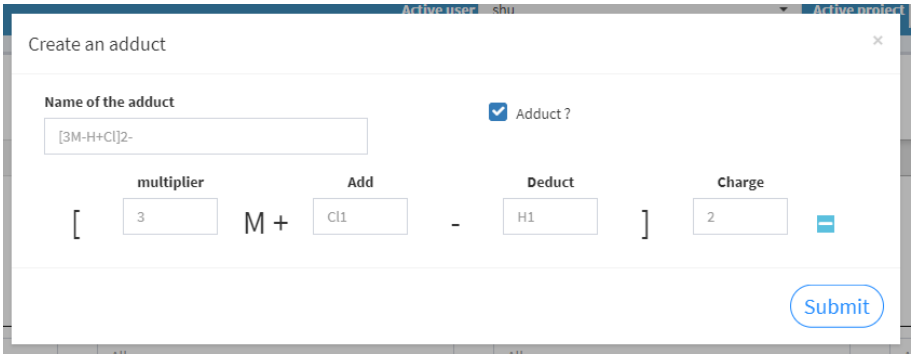


Figure 22 : Modal for the creation of an adduct entry in the HaloSeeker 2.0 database.

It is advised to create adduct types related to labeled standard compounds, particularly deuterated standards analyzed by atmospheric pressure ionization such as electrospray (ESI). Labeled isotopes are noted with square brackets ([13]C, [2]H, [15]N). For example, one may create the adduct type $[M - D]^-$ or $[M - [2]H]^-$ to identify d_{18}-β-hexabromocyclododecane.

5. TAB DECONVOLUTION

5.1. SUB-TAB PROCESS

The "*Process*" sub-tab of the "*Deconvolution*" tab (**Figure 23**) gathers the parameters for the four deconvolution scripts of the workflow in the left area (**1**), which are the peak-picking, the Cl/Br pairing, the intra-file grouping of fragments & adducts and the inter-file alignment. While the two first ones are mandatory, the last two ones remain optional. It is advised to activate all of them to use the application at its maximum potential. By default, the four scripts are activated, but it is possible to unclick the third and/or the fourth one.

The second area (**2**) shows overlapped Total Ion Currents (TICs) of all samples in the project. In order to minimize time to display the chromatograms, data are smoothed to a maximum of 120 points per minute. Still, depending on the number of files, it takes a few dozens of seconds to be displayed, so be patient. A click on the TICs displays the mass spectra (**3**) corresponding to the selected retention time (t_R) coordinate.

Once the parameters are defined, do not forget to click on the "*Launch deconvolution process*" button. The selected deconvolution steps will then be processed at once, including the peak-picking, the pairing and the fragment/adduct search for each sample file successively, and finally the alignment. This processing might take a while (a few dozens of minutes), so be patient. The screen refreshing might not be working at the end of the process, so do not hesitate to move the mouse across the screen window to make it happen. Once finished, a modal stating the success of the process appears. After validation of this modal, the sub-tab "*Result*" is automatically shown (§ 5.2).

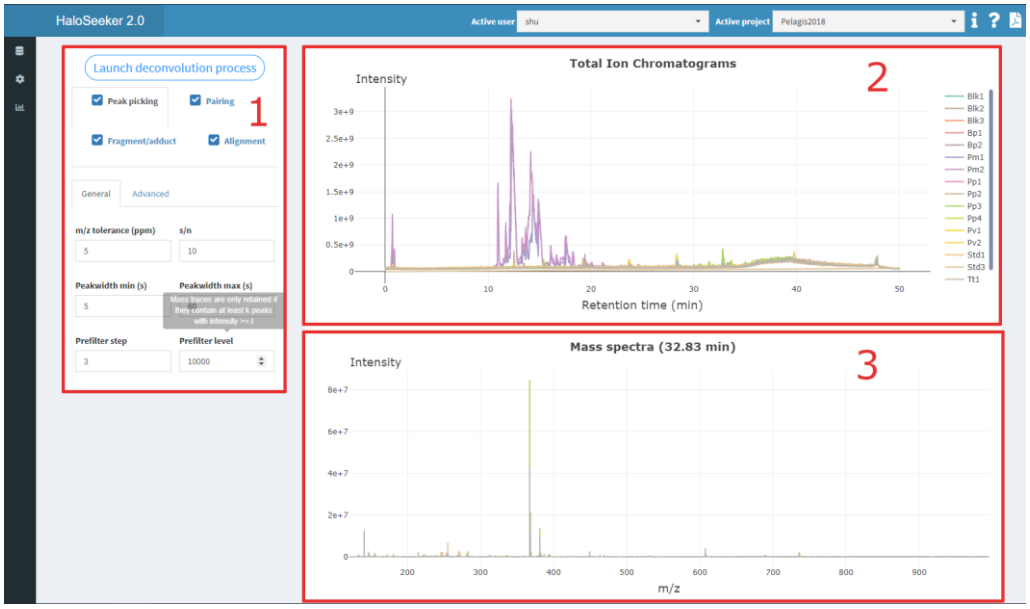


Figure 23 : "*Process*" sub-tab.

5.1.1. Peak-picking

The peak-picking step (or peak identification step) is mandatory. It permits the detection of chromatographic peaks in the project files. For each extracted signal (so-called "feature"), it provides a centroid *m/z*, a retention time and an intensity. This process is performed thanks to the *centWave* algorithm (Tautenhahn et al., 2008) of the *xcms* package (version 3.2.0) without the "*predict isotope ROIs*" functionality. The *centWave* parameters can be adjusted by the user.

About the process itself, *xcms* first searches for so-called "*m/z* Regions Of Interest" (ROI) which correspond to a *m/z* detected along several scans with a ppm tolerance (**Figure 24**). The parameters used for this step are a *m/z* tolerance (ppm), a prefilter step and a prefilter level. The Prefilter step represents the number of mass traces with a *m/z* tolerance in consecutive scans and prefilter level an intensity threshold. Once done, wavelets transformation are used over the ROIs to find peaks (**Figure 24**). Several passes of wavelets are used until the best "fit" is found (Mexican hat wavelets). Two parameters can be set. Peakwidth corresponds to the retention time window to use (minimum and maximum). Snthresh corresponds to a signal to noise ratio cutoff. Many other parameters can be found in the "*Advanced*" tab box. Please refer to the *xcms* package documentation.

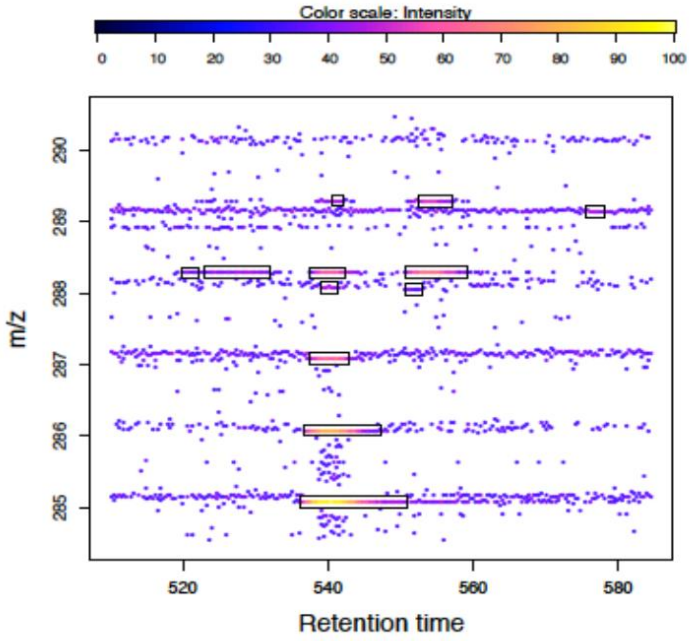


Figure 24 : Region Of Interest (ROI) detection for peak-picking (Tautenhahn et al., 2008).

5.1.2. Pairing

The pairing step (or deconvolution step) is mandatory, as it stands for the core idea of the deconvolution approach of HaloSeeker. It is like a clustering step based on precise *m*/*z* differences between features (output of the peak-picking step). It aims at pairing features corresponding to isotopologue ions ($^{12}C/^{13}C$, $^{35}Cl/^{37}Cl$, $^{79}Br/^{81}Br$) in each project file, based on a retention time tolerance (s) and a *m*/*z* tolerance (mDa) as the two parameters to be defined by the user as regards the instrumental performances.

The pairing script was adapted from Cariou et al. (2016), pending slight modifications, as described in Léon et al. (2019). The list of features is sorted in the decreasing order of intensities. The main loop of the script considers successively each feature as a base peak "A" (the most intense within the isotopic pattern) if not already paired to a more intense feature.

For each base peak "A", the secondary loop of the script tests the occurrence of two less intense features meeting retention time and *m*/*z* tolerances, starting from the most intense feature (**Figure 25**). First, it tests an A+1 feature at a difference of *m*/*z* 1.003355 ± tolerance for the substitution of a ^{12}C by a ^{13}C. Secondly, it tests an A+2 feature at a difference between *m*/*z* 1.99704992 minus tolerance for the substitution of a ^{35}Cl by a ^{37}Cl and *m*/*z* 1.9979521 plus tolerance for the substitution of a ^{79}Br by a ^{81}Br. Indeed, a halogenated ion always exhibits at least A and A+2 peaks, A being the most intense.

In case of an A+2 feature, the application then seeks other isotopologue features at the left side of the A feature (for A-2) and at the right side of the A+2 (for A+4 and A+3). The *m*/*z* difference is always considered from de base peak A, standing as the major modification after Cariou et al., 2016. The secondary loop continues with A-n-2/A-n+1 and/or A+n+2/A+n+1 as long as A±n features are paired (n being an even number), up to A+12 and A-12. At the end of the secondary loop, the matched isotopologue features are tagged as a so-called paired cluster if there is at least an A+2 feature.

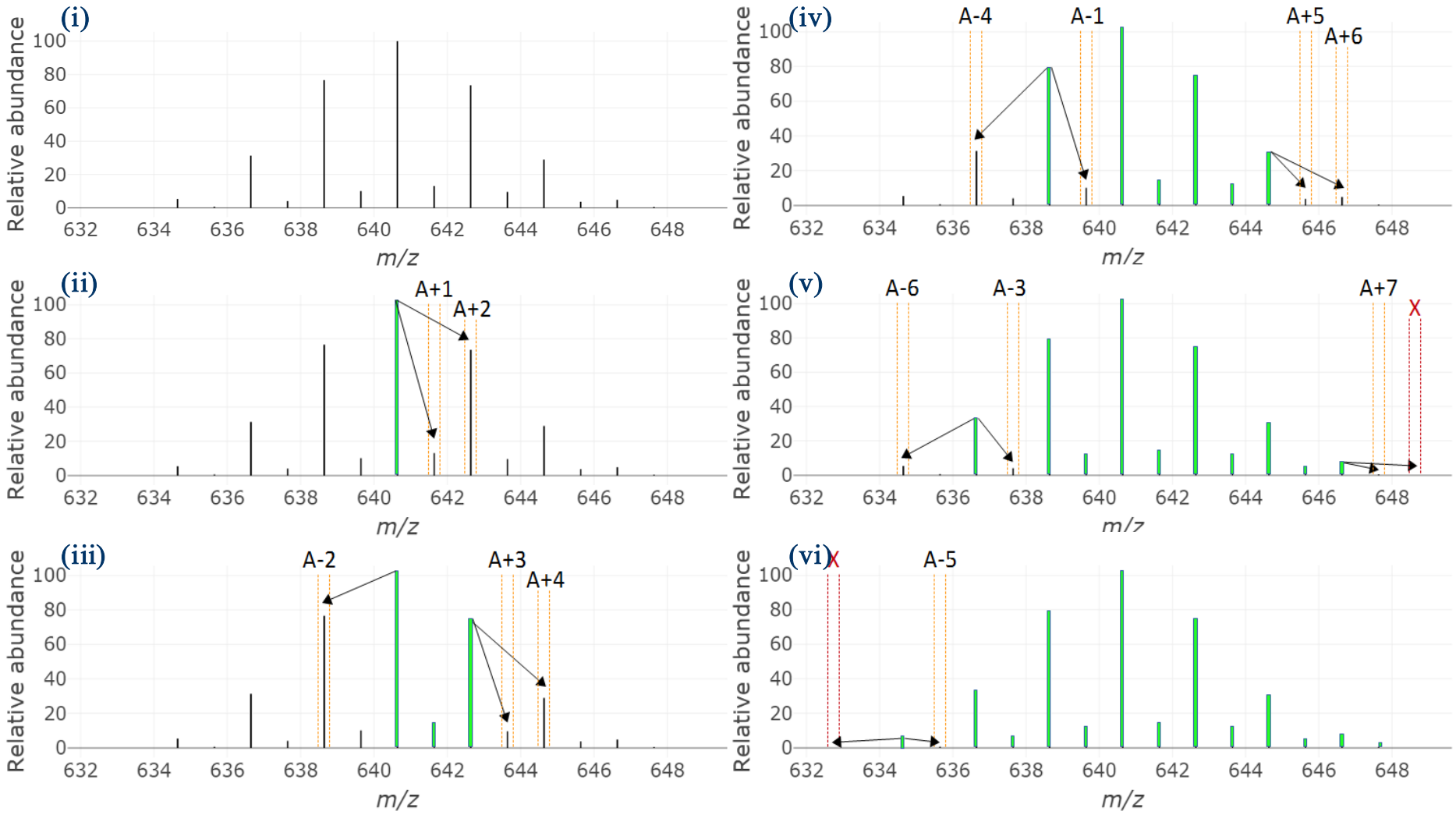


Figure 25 : Pairing script secondary loop. Example on theoretical MS profiles of $C_{12}H_{18}Br_6$ $[M-H]^-$ ion, from (i) to (vi).

5.1.3. Fragment/adduct

The fragment/adduct step is optional and can be deactivated by clicking on the tick box. It aims at pairing clusters (output of the pairing step) arising from the same molecule during the in-source ionization. Necessarily, these fragments and/or adduct ions show theoretically identical relative chromatograms. The script is applied to each data file (intra-file grouping) independently, using the *CAMERA* (Kuhl et al., 2012) *R* package, in which two actions are processed, *i.e.*:

- Group all peaks which eluate at the same time-point, defined as the retention time ± full width at half maximum (FWHM). This FWHM of a peak is defined as the standard deviation * 2.35 (the peak is assumed as normally distributed).
- Partition those groups according to their chromatographic peak shape correlations. It performs all extracted ion chromatogram (EIC) correlations between peaks in one group inside a sample and separates them with a graph clustering method (Highly Connected Subgraphs approach or Label Propagation Community Algorithm).

The original CAMERA script has been adjusted for HaloSeeker needs. Indeed, only the sub-set of base peaks A is considered, as being the best representatives of each cluster. This restriction reduces the number of false positives.

Also, the possibility offered by *CAMERA* to annotate adduct ions and identify the molecular mass is not considered by HaloSeeker, this information being determined later during the formula decomposition process.

5.1.4. Alignment

The alignment step is optional and can be deactivated by clicking on the tick box. It aims at pairing clusters (output of the pairing step) matching from one file to another (inter-file grouping).

In a first step, retention time adjustment is performed using the OBI-Warp algorithm (**Figure 26**) (Prince and Marcotte, 2006).

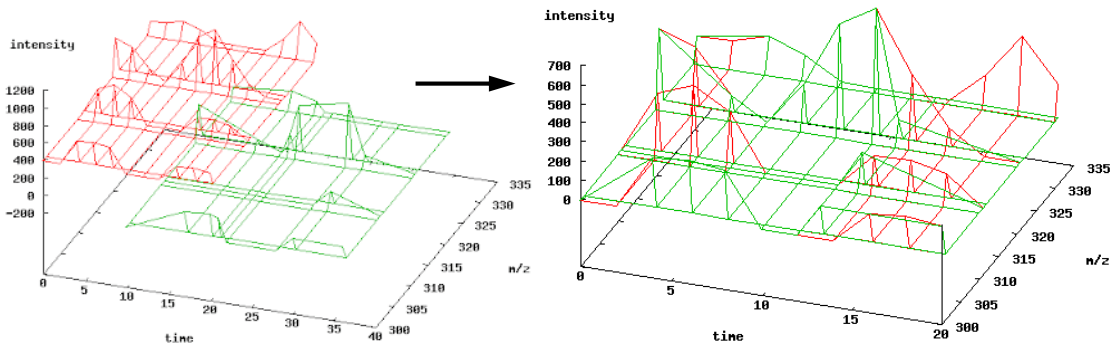


Figure 26 : Representation of the OBI-Warp algorithm (http://obi-warp.sourceforge.net/tutorials/mass_spectrometry_alignment_example.html)

In a second step, the density distribution of the identified peaks is calculated along the retention time axis and peaks from the same or different samples that are close to each other are grouped. The original algorithm has been adjusted for HaloSeeker needs. Indeed, only the sub-set of triplets A-2 / A / A+2 is considered, as being the best representatives of each cluster. This restriction reduces the number of false positives. A triplet of features was preferred to the sub-set of base peaks A in order to address experimental switching between theoretical base peak A and relatively intense A-2 or A+2 isotopologues, misleading the correct annotation of the base peak.

5.2. SUB-TAB RESULT

The "*Result*" sub-tab of the "*Deconvolution*" tab provides information on the deconvolution process result in terms of number of features/clusters in each sample file, according to halogen status (table head, see § 6.1.4.2) and deconvolution step (columns) (**Figure 27**). The last row provides enumeration to the entire set of project files. Each column provides the enumeration of clusters (or features if clusters not applicable).

- Label: sample file label as defined (**Figure 13**).
- Peak Picking (features): number of chromatographic peaks integrated.
- Pairing (clusters): number of paired halogenated isotopic clusters (output of the pairing step).
- Fragment/adduct (groups): number of groups in which the paired clusters were deconvoluted after the grouping steps (output of the fragment/adduct step).
- Alignment (groups): number of paired clusters aligned with clusters from other sample files (output of the alignment step).
- SuperGroups: number of SuperGroups in which the paired clusters were deconvoluted. A SuperGroup gathers clusters sharing a same fragment/adduct ID and/or alignment ID with one another (nearest neighbors).

F0 | F1 | F2 | F2+

label	Peak Picking (features)	Pairing (clusters)	Fragment/adduct (groups)	Alignment (groups)	SuperGroups
Blk1	671	270	221	123	198
Blk2	676	278	235	137	214
Blk3	274	91	60	80	41
Bp1	415	140	104	115	82
Bp2	370	124	92	104	65
Pm1	342	111	85	74	63
Pm2	300	95	72	54	55
Pp1	1077	333	236	182	210
Pp2	1058	256	158	175	130
Pp3	1345	302	184	170	157
Pp4	1086	343	252	184	230
Pv1	416	120	86	108	62
Pv2	446	129	82	108	60
Std1	923	352	267	135	246
Std3	700	236	164	124	108
Tt1	816	210	130	167	98
Total 17 files	**12073**	**3670**	**2591**	**556**	**1423**

Figure 27 : "*Result*" sub-tab.

6. TAB EXPLORE DATA

6.1. SUB-TAB INTERACTIVE PLOT

6.1.1. Overview

The "*Interactive plot*" sub-tab of the "*Explore data*" tab stands for the ergonomic core of the software (**Figure 28**). Multiple parameters and filters can be adjusted by the user, separated in 3 panels (**1**). A table displays some enumeration providing information on the efficiency of selected filters (**2**) The visualization of all features is displayed in an interactive plot (**3**). Do not forget to click on the "*Draw*" button to update the table and the plot according to the selected parameters.

The displayed data can be exported using the button  as an Excel® file containing the deconvolution and display parameters used in a first sheet and the entire data table of the displayed data in a second sheet.

On mouse over, the following information on the selected feature is provided: Label, cluster ID, base peak A *m/z* (if "display *m/z*" activated in "*Axes & Legends*" options, § 6.1.2), *m/z*, retention time (t_R), cumulative intensity of the cluster, isotopologue, and, if annotated, compound name, neutral formula, adduct type or fragment formula, ion formula, patter score, mass accuracy (deviation mDa).

By a click on a feature in the plot, a modal dedicated to the investigating and annotation of the corresponding SuperGroup appears (see § 7).

Figure 28 : Interactive plot window.

6.1.2. Axes & Legends

The option panel "*Axes & Legends*" offers the customization of the axes and the legends by providing the choice of the X-axis and Y-axis by a list pre-defined types of values (IUPAC *m/z*, H/Cl-scale *m/z*, retention time, H/Cl-scale *m/z* fractional part, IUPAC fractional part, intensity) (**Figure 29**). Legend entries can also be sorted by their intensities, order assignment during the pairing process or by increasing *m/z* of the base peak. Also, two options allow the user to display *m/z* of base peak in the plot legend and formula as text labels on the plot.

By default, the application displays an H/Cl scale mass defect plot (MD-plot). A MD-plot represents the fractional part of a *m/z* according to the *m/z* value. Here, feature *m/z* is converted from the IUPAC scale (mass of ^{12}C = 12.0000 uma) in the H/Cl scale (mass of $-^{1}H+^{35}Cl$ = 34.0000 uma) (Taguchi et al., 2010). In such a scale, while alkane series align on an increasing diagonal, monoisotopic dots corresponding to chlorinated compound series align on a horizontal line (*e.g.* polychlorinated biphenyl homologue groups). Conveniently, it also works for brominated homologue series.

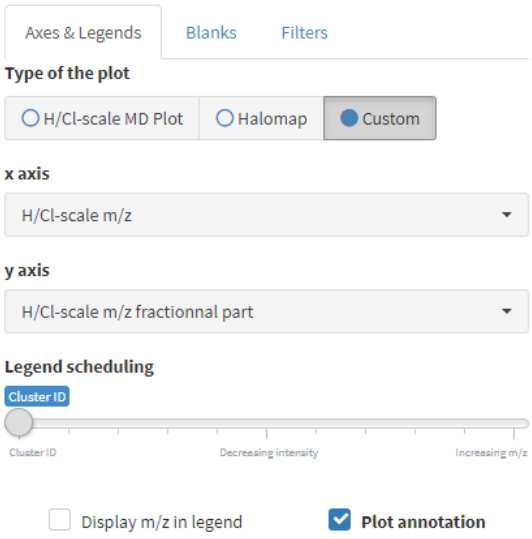


Figure 29 : Display option parameters related to axes and legends.

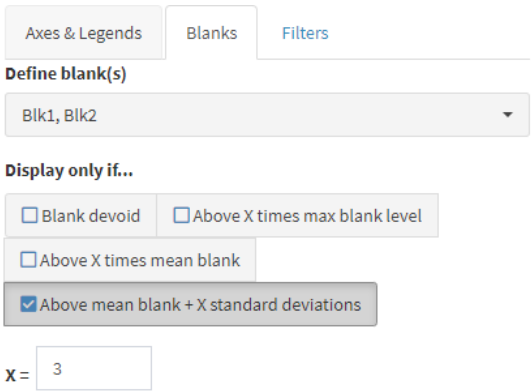


Figure 30 : Display option parameters related to blank comparison.

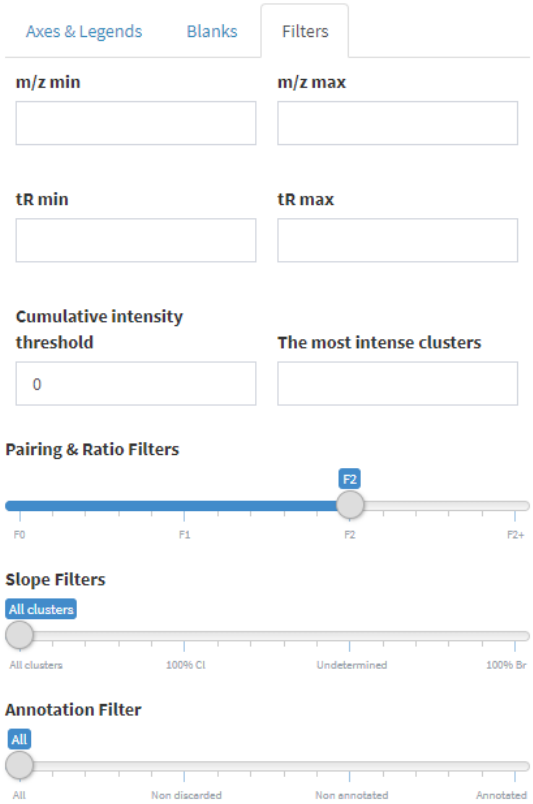


Figure 31 : Display option parameters related to various filters.

6.1.3. Blanks

The option panel "*Blanks*" is a new functionality of HaloSeeker 2.0. It offers the possibility to remove clusters from the displayed plot according to adjustable rules related to differences between selected samples and other samples defined as procedural blanks. Thus, no subtraction if operated: only a comparison of cumulative intensities with a consequence on the displayed clusters. This functionality is possible only if the alignment was processed at the deconvolution step. Indeed, the comparison is performed on cumulative intensities of aligned clusters within a SuperGroup.

Four strategies can be applied after selection of the blank file(s) within the sequence (**Figure 30**):

- "*Blank devoid*": any cluster aligned with a cluster present in a blank is excluded from display.
- "*Above X times max blank level*": any cluster below X times the most intense aligned cluster among the blanks is excluded from display.
- "*Above X times mean blank*": any cluster below X times the average value of aligned cluster among the blanks is excluded from display. The value "0" is considered for blanks with no aligned cluster.
- "*Above mean blank + X standard deviations*": any cluster below the average value of aligned cluster among the blanks plus X times their standards deviation is excluded from display. The value "0" is considered for blanks with no aligned cluster.

6.1.4. Filters

The option panel "*Filters*" displays various options for filtering the data in order to prioritize the annotation (similar to HaloSeeker 1.0). Filters are available for a *m*/*z* range, a retention time range, an intensity threshold (cumulative for paired clusters), and a maximum number of features or paired clusters to be displayed (sorted by their intensities) (**Figure 31**). Three filter sliders are also available for paired clusters, according to their annotation during the deconvolution process, to the cluster slope of the linear regression curve and to their annotation status (discarded, already annotated with a formula and/or a compound).

6.1.4.1. m/z and retention time range

The *m*/*z* and retention time (t_R) range filters apply on the entire cluster considering the base peak A only (to the single feature if not applicable).

6.1.4.2. Pairing and ratio annotation status of the paired clusters

"*Paired clusters*" encompass all paired clusters containing at least an A+2 feature. Complementarily, paired clusters are annotated according to ion ratio rules, aiming at improving filtering of truly halogenated ions.

F0 filter displays all peak-picked features ("*all features*") and F1 filters the "*m/z–paired*" clusters (containing at least an A+2 isotopologue feature). F2 and F2+ filters display clusters complying with additional rules to F1, related to ion ratios in order to improve filtering of halogenated and polyhalogenated clusters, respectively. Based on theoretical considerations, isotopologue area relative to base peak (A) rules are as follows for F2 filter ("*ion ratio*").

- A-2 ratio = 0 (absence) AND A+2 ratio ≥ 25% OR;
- A-2 ratio ≥ 60% AND A+2 ratio ≥ 20% OR;
- A-2 ratio ≥ 27% AND A+2 ratio ≥ 36%.

For F2+ filter ("*polyhalo*"), F2 clusters suspected to be related to monohalogenated ions are removed according to the following rules:

- A-2 and A+4 ratios = 0 (absences) AND;
- A+2 ratio ∈ [25-39] ∪ [77-117] %.

The F2 rules were derived from theoretical patterns of Mixed Cl and Br formulas up to 6 atoms. Intensities of isotopologue of similar nominal masses were cumulated (**Table 2**). Relative intensities of A-2 and A+4 signals over base peak signal A were plotted (**Figure 32**) and domains defined at a minimum of 20% tolerance. The F2+ rules were derived from theoretical patterns of Cl and Br atoms according to a ±20% tolerance. **Figure 33** presents the way these rules are expressed in the script.

Table 2: A-2, A and A+2 cumulative relative intensities for isotopic patterns of Cl_xBr_y.

Cl	Br	A-2	A	A+2
1	-	0	100	32
-	1	0	100	97
2	-	0	100	64
1	1	77	100	24
-	2	51	100	49
3	-	0	100	96
2	1	62	100	45
1	2	44	100	69
-	3	34	100	97
4	-	78	100	48
3	1	52	100	64
2	2	39	100	89
1	3	86	100	48
-	4	69	100	65
5	-	63	100	64
4	1	44	100	83
3	2	93	100	49
2	3	74	100	63
1	4	61	100	79
-	5	51	100	97
6	-	52	100	80
5	1	100	100	51
4	2	80	100	63
3	3	65	100	77
2	4	55	100	93
1	5	90	100	61
-	6	77	100	73

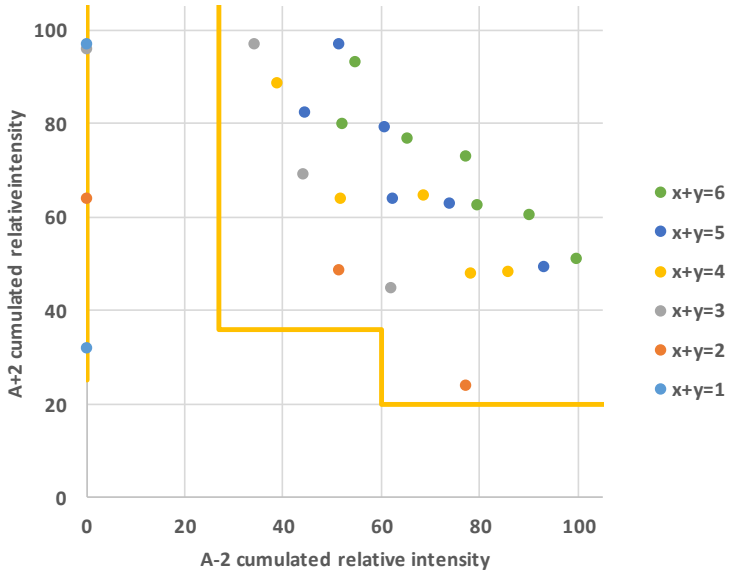


Figure 32 : A+2 versus A-2 cumulative relative intensities for isotopic patterns of Cl_xBr_y formulas (dots) and limits of selected domain for ion ratio rules for halogenated ion (orange).

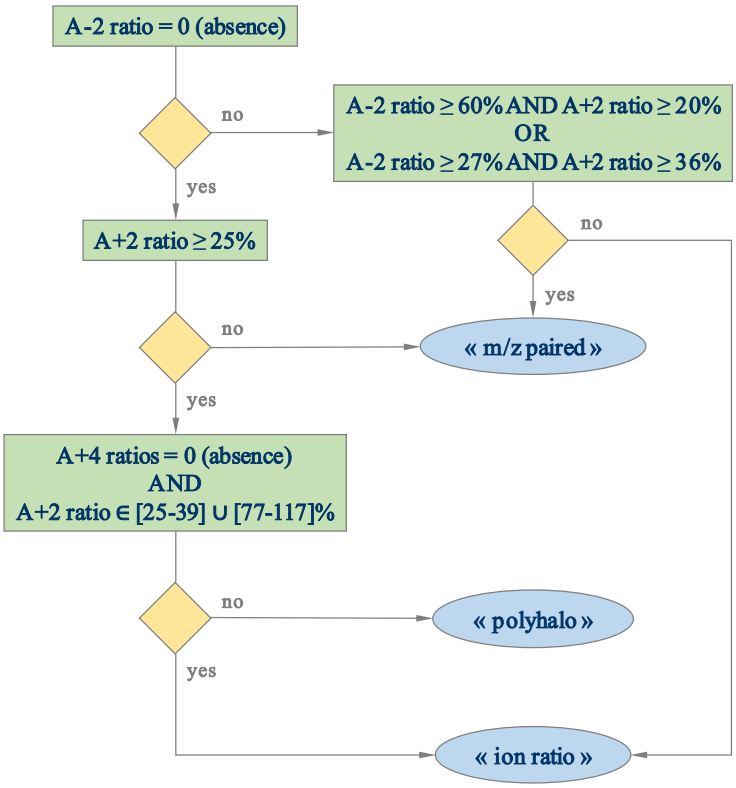


Figure 33 : Logic diagram of the ratio rules within the script.

6.1.4.3. Cluster slope

In H/Cl-scale MD-plots, the cluster slope provides useful information on the nature of the halogens (brominated, chlorinated or mixed). Indeed, the theoretical slope is -3.3×10^{-4} and $+1.2\times10^{-4}$ with correlation coefficient of 1 for (poly)chlorinated and (poly)brominated clusters, respectively, whereas mixed clusters exhibit intermediary slopes with lower correlation coefficients. This latter affirmation is true when resolution does not exceed a few hundred thousand, leading to a unique centroided peak from isotopologues of similar nominal *m*/*z*.

During the pairing process, the script empirically annotates each cluster as fully chlorinated, fully brominated or mixed Cl/Br ion, according to this slope (see also § 7.4). It is thus possible to filter clusters according to the slope.

6.1.4.4. Annotation status

The discard status can be annotated to an entire SuperGroup though the SuperGroup investigation modal (§ 7.7.2).

Formulas and/or a compound can be annotated though the SuperGroup investigation modal (§ 7.7.3 and § 7.7.4). The "*non annotated*" slider option displays clusters belonging to SuperGroups for which none of the clusters is annotated with a formula or a compound. Quite inversely, the "*annotated*" slider option displays clusters with an annotated formula or compound. Thus, be aware that clusters with no annotated formula or compound but belonging to SuperGroups for which at least one cluster is annotated with a formula or a compound are not considered in these two last slider options.

6.2. SUB-TAB TIC/EIC & MS

By default, the "*TIC/EIC & MS*" sub-tab (**Figure 34**) displays the total ion chromatograms (TICs) of the project files (**2**). Sample files to be displayed can be selected using the picklist (**1**). Instead of TICs, extracted ion chromatograms (EIC) can be displayed according to a defined *m*/*z* and associated tolerance (mDa) (**1**). Do not forget to click on the "*EIC*" button to update the chromatogram figure, or on the "*TIC*" button to go back to TICs. Depending on the number of files, it takes a few dozens of seconds to be displayed, so be patient.

A click on the TIC/EIC figure displays the mass spectra (**3**) corresponding to the selected retention time (t_R) coordinate. Each paired cluster is displayed with a unique color.

Figure 34 : Chromatogram and mass spectrum sub-tab.

6.3. SUB-TAB ENUMERATION

The "*Enumeration*" sub-tab reports the achievement of annotation (**Figure 35**). The upper area (**1**) allows to select the project files to be considered and to filter data by a defined *m/z* range, retention time (t_R) interval and/or intensity threshold. Do not forget to click on the "*Enumerate*" button to update the enumeration tables provided in the two panel boxes (**2**).

The panel box "*Resume*" (**2**) provides the number of annotated clusters according to the halogen filters (F0, F1, F2 and F2+, see § 6.1.4.2), as well as their cumulative intensity.

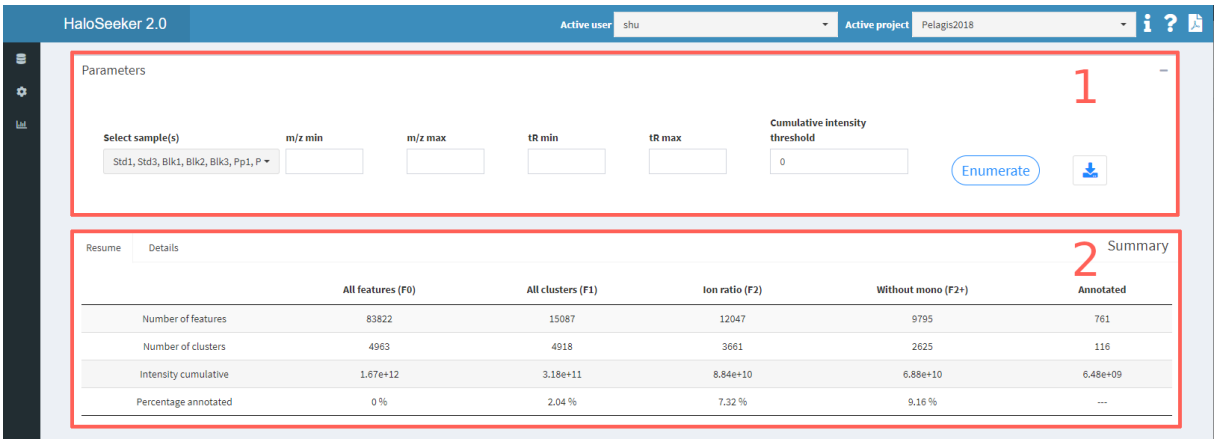


Figure 35 : "*Enumeration*" sub-tab, providing annotation achievement.

The panel box "*Details*" (**Figure 36**) provides the annotation with a table where each row represents a SuperGroup. The button shows a sub-table under those rows displaying information for all clusters belonging to the SuperGroup (t_R, intensity, adduct, pattern score, ….).

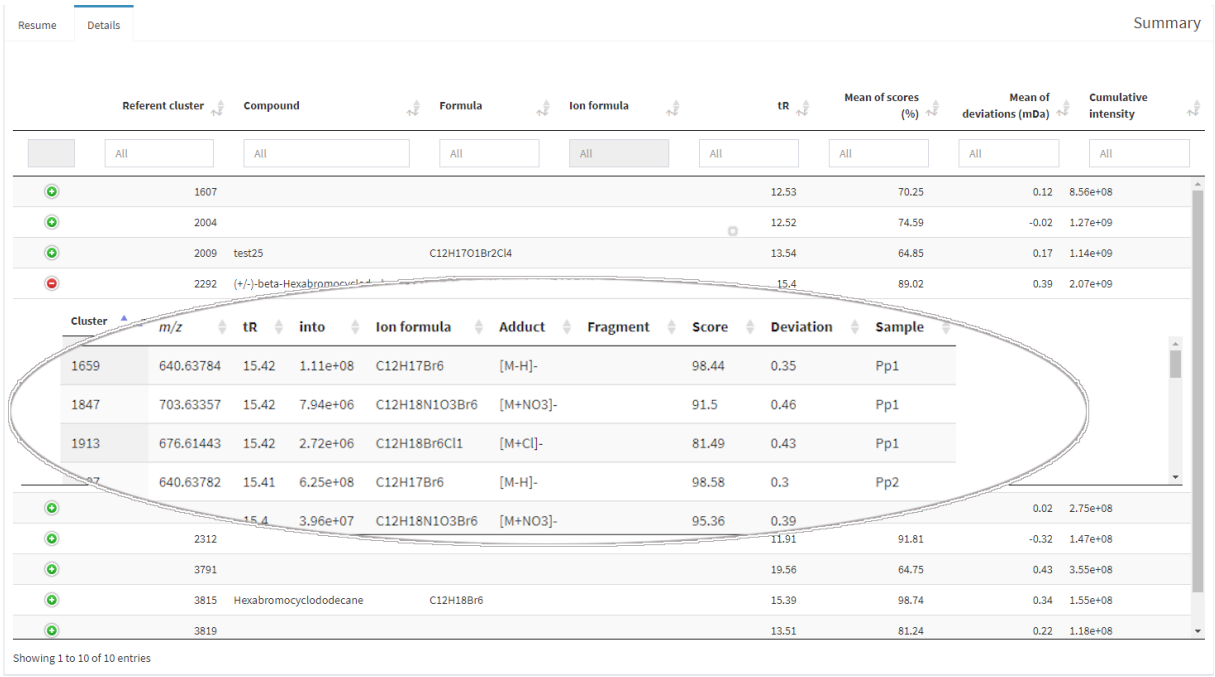


Figure 36 : "*Details*" panel box of the "*Enumeration*" sub-tab, providing SuperGroup annotation achievement.

Those results can be exported using the button as an Excel® file containing an overview of the parameters used in a first sheet, the "*Resume*" table in a second sheet and the entire data table for each file in separated sheets.

7. SUPERGROUP INVESTIGATION MODAL

7.1. OVERVIEW

This modal can be accessed by clicking on a dot in the interactive plot or a mass spectrum peak on the "*TIC/EIC & MS*" sub-tab. The modal is dedicated to the investigation and annotation of the corresponding SuperGroup appears (**Figure 37**). The upper-left area (**1**) gathers the formula decomposition parameters. The obtained results are then displayed (**2**). Information and representations such as EICs, H/Cl-scale MD-plot slope and feature table of the active cluster is provided on the upper-right area (**3**). A table presenting the SuperGroup is available (**4**), as well as a hybrid mass spectrum (**5**). Lastly, annotation can be made using buttons at the bottom of the modal (**6**).

The formula decomposition is performed on a cluster defined as the "*active cluster*". Be aware that by default the "*active cluster*" is the one exhibiting the highest cumulative intensity, no matter the cluster on which the user clicked (interactive plot or "*TIC/EIC & MS*" sub-tab) to open the modal. The "*referent cluster*" is the one corresponding to the displayed results, so that once the decomposition is performed, the "*active cluster*" becomes the "*referent cluster*".

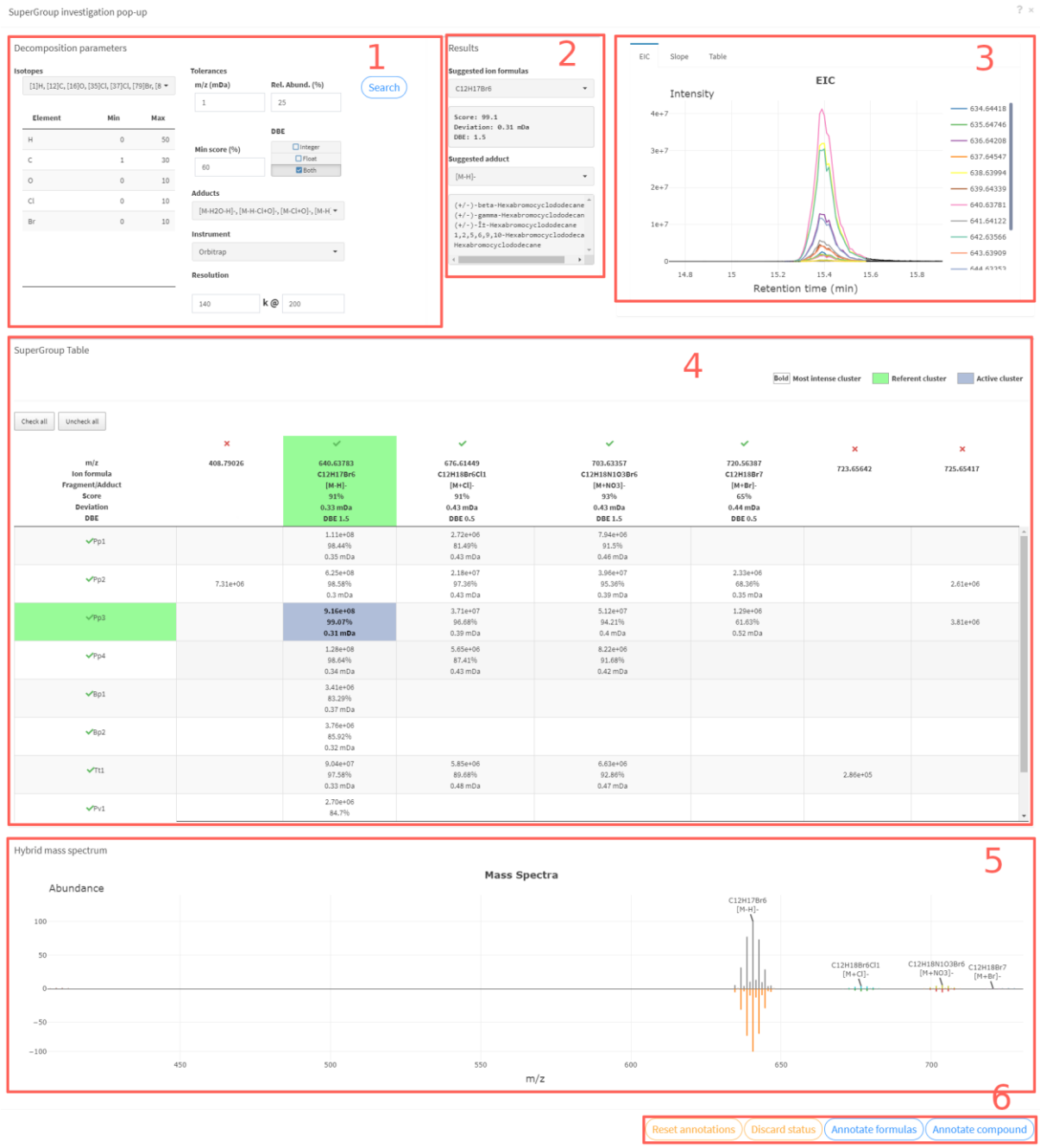


Figure 37 : Modal dedicated to the investigating and annotation of a SuperGroup.

7.2. DECOMPOSITION PARAMETERS

Available parameters for the user are as follows (**Figure 38**):

- Selection of the isotopes. By default, the following elements are selected: C, H, O, Cl, Br.
- Limitation to minimum and maximum (ranges) for each element according to the isotope choices. By default, the minimum value for each element is "0", except for the ^{12}C at "1", in order to save processing time considering that ions arise from organic molecules. This parameter can be set to "0". In case ^{13}C is selected as isotope, the default of minimum value for ^{13}C is "1" and that for ^{12}C is decreased to "0".
- Selection of the *m/z* tolerance for the decomposition.
- Selection of the relative abundance tolerance for each isotopologue relative to the base peak.
- Selection of a pattern score threshold. The scoring formula of HaloSeeker is:

$$100 \times \left(2 - \sum_{i=1}^{i=N} |Abd_i^{theo} - Abd_i^{obs}| \times \left(\frac{Abd_i^{theo}}{\sum_{j=1}^{j=N} Abd_j^{theo}} + \frac{Abd_i^{obs}}{\sum_{j=1}^{j=N} Abd_j^{obs}}\right)\right)$$

  where *N* is the number of isotopologues, *theo* the theoretic and *obs* the observed isotopologue.
- Limitation of the double-bond equivalent (DBE) possibilities (float, integer or both). This parameter depends on the knowledge of the user about the ionization process and the expected ion types. "Float" means that the DBE exhibits a decimal, *e.g.* 1.5 for the $[M-H]^-$ ion of hexabromocyclododecane ($C_{12}H_{18}Br_6$) while the molecule has a BDE of 1.
- Limitation on a pool of adduct ion possibilities, used only in the second part of the script, when seeking adducts or fragments from the "*active cluster*". By default, the sub-set of monocharged adduct types from the database (§ 4.2.4) is activated. It is advised to limit this picklist as much as possible in order to save processing time. Also, the decomposition of HaloSeeker 2.0 script only considers monocharged adduct types, because the pairing script does not pair features based on polycharged mass differences (§ 5.1.2).
- Selection of the HRMS instrument resolution. This information is useful to compute theoretical isotopic profile. For Orbitrap mass analyzers, the nominal resolution and the reference *m/z* for his nominal resolution are considered for calculating theoretical resolution at any *m/z*. For available ToF instruments, cubic interpolations (12 points) were done using the experimental data available from enviPat (**Figure 39**).

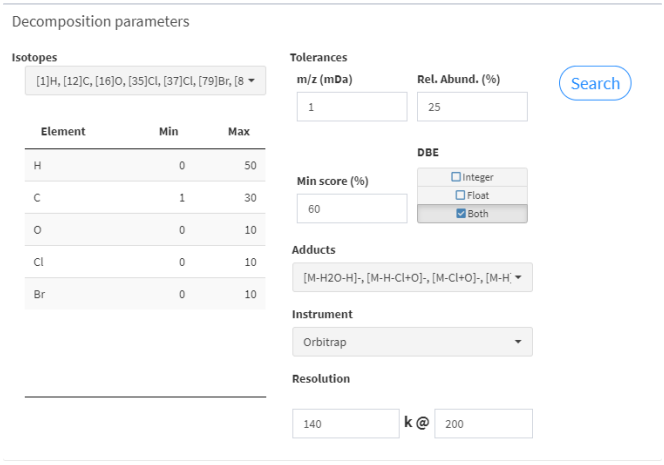


Figure 38 : Formula decomposition parameters.

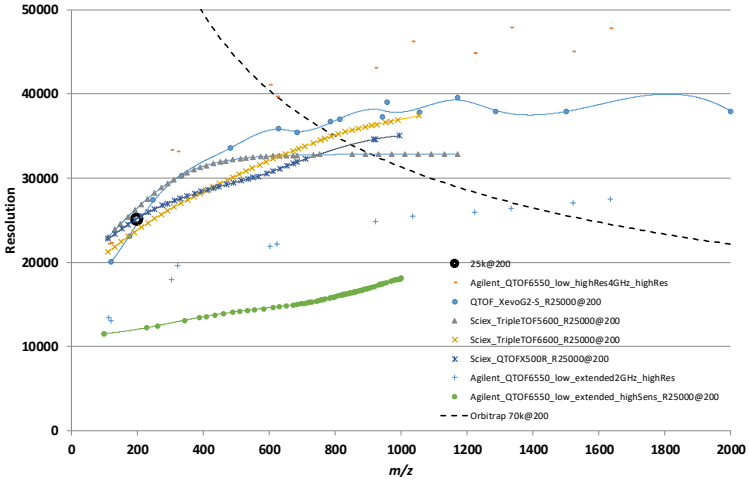


Figure 39 : Modeling of resolution according to *m/z* for a selection of ToF mass analyzers.

Do not forget to click on the "*Search*" button to launch the formula decomposition. The script is subdivided in two successive parts (**Figure 40**):

- Firstly, the *m/z* of the base peak A of the "*active cluster*" is decomposed to retrieve all potential formulas, based on the *Rdisop* package (Böcker and Lipták, 2007), with slight in-house modifications to consider the isotopes rather than the elements. This modification is crucial when working on isotopologues which are not corresponding to the monoisotopic. Then, the solutions are filtered according to most of the seven Golden Rules (Kind and Fiehn, 2007). Rules #1, #3, #4, #5 and #6 are active. Within Rule #2, LEWIS rules were deactivated, SENIOR rules remaining active. Rule #7 on TMS check when trimethylsilylation is used with gas chromatography is not activated. Eventually, solutions are sorted according to an in-house match scoring between the observed and the theoretic isotopic profiles.
- Secondly, *m/z* differences with other clusters are decomposed in order to identify adducts or fragments concordance between those and the "*referent cluster*", enhancing the annotation hypothesis.

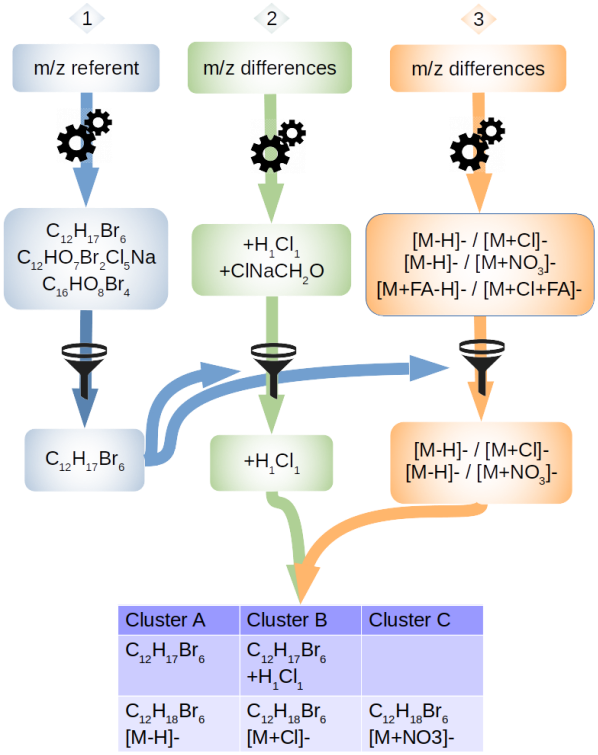


Figure 40 : Illustration of the formula decomposition script.

7.3. RESULTS

Candidate ion formulas for the selected cluster are then sorted according to some information: pattern score, mass deviation and double-bond-equivalent (DBE) (**Figure 41**). Hits with the compound database are given. Also, an adduct can be suggest, along with the number of explained clusters. Thus, for a selected ion formula, do not forget to review the suggested adducts and possibly select one.

Figure 41 : Formula decomposition results.

7.4. EIC, SLOPE AND TABLE

A tabbed box displays some information, considering all the isotopologue features of the "*active cluster*" (**Figure 42**):

- An Extracted Ion Chromatograms (EICs), colored parts representing the integrated ranges.
- A zoom-in of the H/Cl-scale MD-plot on the "*active cluster*". It also displays the linear regression curve of the even features (A, A+2, A-2, A+4, A-4...). The slope provides useful information on the nature of the halogens (brominated, chlorinated or mixed). Indeed, in H/Cl-scale MD-plots, the theoretical slope is -3.3×10^{-4} and $+1.2\times10^{-4}$ with a correlation coefficient of 1 for (poly)chlorinated and (poly)brominated clusters, respectively, whereas mixed clusters exhibit intermediary slopes with lower correlation coefficients. This latter affirmation is true when resolution does not exceed a few hundred thousand, leading to a unique centroided peak from isotopologues of similar nominal *m/z*. Colored areas empirically suggest atoms to be selected for the formula decomposition.
- A table with all the data. It is possible to remove feature(s) to score calculations by selecting a row(s). The pattern score is automatically updated. A click on the "Search" button for a formula decomposition processes the search without the selected rows.

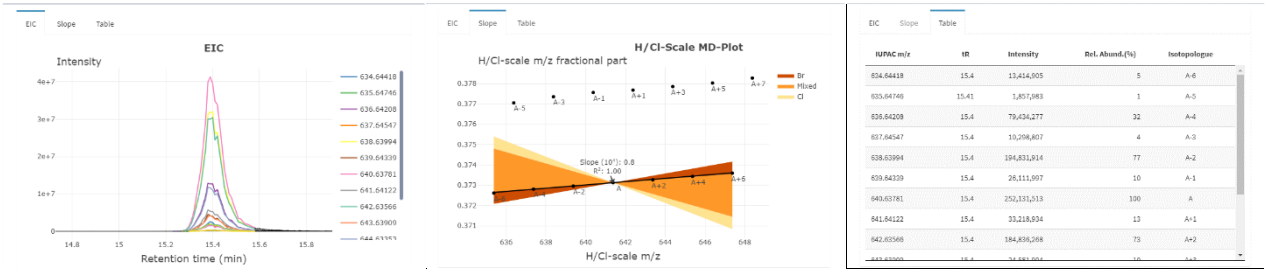


Figure 42 : EICs, regression curve slope and feature table of the "*active cluster*".

7.5. SUPERGROUP TABLE

The SuperGroup table (**Figure 43**) shows all Cl/Br-paired clusters (output of pairing step) belonging to the selected SuperGroup, according to both deconvolution dimensions which are the intra-file grouping of fragments & adducts (output of the optional adduct & fragments deconvolution process) (rows of sample files) and the inter-file alignment (output of the optional alignment deconvolution process) (columns of aligned clusters). Each cell displays the cumulative intensity of paired isotopologue features, if not empty. When a chemical formula is selected after a decomposition search, it also contains the following information:

- The pattern score in comparison with the suggested theoretic isotopic profile.
- The mass accuracy (deviation in mDa) in comparison with the theoretic isotopic profile.

Also each column head contains the following information:

- The average base peak A *m/z*.
- The ion formula according the formula selected from the decomposition search.
- The fragment or adduct ion formula relative to the "*referent cluster*" ion formula.
- The mean pattern score of all the clusters in the column.
- The mean mass accuracy (deviation in mDa) of all the clusters in the column.
- The DBE of the ion formula.

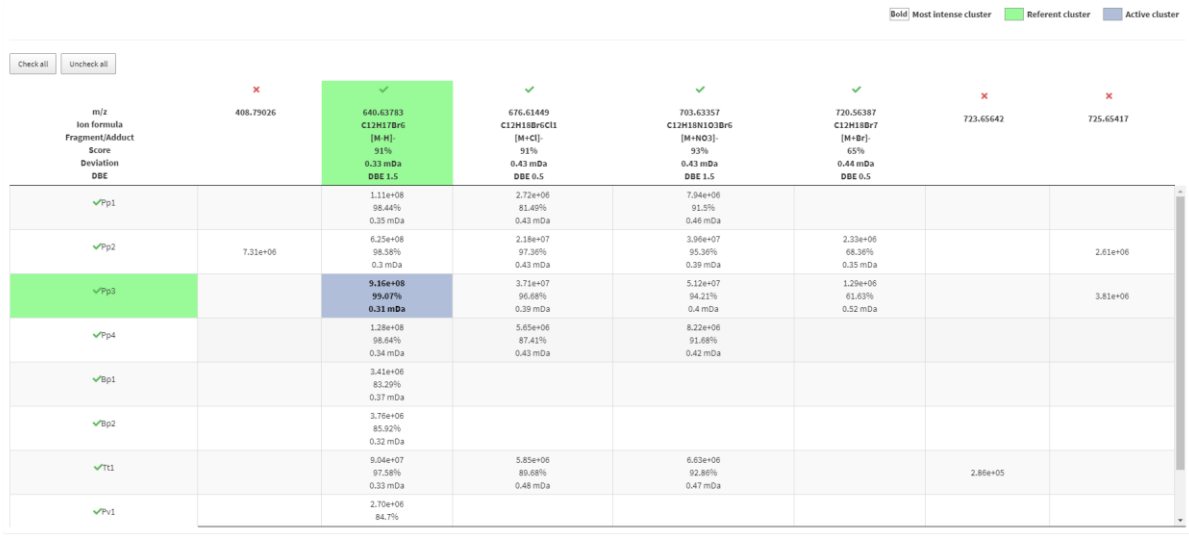


Figure 43 : SuperGroup table.

No matter the cluster from which the SuperGroup investigation modal was opened, the "*referent cluster*" and the "*active cluster*" are defined both as the most intense clusters of the SuperGroup. The most intense cluster of the SuperGroup always remains in bold font. The reference cluster, corresponding to the one from which the decomposition was performed, corresponds to the green background in the column and row head cells. Another "*active cluster*" can be selected with a click, with subsequent automatic update of the EIC, slope and table information. Another formula decomposition search can then be launched from the selected active cluster, which thus becomes the "*referent cluster*".

Each column and row head contains a checkbox. By default, only the columns with a suggested ion formula and the most informative row for each sample file are activated. The cells (clusters) for which both column and row checkboxes are activated constitute a sub-set of the SuperGroup on which the user can apply or not the annotation (see § 7.7).

Depending on the deconvolution parameters, it is possible that one or more clusters belonging to the same sample file are aligned within the SuperGroup (successive rows named identically but with an increment number between brackets). By default, the most intense is considered on the first row which gathers potential fragment and adducts. It is possible to interchange such aligned clusters with a sample file using a simple drag and drop.

7.6. HYBRID MASS SPECTRUM

The hybrid mass spectrum (**Figure 44**) displays the mass spectrum of the most intense cluster of each SuperGroup column. It is assumed that it is the one with the best definition/quality. This way, "*hybrid*" means that those clusters might come from different sample files. Also, do not forget that only paired clusters (i.e. Cl/Br-halogenated) are considered in SuperGroups, so that non halogenated fragments are not searched nor displayed.

After a formula decomposition search, the theoretical isotopic profiles of the selected ion formula, including those of the suggested fragments and adducts, are displayed in the mirror hybrid mass spectrum.

Figure 44 : Hybrid mass spectra of the observed (top) most intense cluster of each SuperGroup column and corresponding theoretical suggested ion formula (mirror).

7.7. FORMULA ANNOTATION

After operating formula decomposition searches, reviewing suggested formulas and selecting clusters to be annotated, four buttons are available at the bottom of the modal for the annotation (**Figure 45**). Once an action is completed, the application goes back to the interactive plot. Do not forget to click on the "*Draw*" button to update the plot.

Figure 45 : Annotation buttons.

7.7.1. Reset annotation

Reset annotations erases previous annotations of all the clusters belonging to the SuperGroup. A confirmation modal is associated (**Figure 46**).

7.7.2. Discard status

Discard status (re-)annotates all clusters of the SuperGroup as being uninteresting, *e.g.* if the user concludes it does not correspond to a halogenated compound. A confirmation modal is associated (**Figure 47**). It is then possible not to display it on the interactive plot by using the annotation status filter slider (§ 6.1.4).

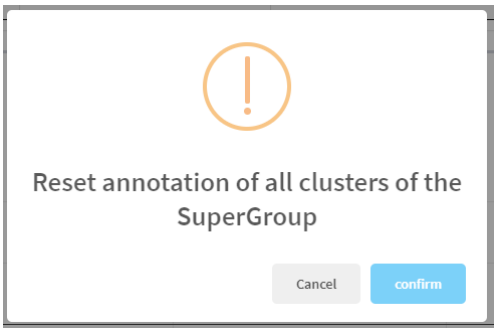


Figure 46 : Warning modal for resetting annotation.

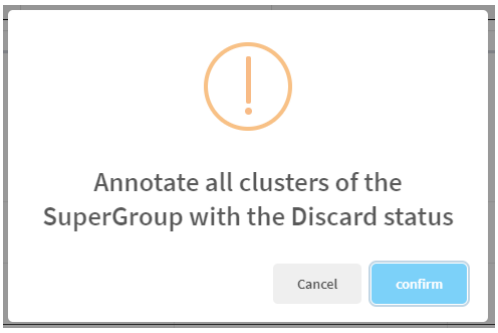


Figure 47 : Warning modal for annotating a discard status.

7.7.3. Annotate formulas

The button "*Annotate formulas*" allows to annotate the suggested chemical and ion formulas to clusters for which the checkboxes of the SuperGroup table are active (§ 7.6). For the export Excel table, a confidence level of 4 according to Schymanski et al. (2014) is annotated.

If active and non-active checkboxes are present in the SuperGroup table, a choice modal is associated to split the SuperGroup (**Figure 48**). The sub-set of clusters for which both column and row checkboxes are activated is kept in the SuperGroup (for which part or all clusters are annotated) while the other sub-set is subjected

to a new SuperGroup determination (clusters sharing a same fragment/adduct ID and/or alignment ID with one another; nearest neighbors) leading to one or several new SuperGroup IDs.

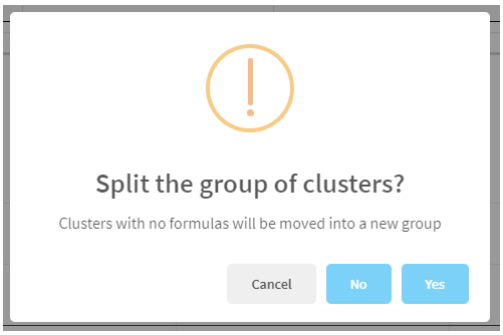


Figure 48 : Choice modal for splitting a cluster SuperGroup.

7.7.4. Annotate compound

Similarly to the button "*Annotate formulas*", the button "*Annotate compound*" allows to annotate the suggested chemical and ion formulas to clusters for which the checkboxes of the SuperGroup table are active (§ 7.6). For the export Excel table, a confidence level of 3 according to Schymanski et al. (2014) is annotated.

Similarly to the button "*Annotate formulas*", if active and non-active checkboxes are present in the SuperGroup table, a choice modal is associated to split the SuperGroup (**Figure 48**). The sub-set of clusters for which both column and row checkboxes are activated is kept in the SuperGroup (for which part or all clusters are annotated) while the other sub-set is allocated a new SuperGroup ID (for which there is no annotation).

In addition, another modal (**Figure 49**) allows to select a compound among the compound database hits (according to the suggested compound hits and formulas derived from the suggested list of adducts). If there is no hit, then the entire compound database is suggested.

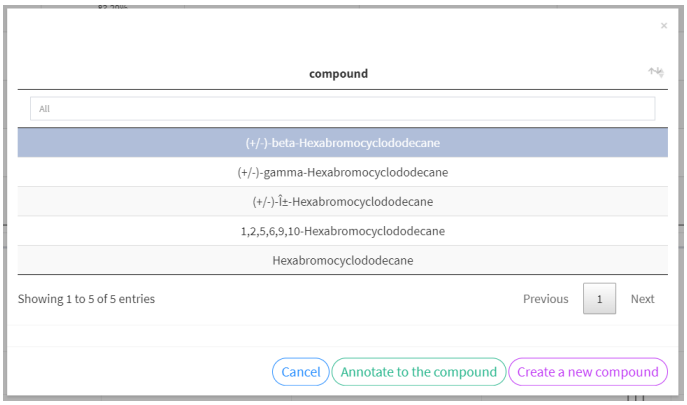


Figure 49 : Modal for the selection of a compound to be annotated, including possible creation of an entry in the compounds database.

7.7.4.1. Create a new compound

The user can create an entry in the compounds database using the "*Create a new compound*" button. Required information consists in a name and a chemical formula (**Figure 50**).

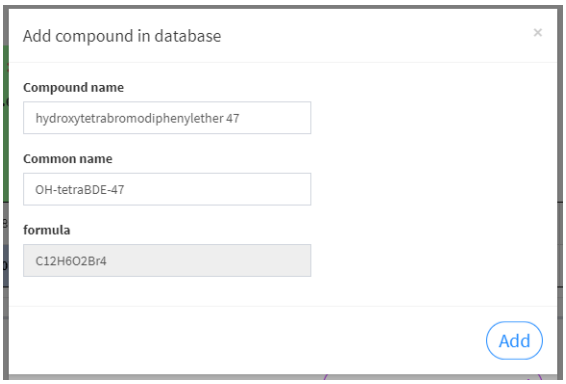


Figure 50 : Creation of a new compound entry.

8. FREQUENTLY ASKED QUESTIONS (FAQ)

The shortcut to launch HaloSeeker is not working anymore

Sometimes, antiviruses like Avast can mark the software *R* as a virus. If the application does not launch, check if your antivirus has been assigned to quarantine.

How to install the Microsoft update to make MSConvert work?

MSConvert requires updated Microsoft.NET framework 3.5 SP1 and 4.0 that we cannot embed in HaloSeeker. Indeed, full admin rights are needed to update it. They can be downloaded at:

https://www.microsoft.com/en-US/download/details.aspx?id=22

https://www.microsoft.com/en-US/download/details.aspx?id=17851

I face difficulties with my wiff files

In the case of a wiff file, please check the existence of its corresponding wiff.scan in the same directory.

I face difficulties with my Waters files

Depending on the MS method, converting Waters raw data files to *mzXML* is a bit tricky but manageable with MSConvert. First, process the files with Masslynx\Tools\Accurate mass measure and save it (**Figure 51**). That step will avoid one of the known ppm deviation regularly seen during Waters file conversion to *mzXML*. In addition, if the method contains lockmass, then remove it before using MSConvert by renaming the FUNC002.dat to FUNC002.old (here we assume that FUNC002.dat contains lockmass data but it is not always the case depending on the MS method).

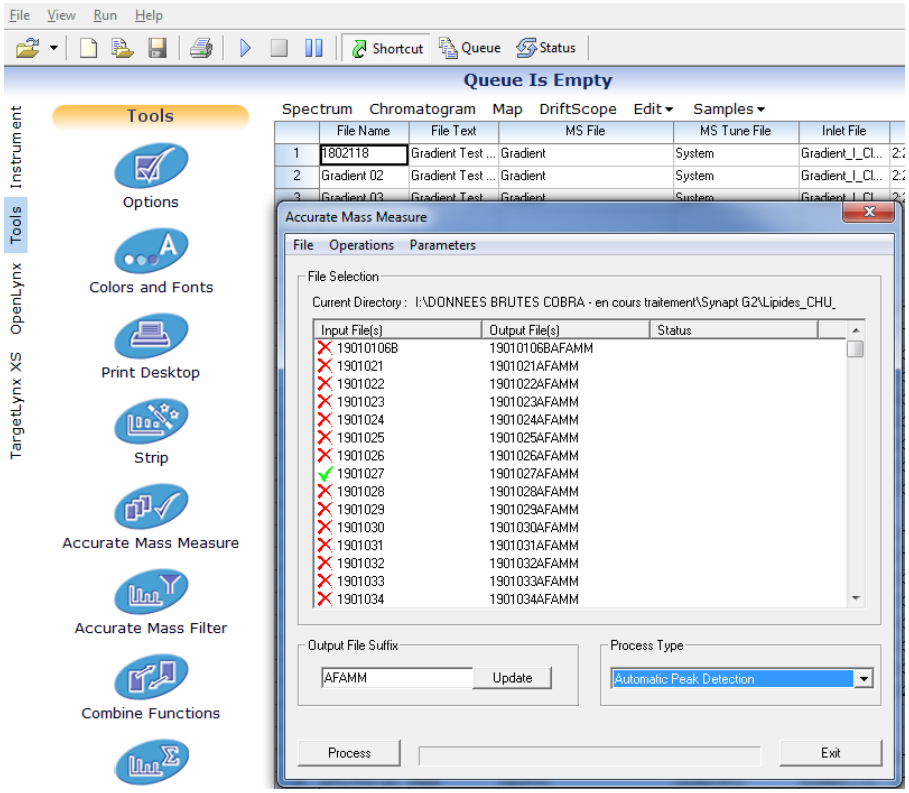


Figure 51 : Accurate mass measure window of MassLynx to adapt Waters data files.

The Chrome browser user interface (UI) turns gray and does not respond

The UI turning gray indicates that something went wrong and that *R* disconnected from the interface (and possibly even closed) for various reasons. Nevertheless, this does not imply your project or intermediate calculation results have been lost. Possibly, it happens when the computer is automatically paused. You can restart the application, possibly after forcing the shutdown of the *R* application in the task manager (ctrl+alt+suppr). Alternatively, contact us by e-mail, attaching the file error.log to help us improve the software. The file error.log can be found in the directory C:\Users\User\.HaloSeeker 2.0.

A weird and incomprehensible error message appears

This means that an error occurred during a process from a non-expected action. Contact us by e-mail, attaching the file error.log to help us improve the software. The file error.log can be found in the directory C:\Users\User\.HaloSeeker 2.0.

An error message says that the database is corrupted

Well, this is a recurrent trouble on Windows systems. The application records backups of the database every day and stores daily, weekly and monthly backups. You can retrieve these backups in the folder backup of the installation directory. Do not forget to rename it "*database*" and put it in the installation directory to replace the corrupted one.