MergeION spectral library search GUI

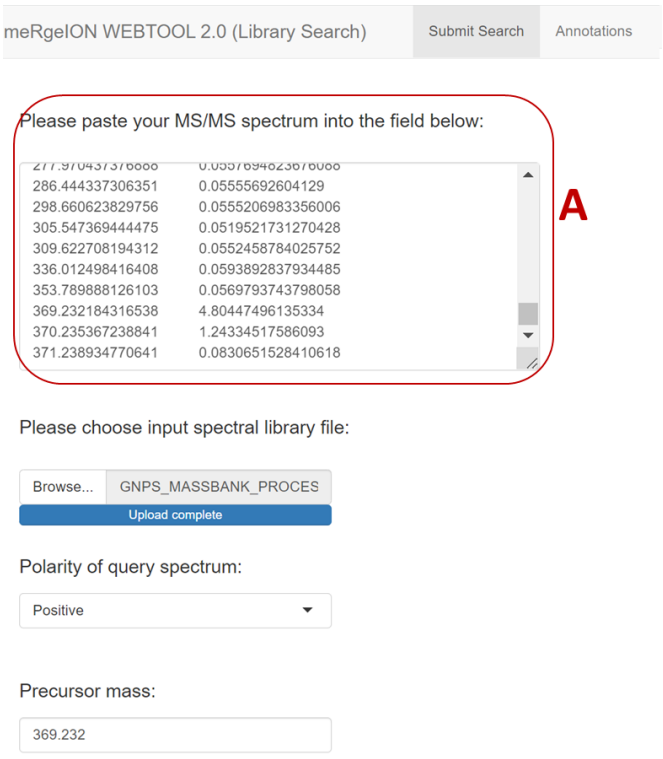
2023-02-02

The library search GUI enables automated structure proposal from unknown compound by searching their high tandem mass spectra against a user-uploaded spectral database. The output of library search is a list of structural candidates whose reference MS/MS spectra resemble the query spectrum based on user-defined spectral similarity metrics. The structural candidates can be either putative structure hits (exact search) or a list of analog compounds (analog search).

0. Calling the GUI after installing MergeION

```
library(MergeION) # First load MergeION
library(RChemMass) # Load the RChemMass package for structure visualization
runGUI()
```

1. Copying query spectrum from vendor software



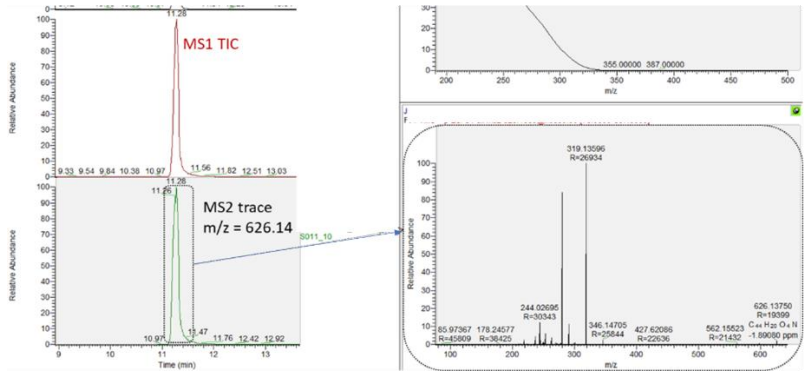
Query spectra from Thermo, Bruker, and Water instruments in the centroid mode are accepted. Please following the instructions for copying data from different vendors:

1.1 From Thermo XCalibur QualAnalysis

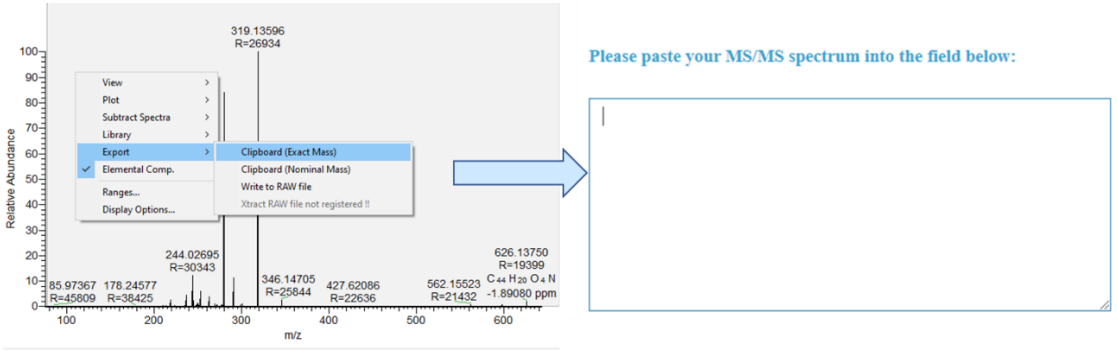
1) Open LC-MS/MS data, display MS2 chromatogram trace of precursor m/z of interest

Type	Range	Scan filter	Delay (min)
☑ Wavelengt...	260.000...	-	0.00
☑ TIC	-	FTMS + p ESI Full ms [200.0000-2000.0000]	0.11
☑ TIC	-	FTMS + p ESI d Full ms2 626.1389@hcd30.00 [76.0000-637.0000]	0.10

2) Activate the mass spectra cell. Averaging or select a point (usually the peak) of MS2 chromogram trace. MS/MS spectrum will be displayed:



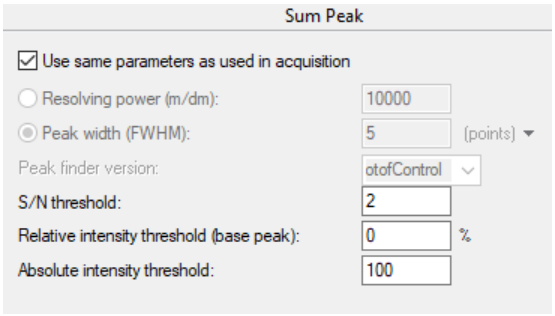
3) Inside the mass spectra cell, right click → export → Clipboard (Exact Mass) → Ctrl + V into GUI



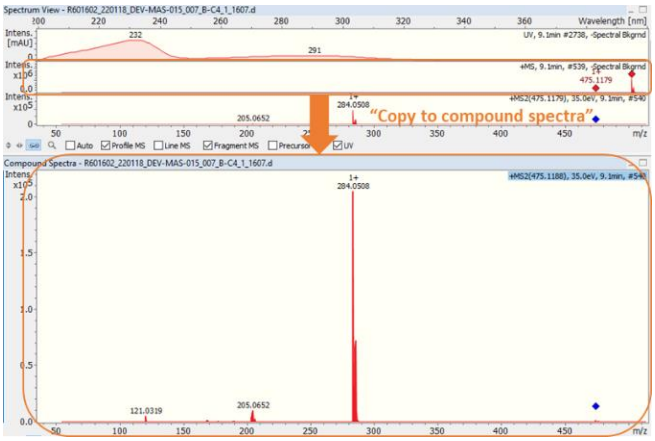
1.2 From Bruker DataAnalysis

1) Open and recalibrate the LC-MS/MS data if necessary

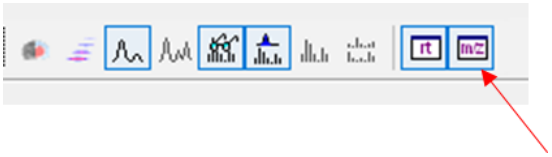
2) MassList → Parameters → Select “Sum Peak” as the algorithm for mass peak picking → Setting “Sum Peak” parameters as follows:



3) Copying the MS/MS spectra of interest from “Spectrum View” to “Compound Spectra”:

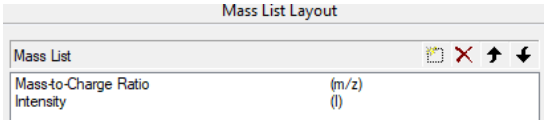


4) Make sure Spectrum data view is activated:

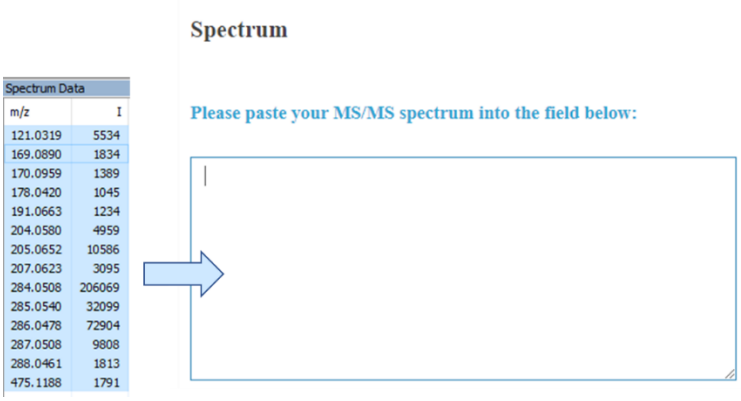


5) Select compound spectra → MassList → Find

6) In Spectrum Data - Mass List panel, right click → select “Layout”, layout should be set:



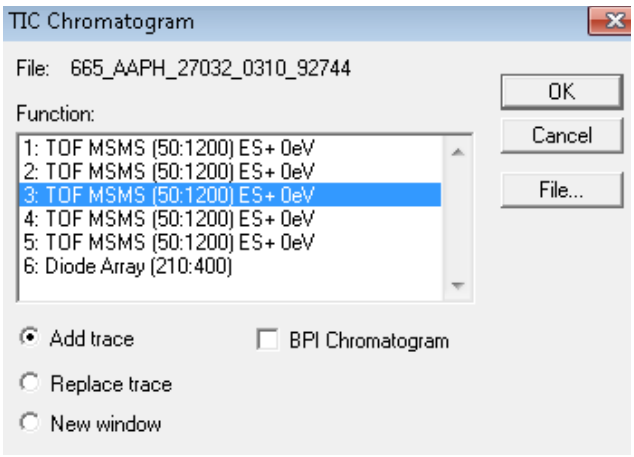
7) In Spectrum Data - Mass List panel, Ctrl + Alt (select all) → Ctrl + C (Copy) → Ctrl + V to webtool:



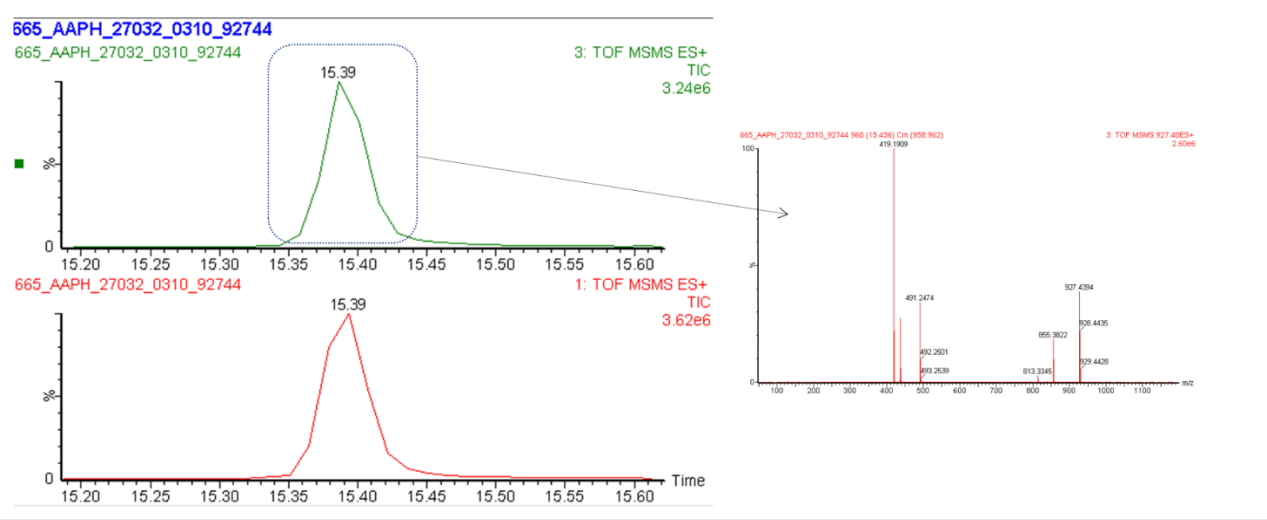
1.3 From Water MassLynx

1) Open LC-MS/MS data file, Display → TIC…→ Select function → Add trace

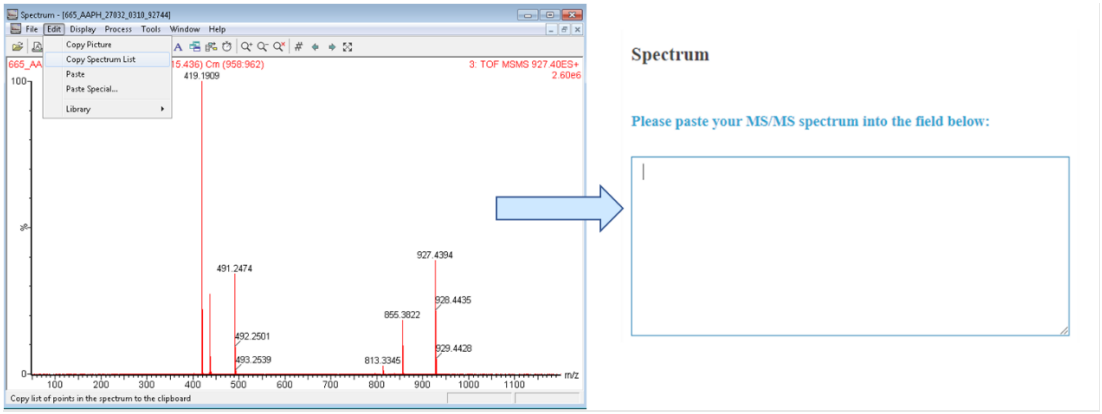
Note The selected function must be a MS/MS trace that provides sufficient fragments in the query spectrum. Depending on instruments, collision energy label can be missing when imported into MassLynx. Be sure to verify each MS/MS trace manually.



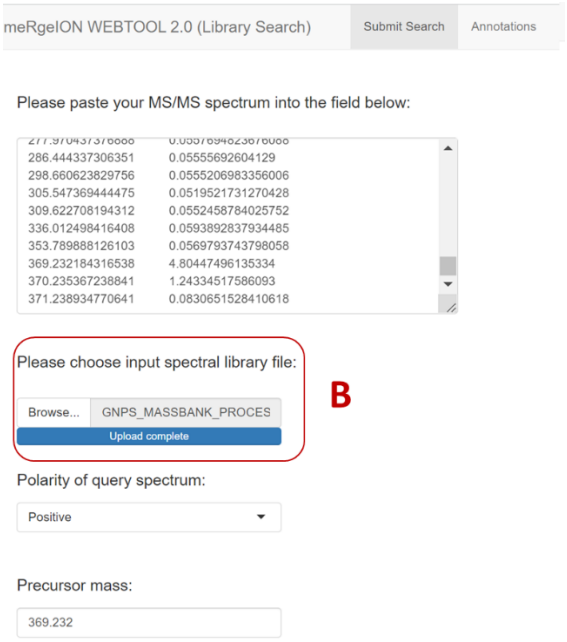
2) In the selected MS/MS trace (chromatogram), create an averaged spectrum surrounding the peak using right mouse button.



3) Edit → Copy Spectrum List → Ctrl + V to webtool:



2. Uploading spectral library



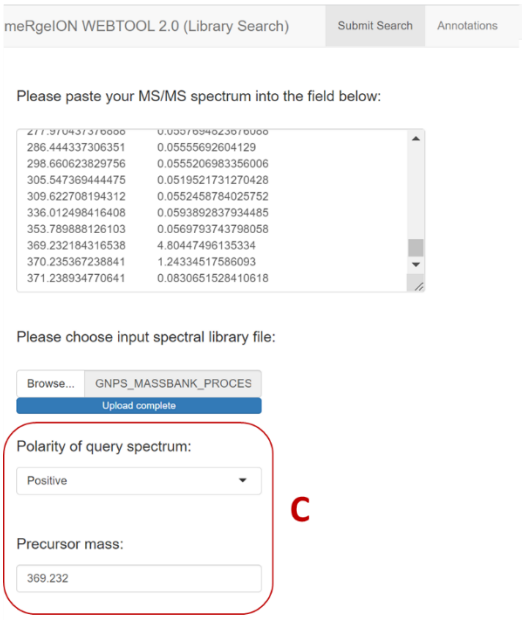
Please upload the spectral library you would like to search against. We support spectral library files in mgf, msp or RData format. Each spectral record must be labelled with a unique identifier (ID = xxx), precursor mass (PEPMASS = xxx) and polarity (IONMODE = Positive/Negative). To annotate unknowns with confidence, we recommend using pre-compiled spectral databases available at: https://zenodo.org/record/7057435/#.Y_KDmHbMLGI

For metabolite and natural product annotation, we have pre-compiled "GNPS_MASSBANK_PROCESSED_POS_CONSENSUS1.RData" containing MS/MS spectra of 11,642 metabolites, natural products and drugs. The database combines public repositories such as GNPS, MassBank and inhouse reference standards. All spectra are in positive ion mode. We are constantly populating this spectral database by combining various repositories after careful inspection of spectra and metadata quality.

We also provide a pharmaceutical spectral database called *Drug+* which contains 1,959 drug substances. This positive ion mode library is available under mgf, msp or RData format.

3. Setting up search parameters

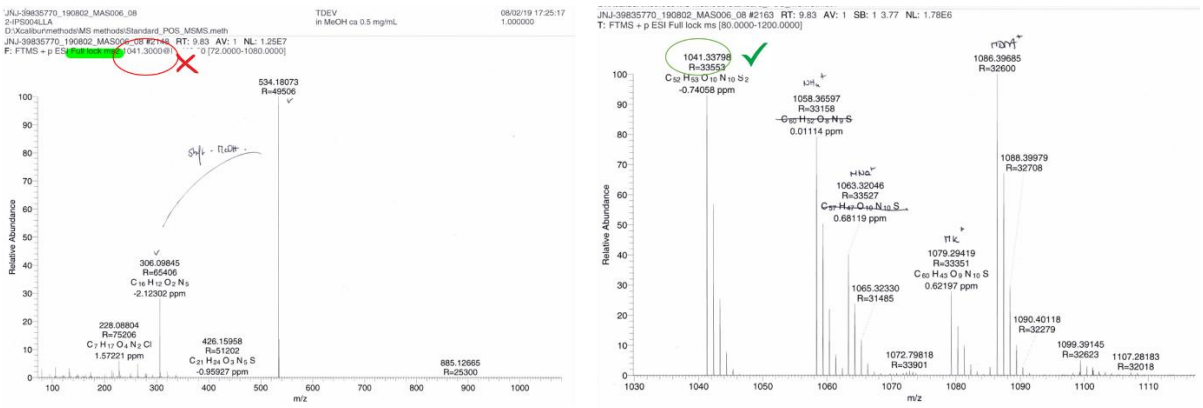
3.1. Spectrum information



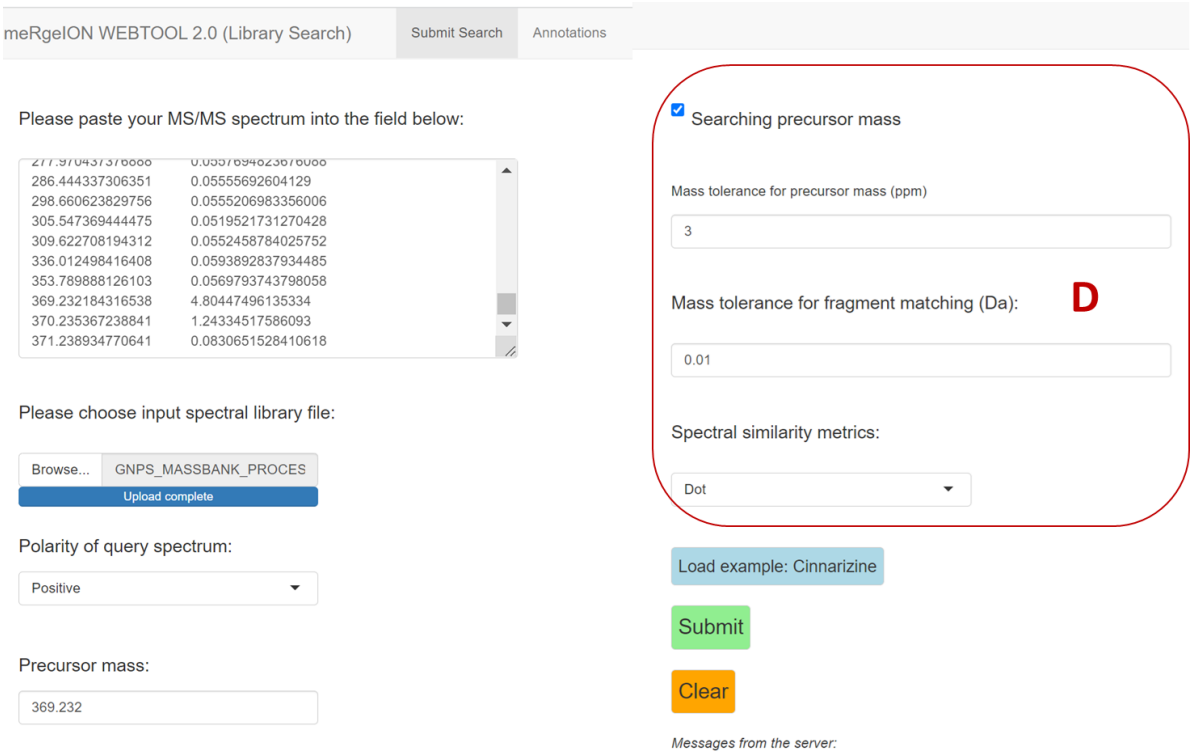
Polarity of query spectrum Ion mode of query spectrum, either positive or negative. It **MUST** be correctly provided if your input spectral library contains spectra from two polarities

Precursor mass Numeric value. Precursor mass of query spectrum. **STRONGLY RECOMMENDED** so that the algorithm will correctly calculate the neutral losses for spectral similarity calculation. Precursor mass is also used for exact library search (see 3.2.3).

Users must enter **EXACT** *m/z* value (at least 3 decimal places) of precursor ion. Error can often be made in target mode MS/MS analysis. In the following example, the query spectrum is labelled with approximate *m/z* that is **NOT** adapted for spectral library search. Please always confirm the *m/z* of precursor ion peak in original MS1 or MS2 scans.



3.2. Search conditions



Searching precursor mass

Activate this option for "Exact search": prior selection of database candidates via precursor *m/z* of the query spectrum and score them by spectral similarity. Please deactivate this option for "Analog search": compare and score all structures within the spectral database through spectral similarity without filtering on precursor m/z. We recommend starting with "Exact search", since very often, the query spectrum matches with a compound that has been analyzed and reported by the community. If no candidate is found, switching to "Analog search" can lead to compounds that resemble the structure to be identified. The reasoning is that MS2 spectral similarity is often a reflection of structural similarity.

Mass tolerance for precursor mass (ppm)

Experimental mass error (in ppm) allowed for searching precursor ion m/z in the database, used for "Exact search".

Mass tolerance for fragment matching (Da)

Experimental mass error (in Dalton) allowed for matching product ions/neutral losses in query spectrum with reference spectra. We recommend 0.01 Da for most applications.

Spectral similarity metrics

The webtool offers three ways to calculate spectral similarity to compare query and reference spectra from the inhouse spectral database. **F1**: harmonic mean of percentage of overlapping peaks and neutral losses in query and reference spectrum. N_X and N_Y are product ion/neutral loss counts from X and Y, while N_{XY} is number of product ions/neutral losses present in both spectra.

$$S_{XY} = 2\ \frac{\frac{N_{XY}}{N_X}\cdot\frac{N_{XY}}{N_Y}}{\frac{N_{XY}}{N_X}+\frac{N_{XY}}{N_Y}}$$

Cosine: Cosine Correlation R_{XY} based on the intensity profile of matched product ions/neutral losses:

$$R_{XY} = \frac{I_X \, I_Y}{\|I_X\| \cdot \|I_Y\|}$$

NIST (or Modified Cosine Correlation) considers both mass and intensity through weighted vectors. This approach puts more importance on product ions/neutral losses with higher molecular weight:

$$R_{XY} = \frac{M_x I_x \times M_y I_y}{\|M_x I_x\| \cdot \|M_y I_y\|}$$

Since Correlation is between -1 and 1, "Cosine" and "NIST" must be normalized to be in the [0, 1] range:

$$S_{XY} = R_{XY}/2 + 0.5$$

Other spectral similarity metrics available for tuning are dot product (DOT), harmonic mean (HM), MassBank and Spectral entropy.

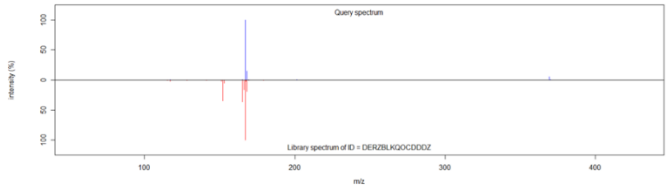

The structural candidate ranking can vary tremendously based on the spectral metrics. **We recommend starting with F1.** With our experience, F1 similarity outperform most other metrics for Janssen compounds.


Several parameters are default settings of library search. They are currently unavailable for tuning:

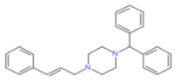
a. Query spectrum quality check: at least 3 peaks smaller than precursor ion m/z
b. Minimum number of peak/neutral loss matches for all candidates found: 5.
c. Maximum peak kept in query spectrum: top 200 most intense.
d. 0.1% relative intensity

4. Spectral library search output

E



F



G

List of annotated candidates

Structure candidates from spectral library ranked by spectral similarity score (SCORE_MERGEION). The exact search will lead to one or a few candidates with identical/similar precursor *m/z*, while analog search will generate more candidates with similar structure.

Mirror plot

By selecting one structure in the candidate table, a mirror plot will be displayed comparing input query spectrum (upper blue plot) against the candidate spectrum (lower red plot).

Candidate structure

The candidate structure can be visualized if its SMILES code is available as metadata.