MSdeCIpher is a tool to connect a chromatography-MS fragment dataset with a molecular ion dataset obtained in two different runs with two different ionization methods.

What MSdeCIpher does:

- Filter possible molecular ions based on intensity or adduct/ neutral loss criteria you define
- Calculate the sum formula of all ions
- Provide a score for every molecular ion based on % of fragments explained
- Save you time compared to manual data curation

What MSdeCIpher DOESN'T do:

- Deconvolution/ peak picking from raw data
- Provide absolute certainty about molecular ion identification
- Structure elucidation or database searching
- (Currently) the computation of heavily chlorinated of brominated compounds. Results are still usable but may be faulty.

Installation

For sum formula calculation to work, you need Java. Download and install Java https://www.java.com/en/download/ if you don't already have it.

Get R from https://www.r-project.org/ if you don't already have it. MSdeCIpher was tested on R versions >= 3.6.1

Execute the following commands in R to install MSdeCIpher (you only need to do this once):

```
install.packages("devtools") #installs the devtools package
devtools::install_github("Pohnert-Lab/MSdeCIpher") #installs MSdeCIpher from GitHub
```

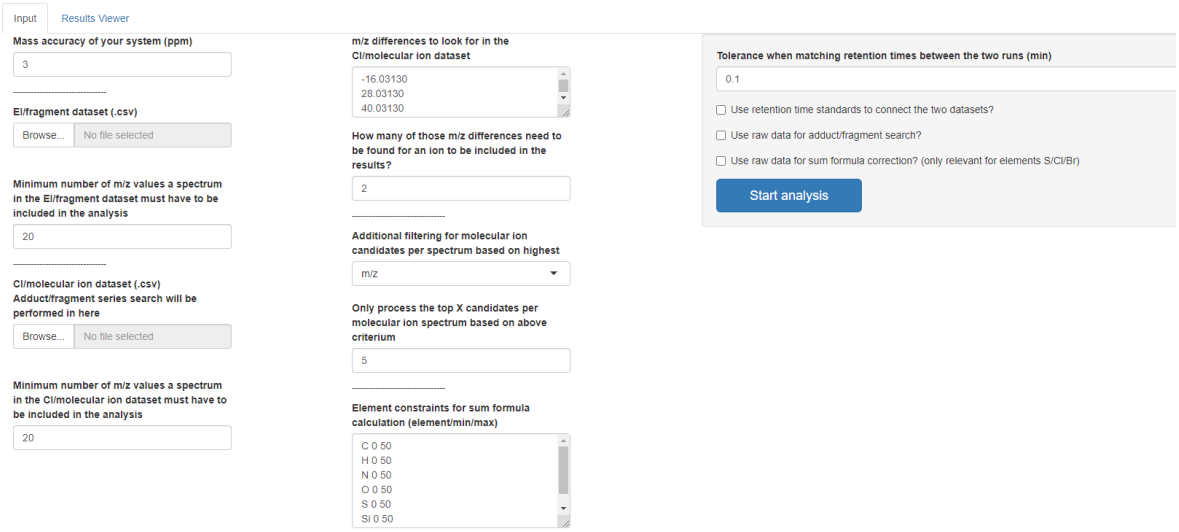
To start the MSdeCIpher user interface, execute:

```
MSdeCIpher::RunMSdeCIpher()
```

More information and help: https://github.com/Pohnert-Lab/MSdeCIpher

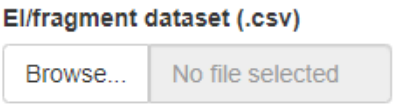
Usage

After executing MSdeCIpher::RunMSdeCIpher(), the shiny session window should open in your default browser and look like this:



Input

The overall mass accuracy of your measurements. All calculations in this tool (sum formulas, isotope and adduct/loss Δ*m*/*z*) are basing their tolerances on this value. This needs to be in the single-digit ppm range, otherwise sum formula candidates for small fragments increase to more than one and Δ*m*/*z* tolerances become too large, defeating the purpose of this tool.



A .csv file containing the deconvoluted data from the fragment data set as shown here:

mz	rt	into	pcgroup
50.01462	21.07667	11280.85	1437
50.01867	39.70833	2825.915	5324
50.0246	5.88915	6256.604	1126
50.05024	21.081	11911.55	1438
50.05469	21.08983	7419.659	1439
50.08791	8.481367	31184.96	971
50.10639	21.07667	11121.89	1440
50.10696	8.468233	19771.15	972
50.1315	39.6995	2762.129	5325
50.13969	8.498883	4911.855	3400
50.15697	28.8325	8266.524	1275
50.19559	7.974767	5675.499	1800

(The nomenclature here is borrowed from the package CAMERA)

Do not change column names!

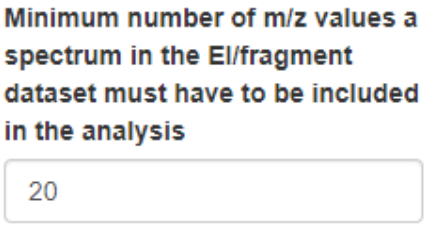
m/z – individual m/z features

rt – Retention time of the intensity apex of the feature (minutes!)

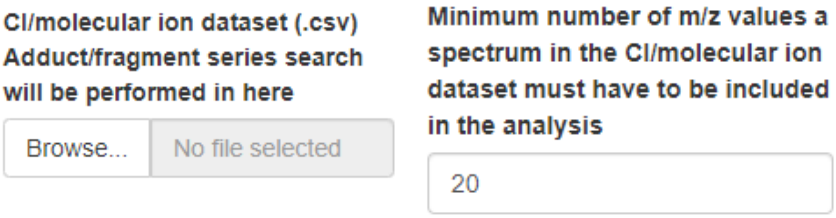
into – Integrated peak area of the feature

pcgroup – The chromatographic peak this feature was assigned to

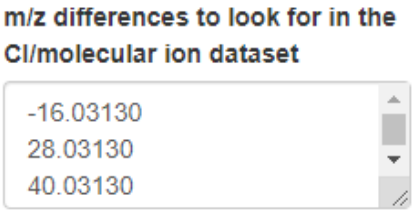
Example files can be found at https://github.com/Pohnert-Lab/MSdeCIpherThis data can be obtained from peak picking software processing the raw data from both runs. It doesn't matter which software (vendor or freely available) as long as the output can be transformed into this format (most likely a bit of spreadsheet juggling is required). The example files here have been created from GC-Orbitrap EI and CI runs with a custom R script based on XCMS and CAMERA (https://doi.org/10.3390/metabo10040143, supplementary).



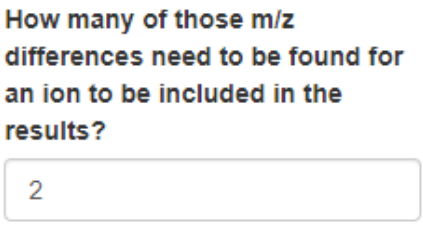
Depending on the peak picking software you use, there will be more or less junk data present. A fragment "spectrum" containing only 1 or 2 fragments is useless, at least for the purpose of molecular ion identification or structure elucidation. Here, you can define the threshold for that. Any fragment spectrum with a number of fragments below this number (isotopic ions included) will be skipped.



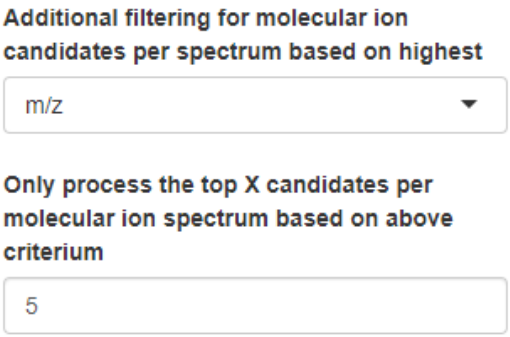
The same as above, just for the dataset containing your molecular ion candidates. If this dataset doesn't contain a lot of fragmentation, a low number is reasonable here.



Adduct/ neutral loss criteria to identify possible molecular ions. If the ions you are looking for have special neutral loss or adduct formation, you can filter the available candidates with these criteria. The default values given here are for $[M+H]^+$ molecular ions obtained by methane GC-CI of TMS-derivatized compounds, which will often display a loss of CH_3 (-16.03130) and the adduct formation of $+C_2H_4$ (+ 28.03130) and $+C_3H_4$ (+ 40.03130). If you don't want to filter based on any of those criteria, put a 0 here.

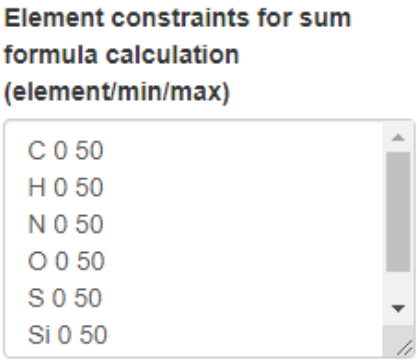


Depending on the abundance and the nature of the molecular ion, it might not always contain *all* the fragments and adducts defined in the previous step. You can adjust here how many of those criteria need to be fulfilled for an ion to be considered a molecular ion. If you want all to be fulfilled, pick the number that matches the number of criteria.

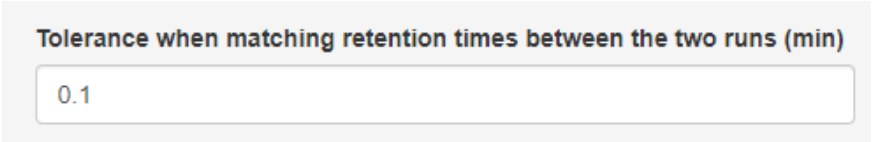


The only sensible ions to be molecular ions in a deconvoluted spectrum are those with the highest *m*/*z* values. This option will save computation time by just considering the top X molecular ion candidates with the highest *m*/*z* per spectrum. Be careful however, if you didn't define adduct/ neutral loss rules to identify molecular ion candidates. Deconvolution algorithms will often detect high *m*/*z* ions that turn out to be unexpected adducts or artefacts, thus potentially dropping the actual molecular ion out of the top X results!

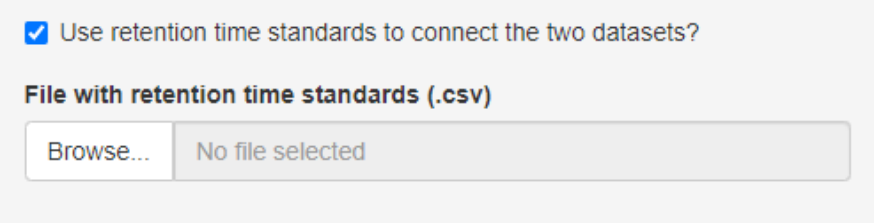
Alternatively, ionization techniques like APCI often result in the molecular ion to have the highest intensity in a spectrum. In those cases, you can filter by high intensity instead of *m*/*z*.



Constraints for calculating sum formulas of fragments and molecular ions in your dataset. You can delete lines or add more. If so, write in a new line in exactly the format depicted.



Look for specific peaks that are recognizable in both runs to determine if you have retention time shift. If you observe no retention time shift, a value of 0.05 should be fine. Increase if you see small discrepancies. In long methods with increasing retention time shift, it might be a good idea to use retention time standards to normalize retention times (see below).



standard	ei rt	ci rt
11	5.47	5.47
12	6.89	6.89
13	8.62	8.58
14	10.61	10.55
15	12.75	12.66
16	14.95	14.84
17	17.12	17.01
18	19.24	19.12
19	21.28	21.16
20	23.24	23.13
21	25.12	25
22	26.92	26.8
23	28.66	28.53
24	30.33	30.23
25	31.88	31.76
26	33.17	33.06
27	34.27	34.17
28	35.25	35.15
29	36.13	36.03
30	36.94	36.85
31	37.7	37.6
32	38.42	38.31

If you observer significant retention time shifts between your two runs, you can add retention time standards that are visible in both ionization techniques to normalize retention times (i.e. *n*-alkanes for GC EI and CI).

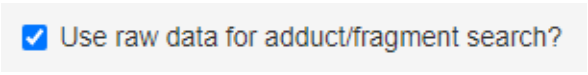
Retention times of those standards need to be manually extracted from the raw data and added to the tool as a .csv file (see left).

Do not change column names!

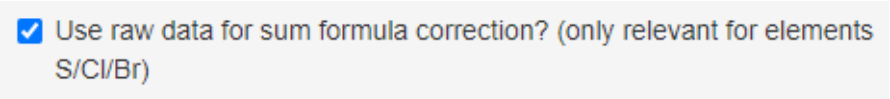
standard – number that helps you keep track of which standards is which (i.e. alkane chain length).

ei rt – retention time in fragment data set (minutes!)

ci rt – retention time in molecular ion data set (minutes!)

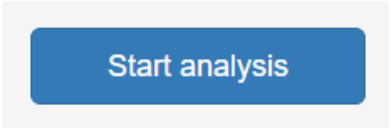


Depending on the performance of the deconvolution software used, it might here and there miss single ions from a spectrum. If it happens to be one of adducts/ neutral losses MSdeCIpher searches for, you will lose a potential molecular ion candidate because of one erroneous deconvolution. Because of that, MSdeCIpher offers the option to check raw data instead of deconvoluted input data for adducts/ neutral losses. It increases computation time but offers more accurate results. Only .mzXML files in centroid format are supported. Check http://proteowizard.sourceforge.net/ or similar tools for more information on how to convert your files to .mzXML.



Elements like S, Cl and Br appear less often in actual compounds than in theoretical sum formulas of ions (if you put them in the element constraints). However, the latter are easily falsified by checking the distinct isotopic peaks of elements in question. This option enables a simple check in the raw data if distinct M+2 isotopic peaks of S, Cl or Br are present. **This option is only useful for MS that have sufficient resolution to distinguish between the isotopic peaks of different elements (Orbitraps or similar)**. Only .mzXML files in centroid

format are supported. Check http://proteowizard.sourceforge.net/ or similar tools for more information on how to convert your files to .mzXML.

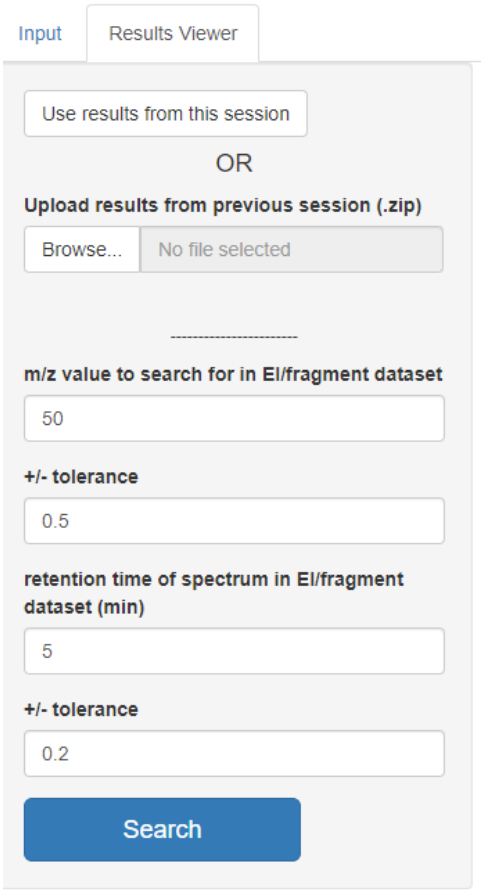


Starts the computation. In the lower right corner, you should see progress reports appearing, giving you an estimate of the overall progress. Once the computation is done, a download button will appear, delivering the result files as a .zip file.

Result Viewer

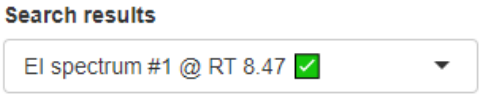
The downloadable output of MSdeCIpher consists of a list of .csv files, each containing one spectrum of the fragment data set together with all potential molecular ions for this spectrum. While this format may be suitable if you wish to feed this output into a subsequent computational pipeline, it is tedious to manually sift through.

Instead, the Result Viewer provides a graphical interface with search functions to check for specific results.

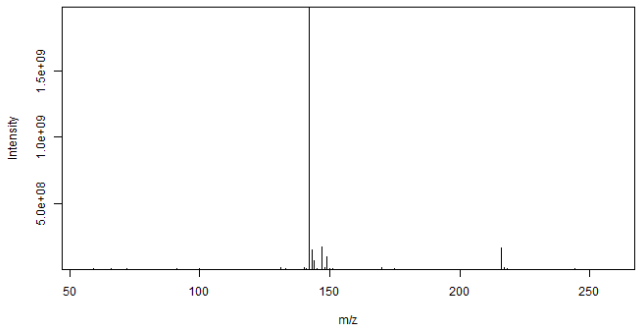


You can either directly use the results from the current session after analysis or you can upload the results from a previous session. In the latter case you must upload the original .zip file as downloaded from the tool.

You can search for results connected to compound of interest. You need the retention time and *m/z* value of one fragment from the fragment spectrum of interest.



A dropdown menu will appear, showing all fragment spectra found which contain an ion matching the search parameters. A green tick means that molecular ion candidates for this fragment spectrum have been found.



Choosing one spectrum from the dropdown menu will display it graphically, allowing quick identification whether this is the spectrum you were looking for.

Possible molecular ions for this EI/fragment spectrum:

Show 10 entries Search:

m/z	retention time	area	spectrum #	sum formula	probability (%)
260.14993	8.45	776147195.2	2 CI	[C12H26NOSSi]+	93.62
218.10278	8.45	213711.66	2 CI	[C9H20NOSSi]+	70.08
188.11024	8.45	196829956.2	2 CI	[C8H18NO2Si]+	57.86
186.09462	8.45	34548286.15	2 CI	[C8H16NO2Si]+	57.1
174.09456	8.46	8942938.3	2 CI	[C7H16NO2Si]+	56.49
191.0919	8.45	10340923.08	2 CI	[C8H19OSSi]+	32.45
163.09698	8.45	940785.64	2 CI	[C6H19OSi2]+	19.77
130.06836	8.47	1258807.8	2 CI	[C5H12NOSi]+	7.84
145.06799	8.46	886648.77	2 CI	[C6H13O2Si]+	7.52
119.05242	8.45	1316249.04	2 CI	[C4H11O2Si]+	5.4

Showing 1 to 10 of 11 entries Previous 1 2 Next

The resulting table will display all potential molecular ion candidates for the chosen fragment spectrum, based on all your criteria. “spectrum #” tells you which deconvoluted spectrum (“pcgroup” in your input data) of the CI/ molecular ion dataset the candidate is part of.

Interpretation of results

MSdeCIpher will find potential molecular ions based on matching retention times between the two datasets. It filters the list based on the intensity and adduct/ neutral loss criteria you define. Then it checks whether the calculated theoretical sum formulas of fragments fit into the theoretical sum formula of the molecular ion candidate. This results in the probability score.

The probability score is calculated by summing up the intensity of all fragment ions (= 100%). Those fragments that have fitting sum formulas now contribute to the probability score with their % intensity. However, bigger fragments are usually more informative than small ones. That is why the contribution by each fragment is additionally weighted by its *m*/*z* value – bigger ions will contribute more probability score than smaller ions relative to their intensity.

Knowing how the probability score is calculated, there are a few caveats that should be considered when evaluating your results:

1) The probability score is highly dependant on the size of the molecular ion candidate. When the candidate is +300 *m*/*z* higher than the biggest fragment in the fragment spectrum, it will nearly always exhibit a high probability score. Simply because the possible sum formula combinations for big ions are so vast, there will always be one that explains most of the fragment sum formulas. Most informative are smaller molecular ion candidates with high probability scores. The closer together molecular ion and biggest fragment *m*/*z* are, the more informative the probability score is.

2) There is, unfortunately, no way to detect if a molecular ion is at all present in the molecular ion dataset. Maybe the compound in question is not ionized at all with the molecular-ion-generating technique. Maybe the compound doesn't stay intact during ionization and what appears to be the molecular ion is just a large fragment. MSdeCIpher has no way of knowing whether this is the case or not. The candidate with the highest score might thus be just a co-eluting compound or a fragment of the actual compound. MSdeCIpher does not provide absolute certainty.
3) Take sum formulas with a grain of salt. While this is dependent on the mass accuracy of the MS used, even with 2 ppm accuracy (Orbitrap or better), sometimes, high scoring molecular ion candidates will have wrong sum formulas. This is especially true the bigger the molecular ion candidate is.
   This is because, somewhere down the line, a fragment got assigned a wrong sum formula and that wrong assignment got transferred to the higher *m*/*z* regions. However, that does not mean that the molecular ion candidate is necessarily false! A wrongly assigned fragment sum formula with a *m*/*z* that is nearly identical to the *m*/*z* of the right sum formula will still lead the way to the right molecular ion. Because of how sum formulas work (elemental building blocks), the right molecular ion will have a possible (wrong) sum formula that fits to the (wrong) sum formulas of the fragments.

   tl;dr: Take the displayed sum formulas only as a hint, they are less accurate the bigger the ion is. But wrong sum formulas mostly don't influence the probability score.

All in all, be mindful when publishing results obtained with MSdeCIpher. The tool is designed to save you time and provide you with few candidate *m*/*z* values per spectrum for molecular ions. However, to really confirm the assignment, you need to perform additional experiments. Measure standards or perform MS^2 experiments of the molecular ions in question. MSdeCIpher helps you with generating hypotheses about your data – it doesn't validate them.