Open tools for

FTICR-MS data processing

EU-FTICR-MS European infrastructure
End User School - 13 Dec 2022 Lille

M-A Delsuc - IGBMC - Université de Strasbourg

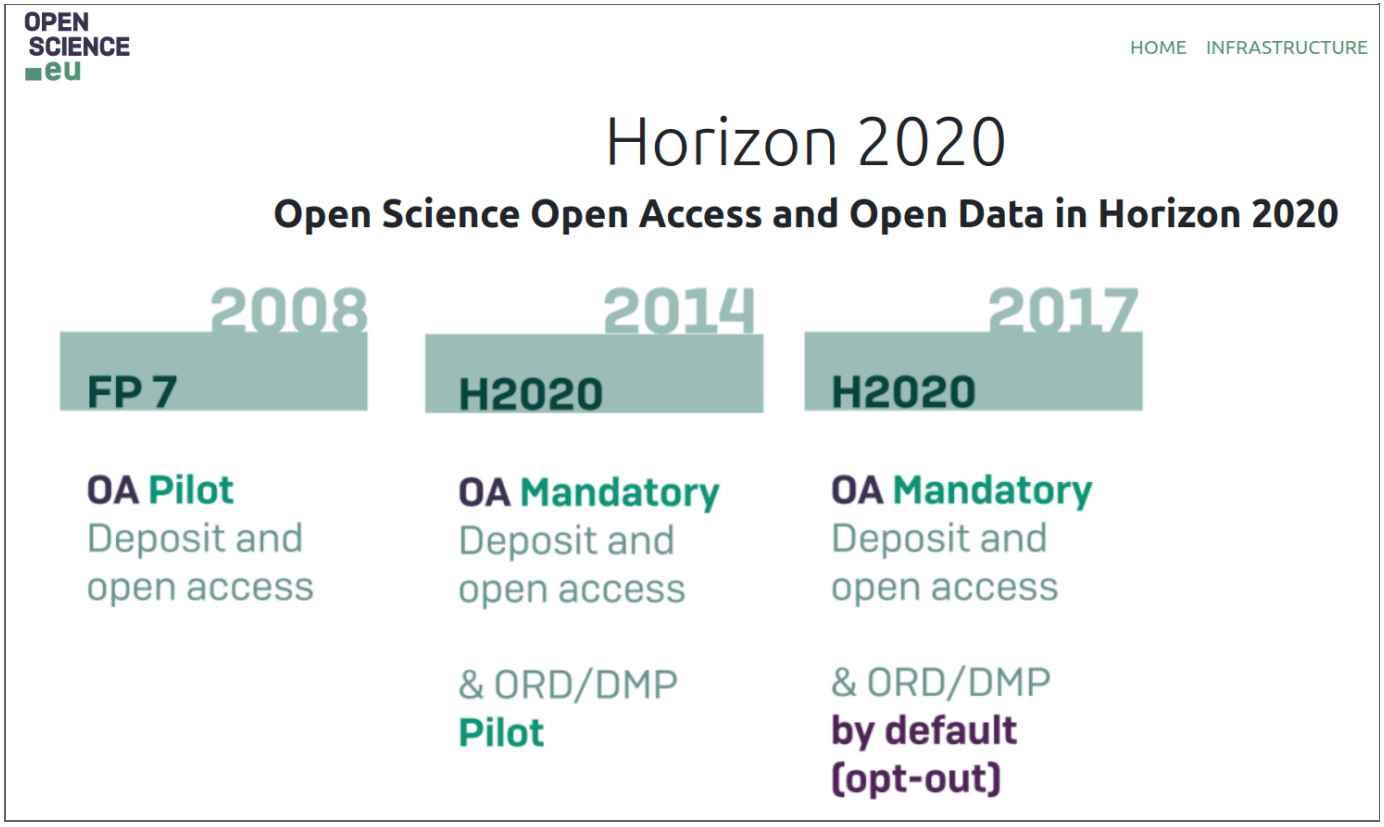
Overview

- Open Science
- Open Data
- Open Source Software
- Data Mining
- Interactive Program
- Larger datasets

Open Science

Open Science

Required by EU:



source: ***openscience.eu***

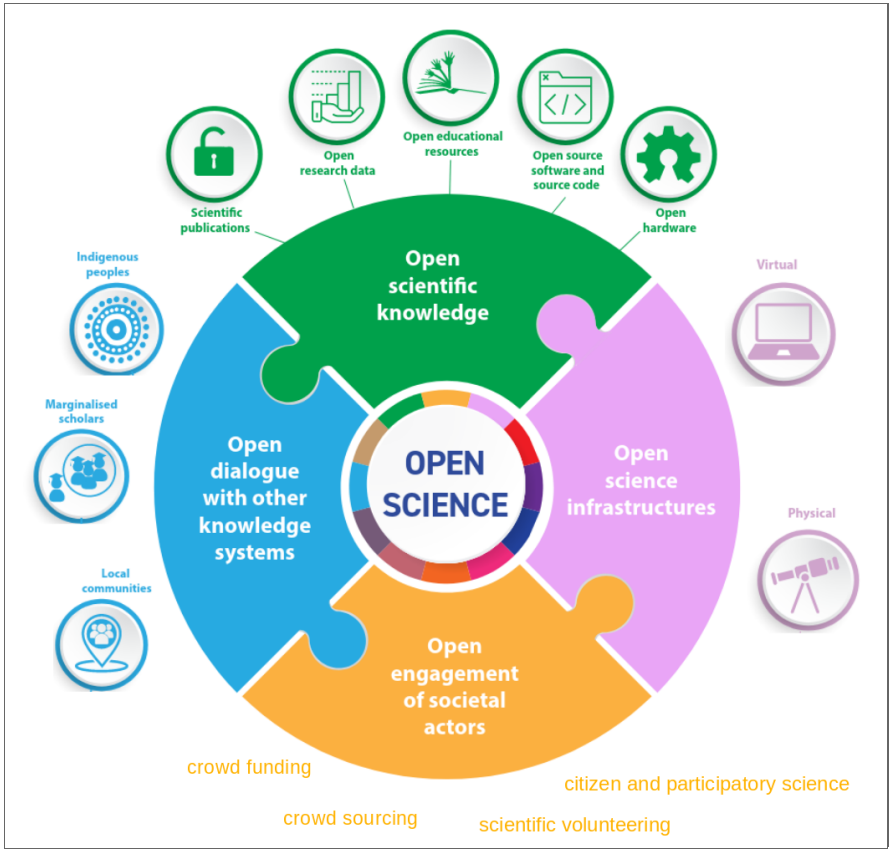
Open Science

and by **UNESCO**:
Recommendation on Open Science adopted by the 41^{st} session of UNESCO General Conference in Nov. 2021

Building on the essential principles of academic freedom, research integrity and scientific excellence, open science sets a new paradigm that integrates into the scientific enterprise practices for reproducibility, transparency, sharing and collaboration resulting from the increased opening of scientific contents, tools and processes.

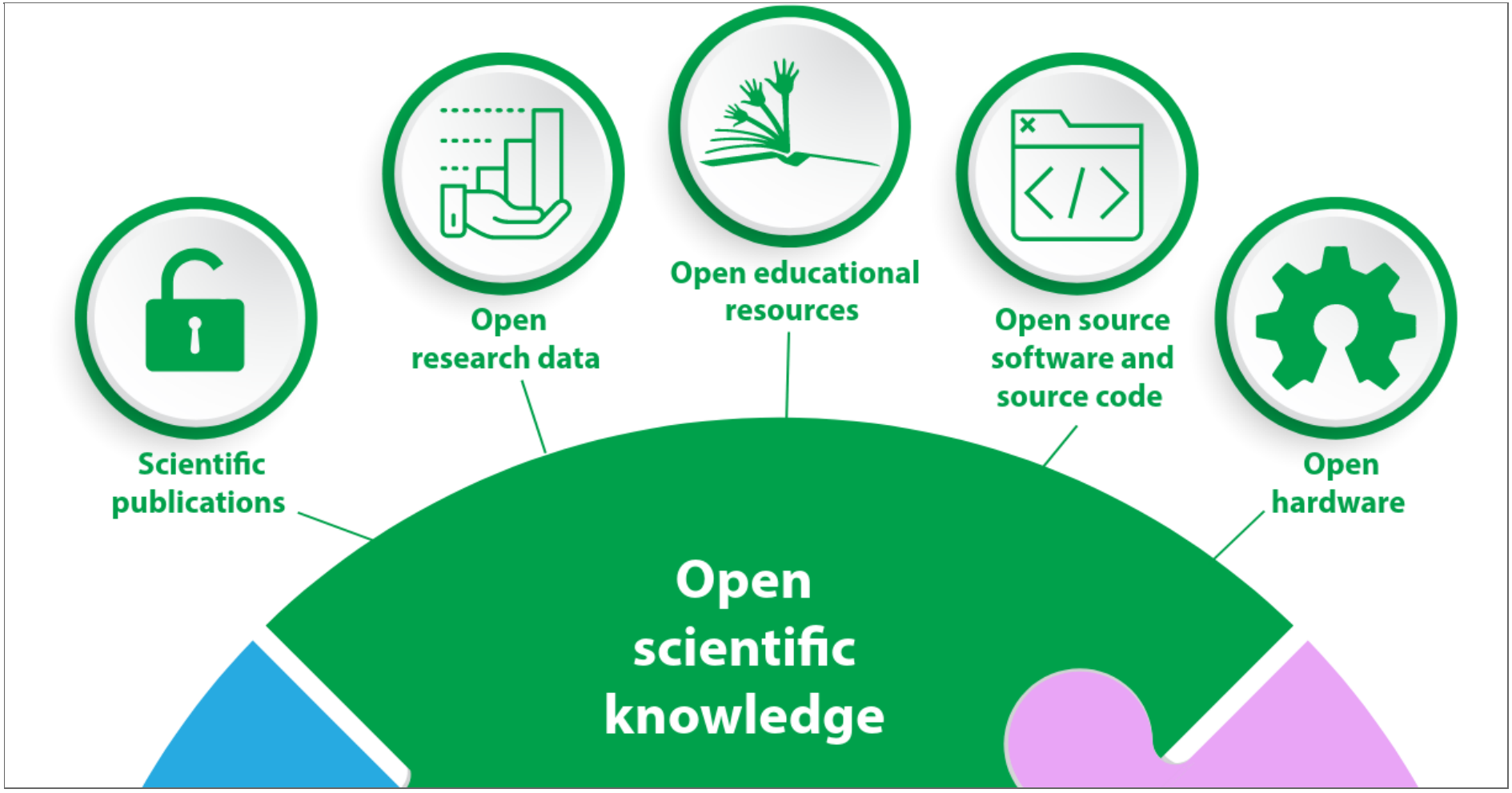
source: www.unesco.org/en/natural-sciences/open-science

Open Science Wheel

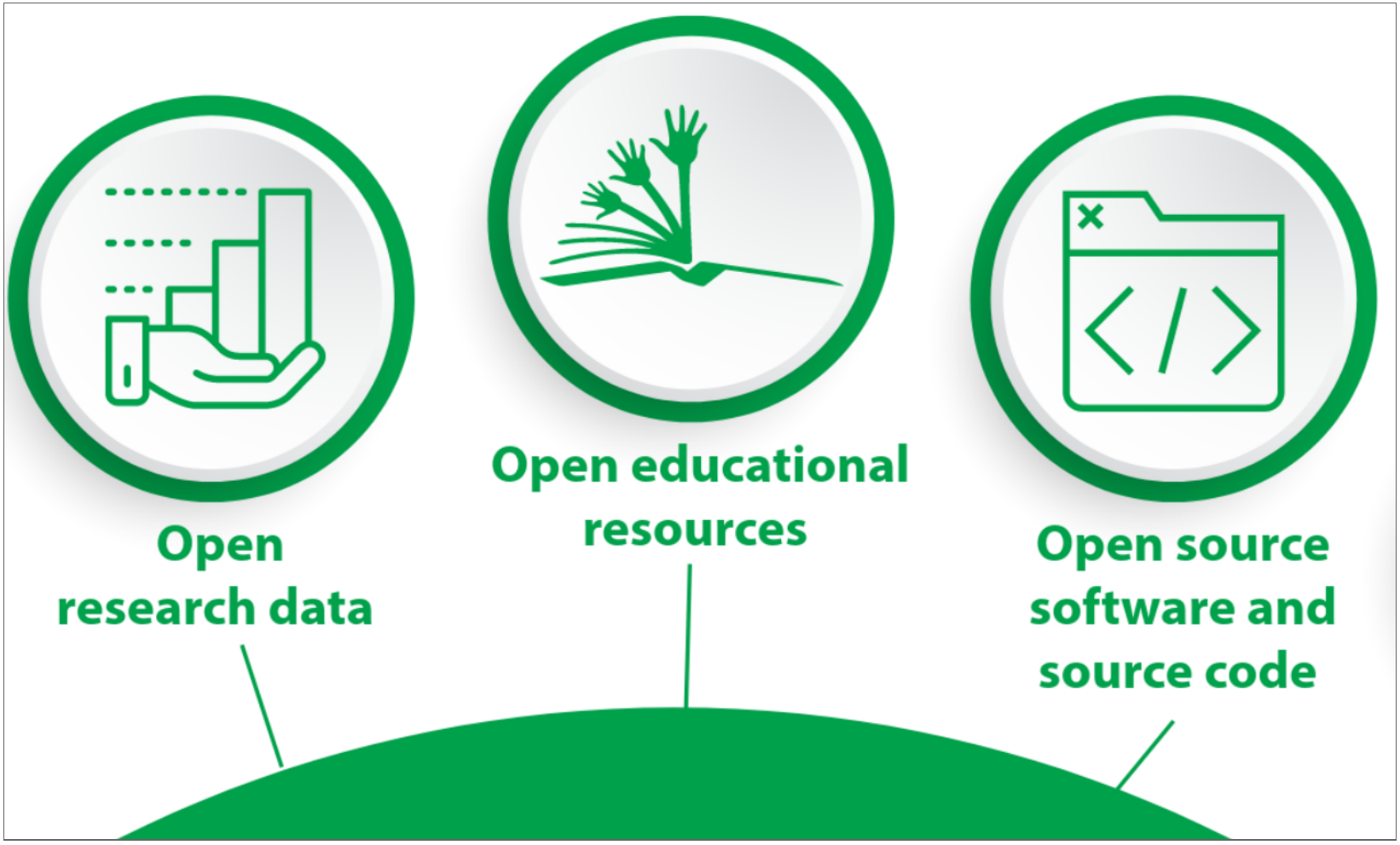


source: www.unesco.org/en/natural-sciences/open-science

Open Science Wheel

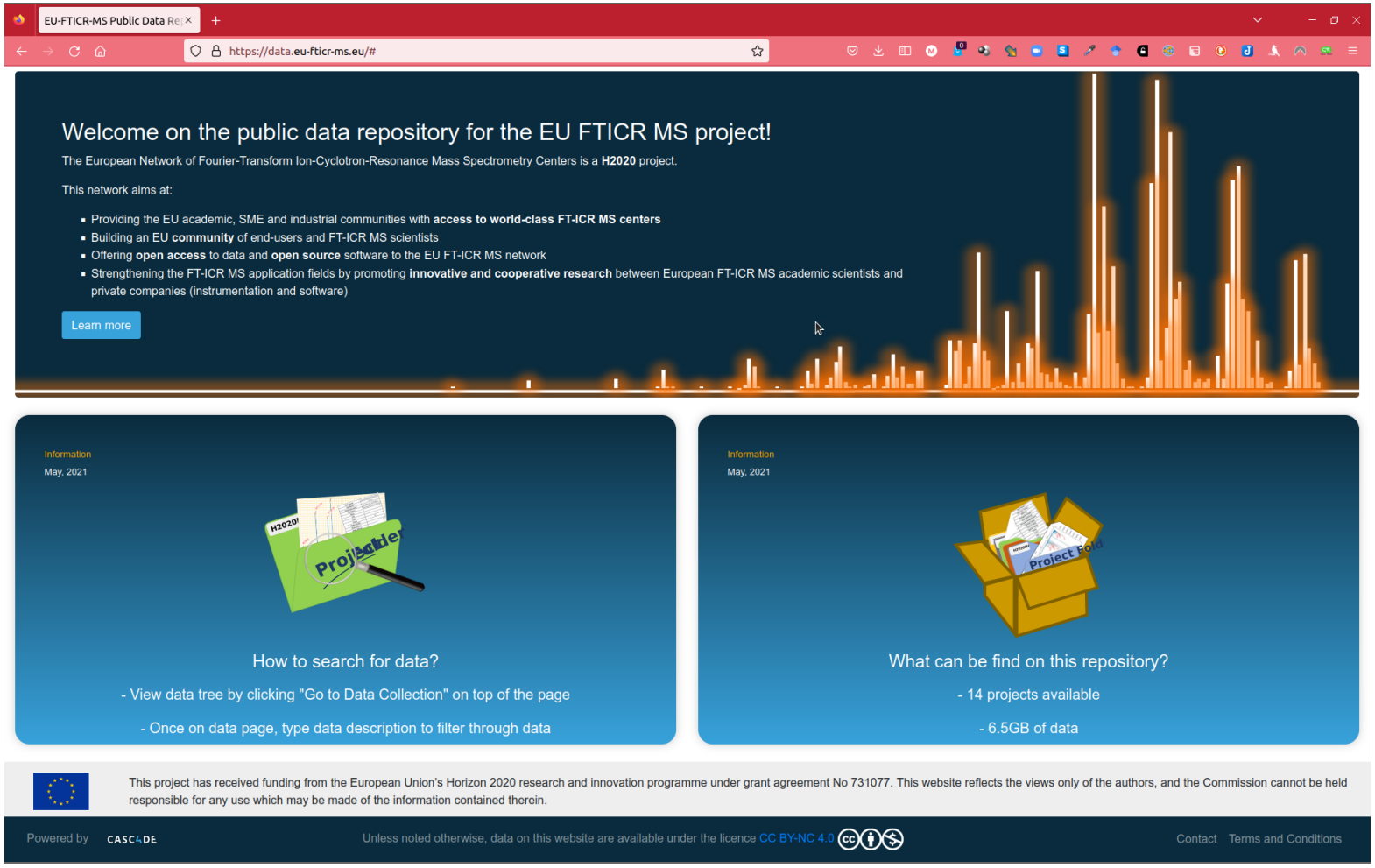


Open Science Wheel

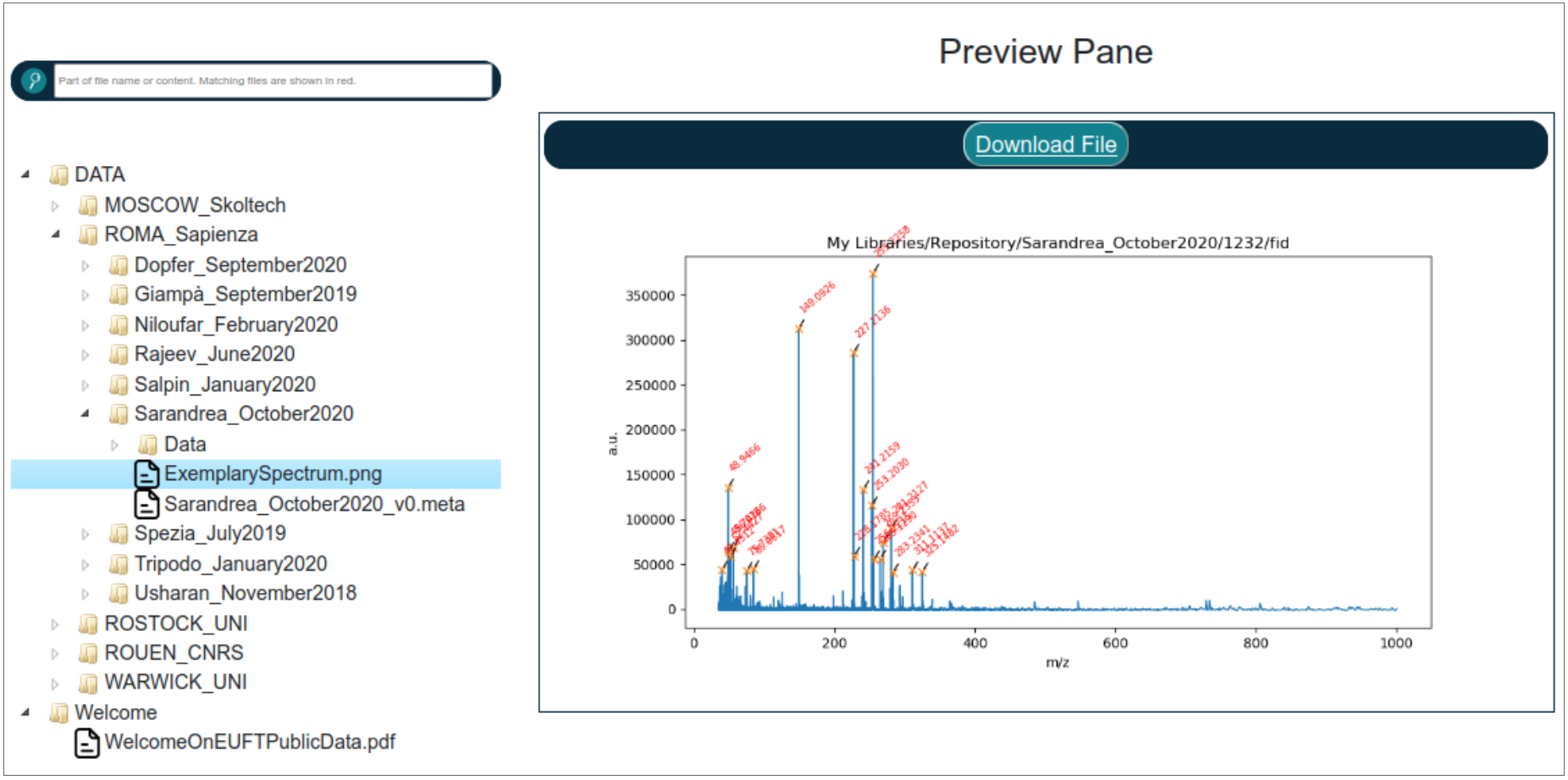


Open Data

EU-FTICR-MS Data repository

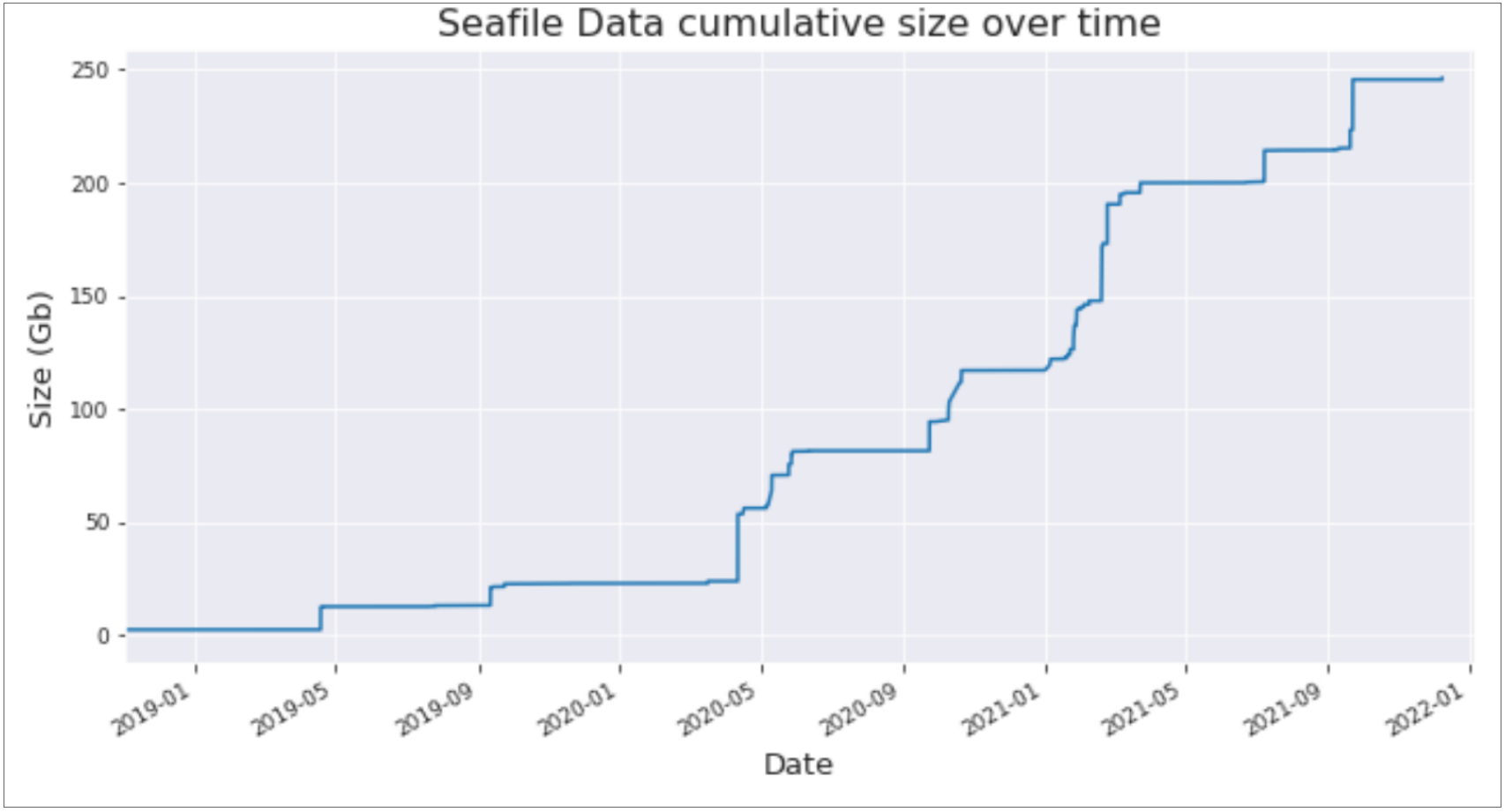


EU-FTICR-MS Data repository



data.eu-fticr-ms.eu

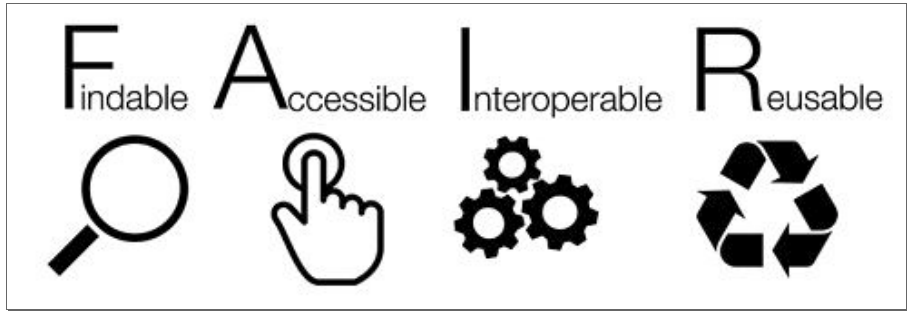
EU-FTICR-MS Data repository



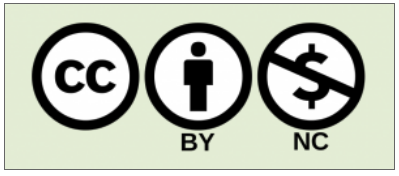
data.eu-fticr-ms.eu

EU-FTICR-MS Data repository

- **Data Management Plan**
  describes repository policy
- **FAIR**
  - Findable
  - Accessible
  - Interoperable
  - Reusable



- **Open**: Data in free access

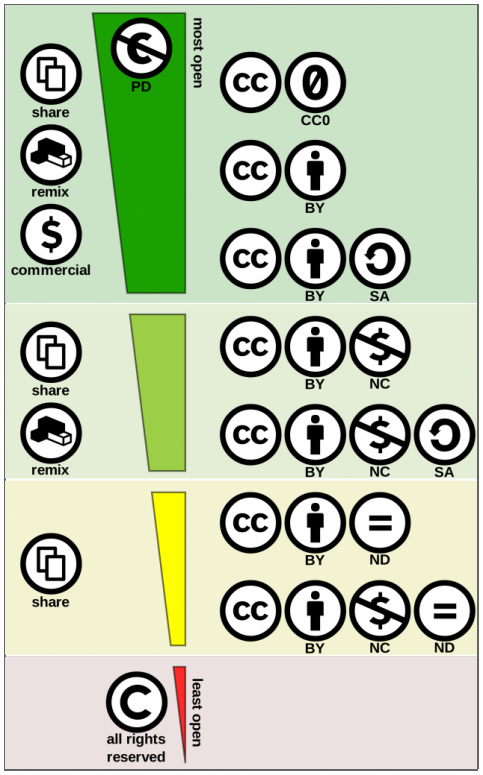


CC 4.0 BY NC

- Open $\neq$ Free

Creative Common licences

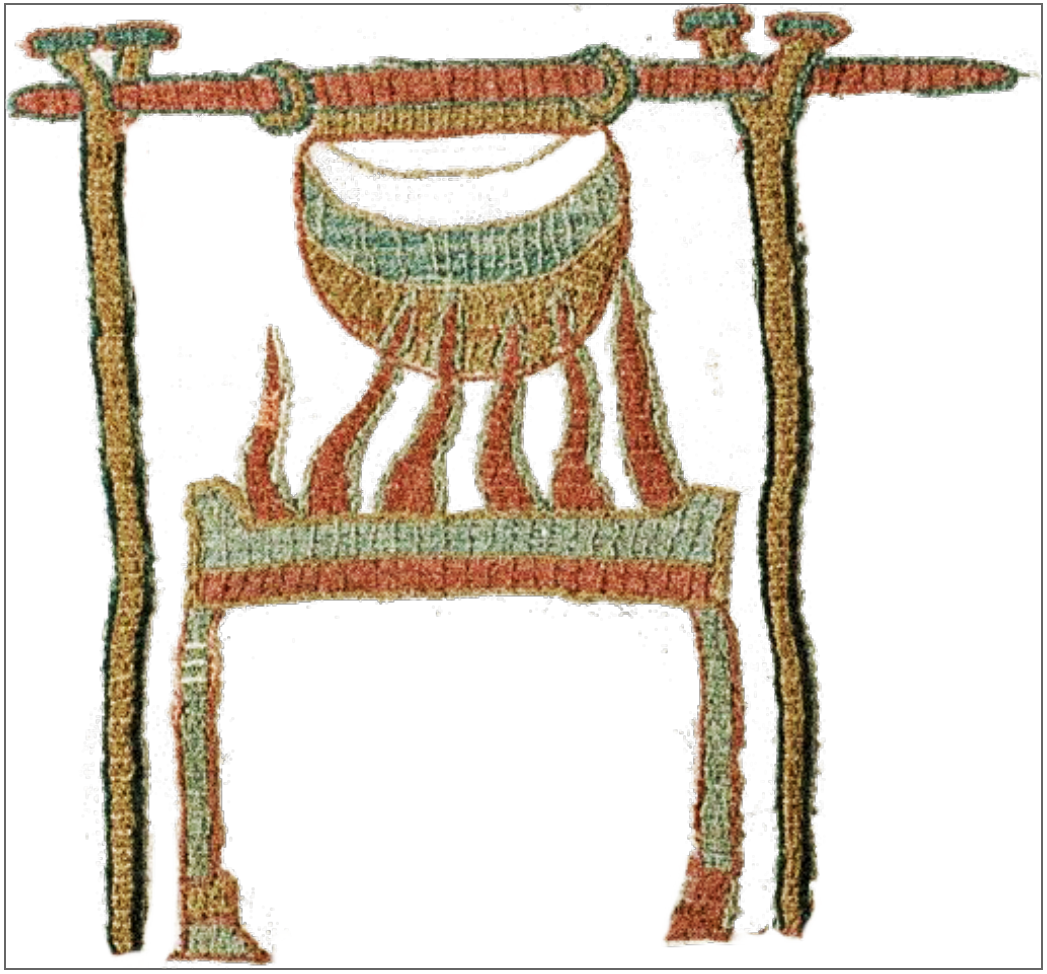
An generic flexible open licence for texts



*source : **creativecommons.org***

Public Domain

Open Source Software

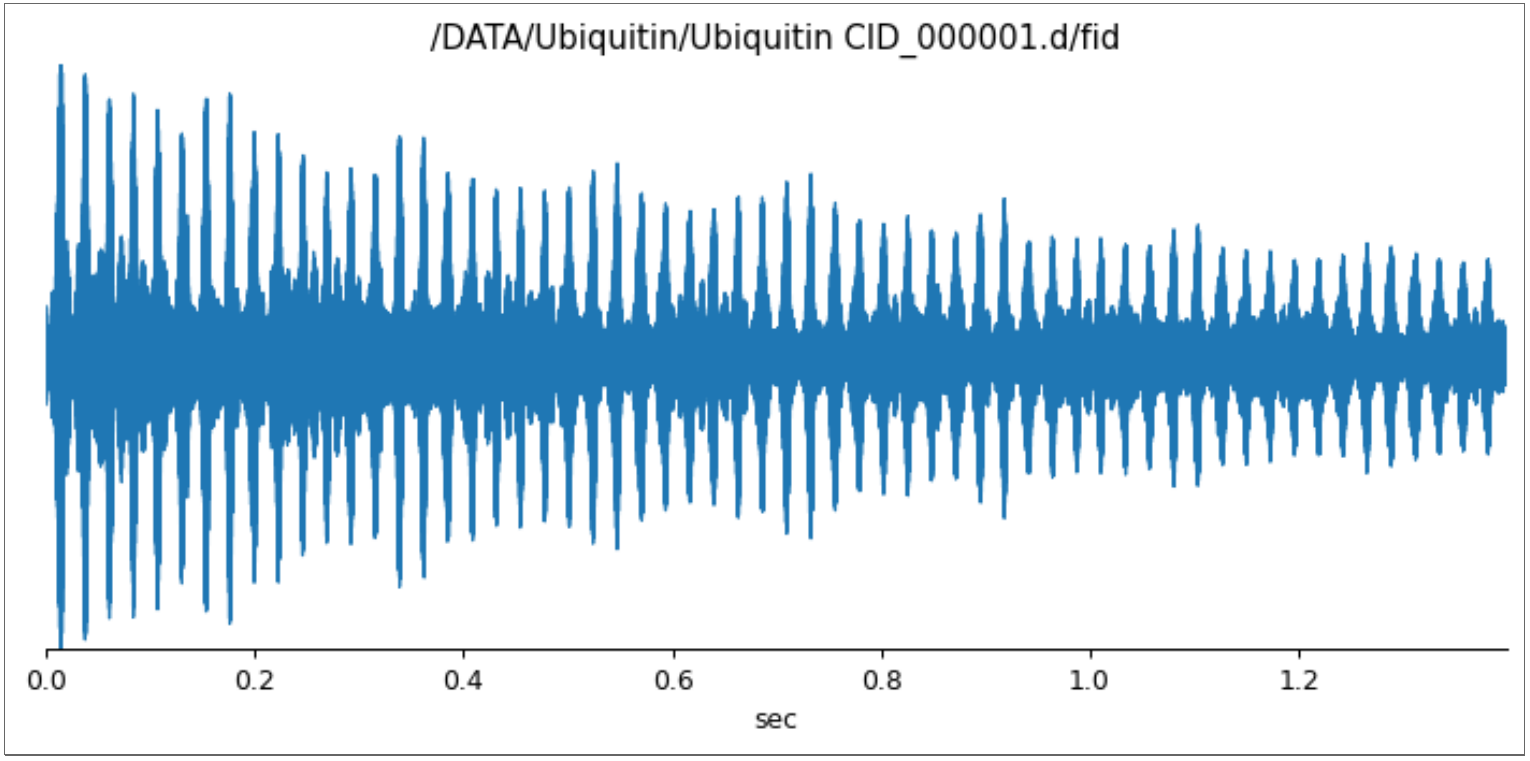


UNESCO Recommendation on Open Science

Building on the essential principles of academic freedom, research integrity and scientific excellence, open science sets a new paradigm that integrates into the scientific enterprise practices for reproducibility, transparency, sharing and collaboration resulting from the increased opening of scientific contents, tools and processes.

reproducibility...

Ion Cyclotronic Resonance-MS is special !
FT-ICR does not measure a Mass Spectrum !
It mesures a **Transient** 😕

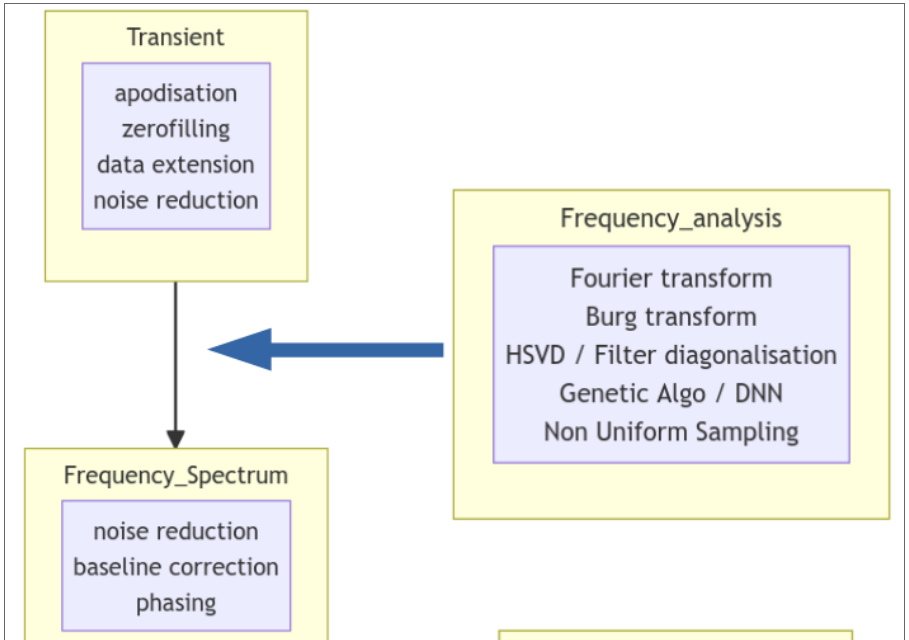


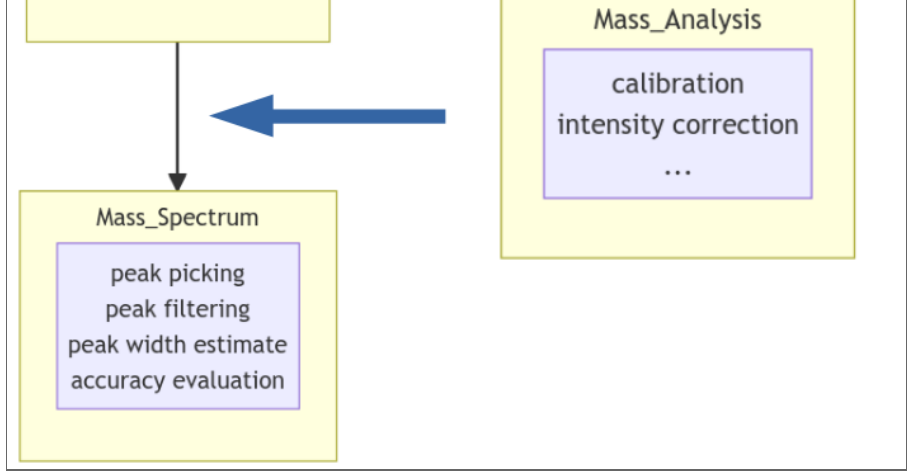
Processing

- **Transient**
- $\rightarrow$ Fourier Transformation
- Frequency Spectrum
- $\rightarrow$ to Mass Spectrum
- **Mass Spectrum**

Data Size

- Peak list << Transient
- Transient < Spectrum

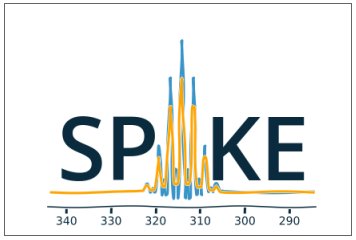



Mass_Analysis
calibration
intensity correction
...
Mass_Spectrum
peak picking
peak filtering
peak width estimate
accuracy evaluation

SPIKE

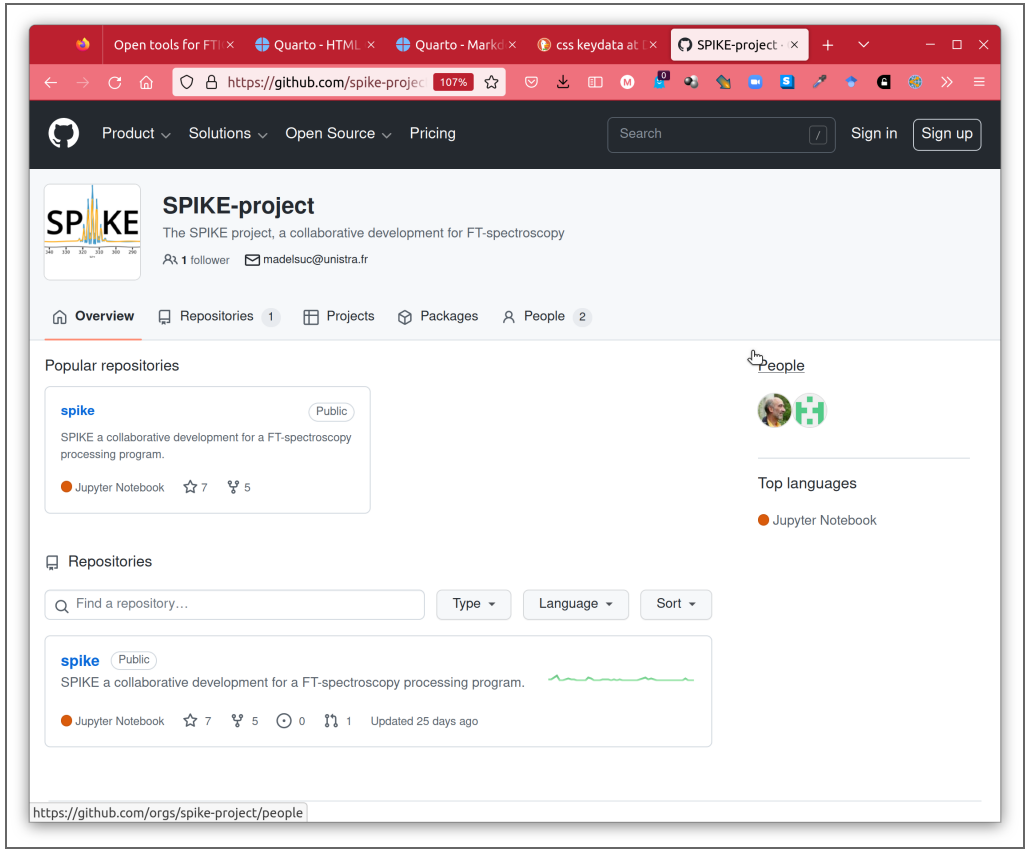
A data-processing library devoted to Fourier spectrometries.

- Nuclear Magnetic Resonance
- FT-ICR Mass Spectrometry
- Orbitrap Mass Spectrometry
- …



SPIKE a processing software dedicated to Fourier spectroscopies
Chiron L., Coutouly M-A., Starck J-P., Rolando C., Delsuc M-A. **(2016) *arXiv* 1608.06777**]

SPIKE



SPIKE is open-source, under GPL licence, distributed on GitHub

SPIKE in details

- written in python
  - multi platform (Windows, MacOS, Linux)
  - based on standard library (`numpy, scipy, ...`)
  - multiprocessing
  - GPU capabilities (`numba`)
  - large files and off-memory processing (`HDF5`)
  - internal compression
- specific methods
  - urQRd noise reduction,
  - Hypercomplex arithmetics
- complete access
  - metadata
  - batch processing
  - some interactive modules

MS & NMR spectrometries !

SPIKE Features

- Transient
  - apodisation
  - zerofilling
  - data extension
  - noise reduction
- Frequency_Spectrum
  - noise reduction
  - baseline correction
  - phasing
- Mass_Spectrum
  - peak picking
  - peak filtering
  - peak width estimate
  - accuracy evaluation
- Frequency_analysis
  - Fourier transform
  - Burg transform

    - *HSVD / Filter diagonalisation*
    - *Genetic Algorithm / DNN*
    - Non Uniform Sampling
- Mass_Analysis
    - calibration
    - *intensity correction*
- More
    - MS series / LC-MS
    - 2D FTICR-MS

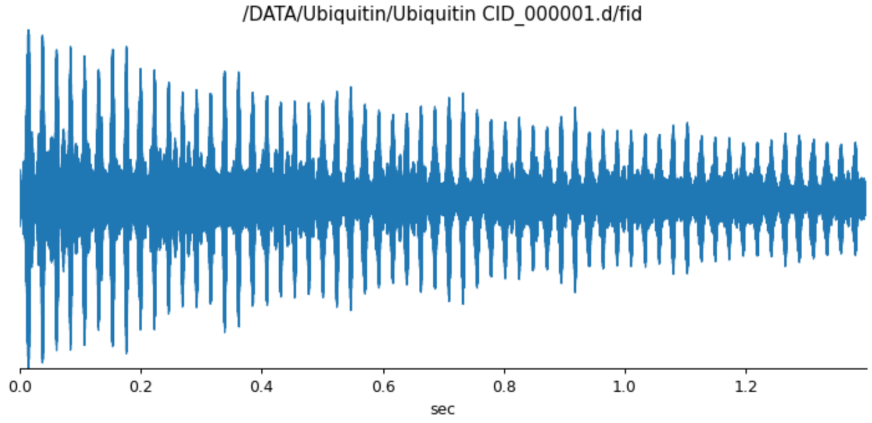
But...

user interface: 🤔

```
print('Reading file ',FC.selected)
d1 = BrukerMS.Import_1D(FC.selected)          # Import_1D creates a SPIKE FTICRData object, from which everything is available
d1.set_unit('sec')                            # it can be acted upon
d1.filename = FC.selected                     # and be extended at will
print(d1)                                     # print() of the dataset shows a summary of the parameters

I.Show1D(d1, title=FC.selected, yratio=1)     # and display  (yratio=1 to have it centered)
# alternatively you can use the low-level tool below:
# d1.display()  # with many options, plus access to matplotlib details
```

```
Reading file  /DATA/Ubiquitin/Ubiquitin CID_000001.d/fid
FTICR data-set
Bo: 15.00
Single Spectrum data-set
FT-ICR axis at 1500.000000 kHz,  4194304 real points,  from physical mz =  153.524   to m/z = 3000.000  R max (M=400) = 1609819
Show1D(children=(VBox(children=(Button(button_style='success', description='Reset', layout=Layout(width='80px'…
```



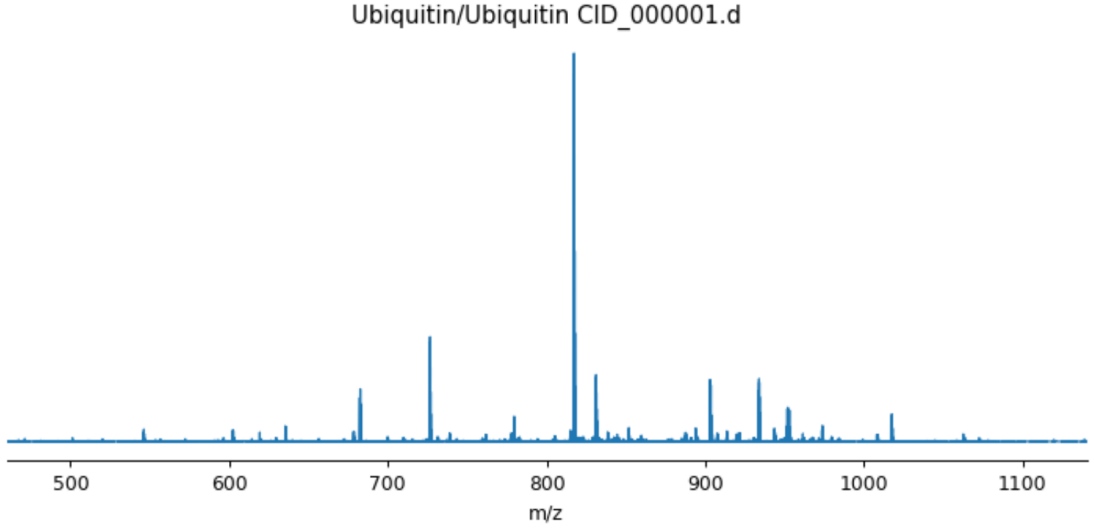
In [5]:

```
D1 = d1.copy()                                # copy the imported data-set to another object for processing
D1.center().kaiser(4).zf(4)   # chaining  centering - apodisation - zerofill
                              # kaiser(4) is an apodisation well adapted to FTICR, slightly more resolution than hamming - try varying th
if D1.axis1.itype == 0:       # means data is real (common case)
    D1.rfft().modulus()           # chaining  real FT - modulus
else:                         #       data is complex, in Narrow-band acquisition
    D1.fft().modulus()            # chaining  complex FT - modulus
D1.bcorr(xpoints=50)          # flatten the baseline
D1.set_unit('m/z')

FI.Show1D(D1, title=FC.nmrname)  #  and display

# D1.display(title=FC.selected_path)  #  alternative, lower level display method with more options
```

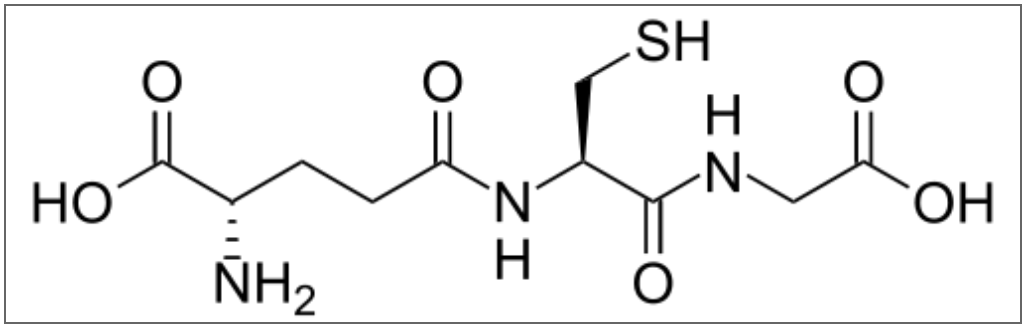
```
Show1D(children=(VBox(children=(Button(button_style='success', description='Reset', layout=Layout(width='80px'…
```

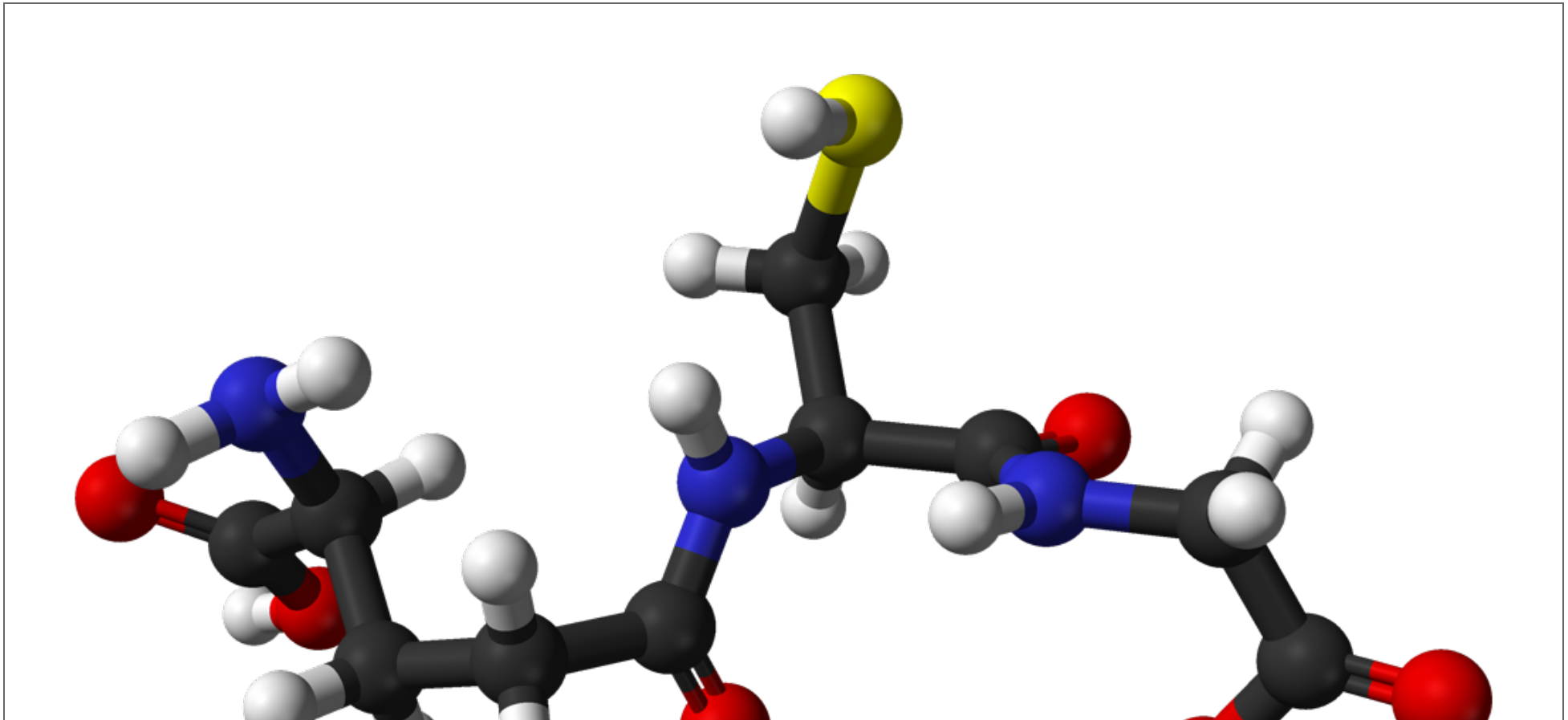


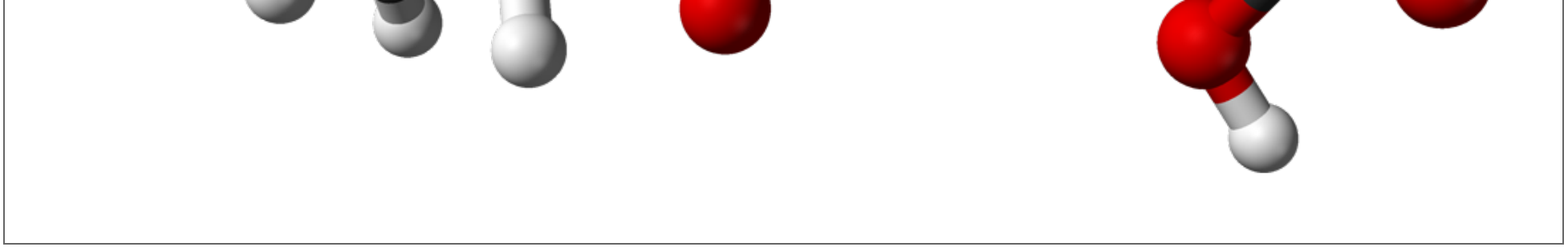
Data Mining

Data Mining

A *Round Robin* test performed over all Instruments in the Network.

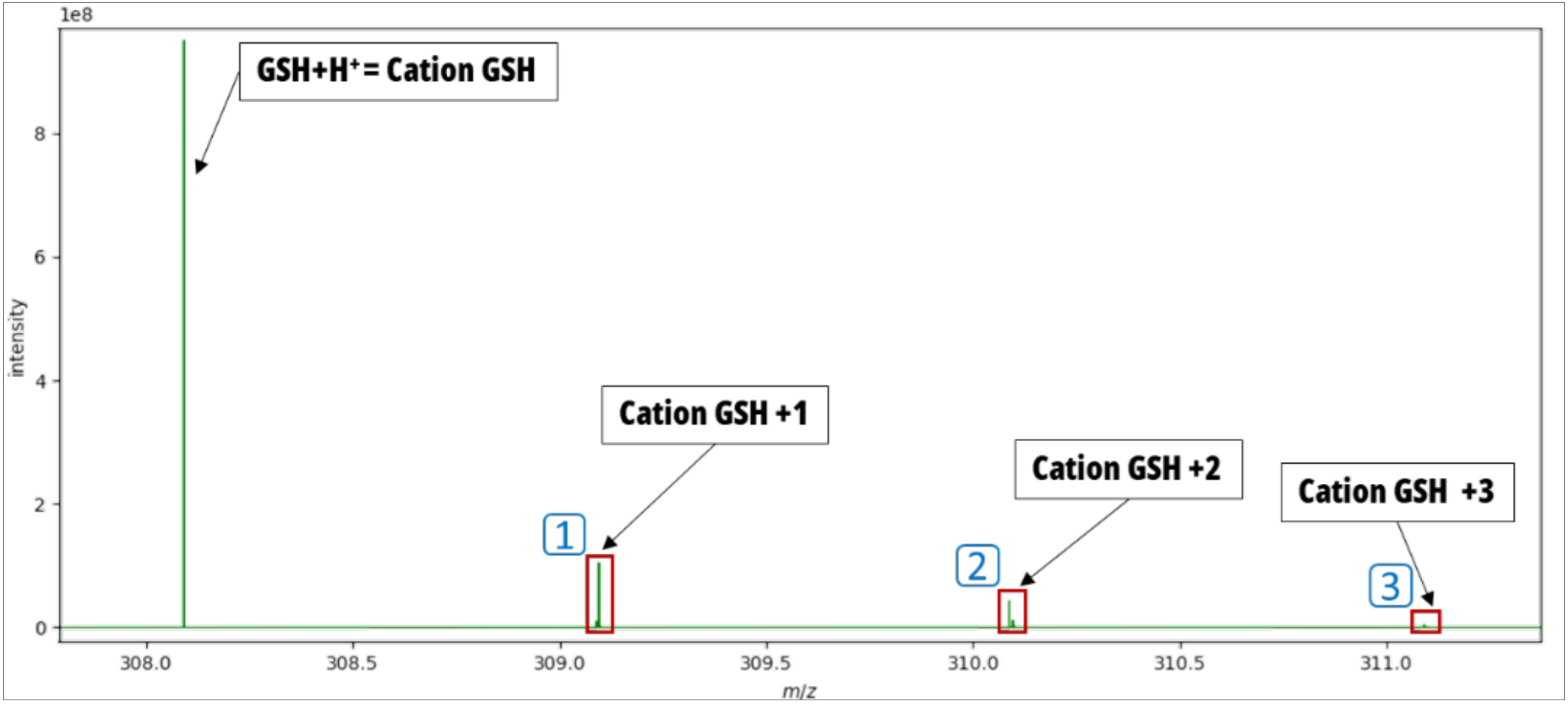




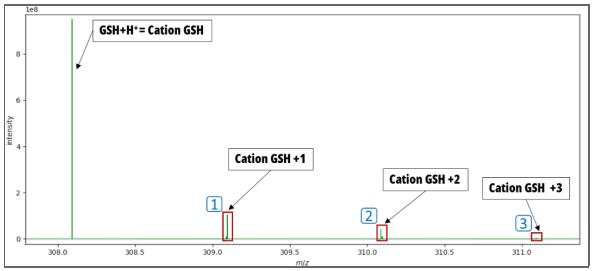


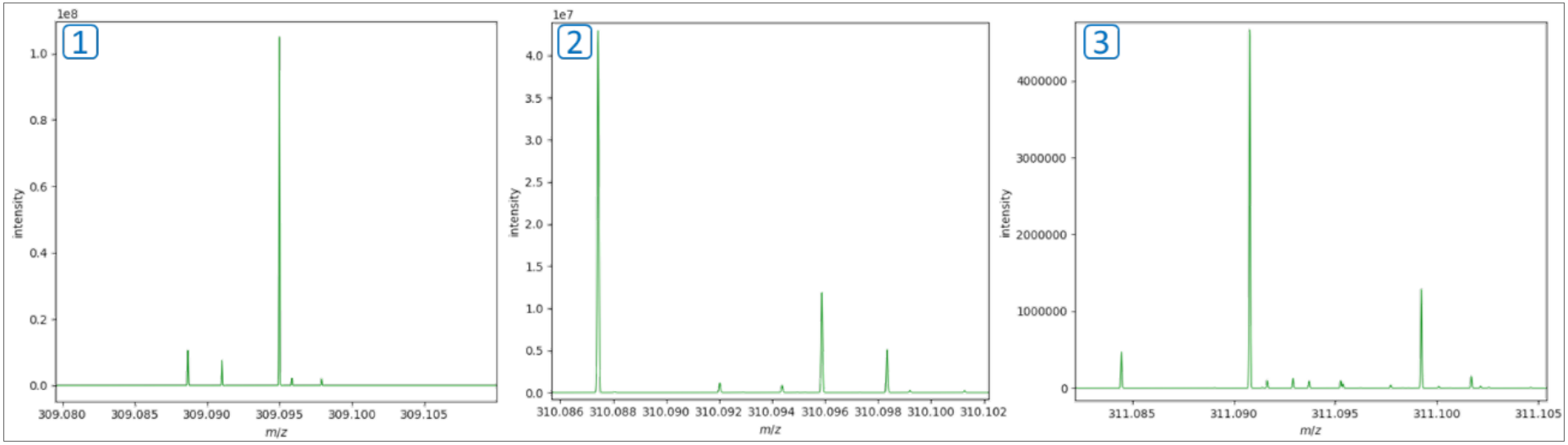
Glutathione molecule

chosen for its fine isotopic pattern



fine isotopic pattern of Glutathione





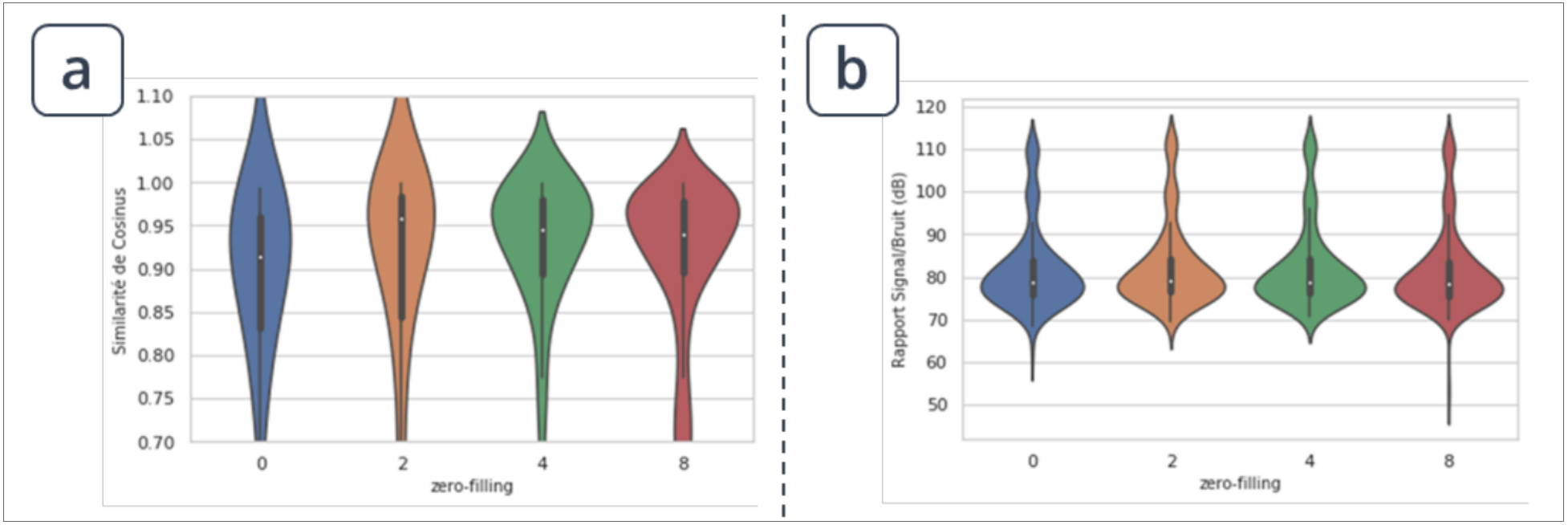
fine isotopic pattern of Glutathione

DATA Mining

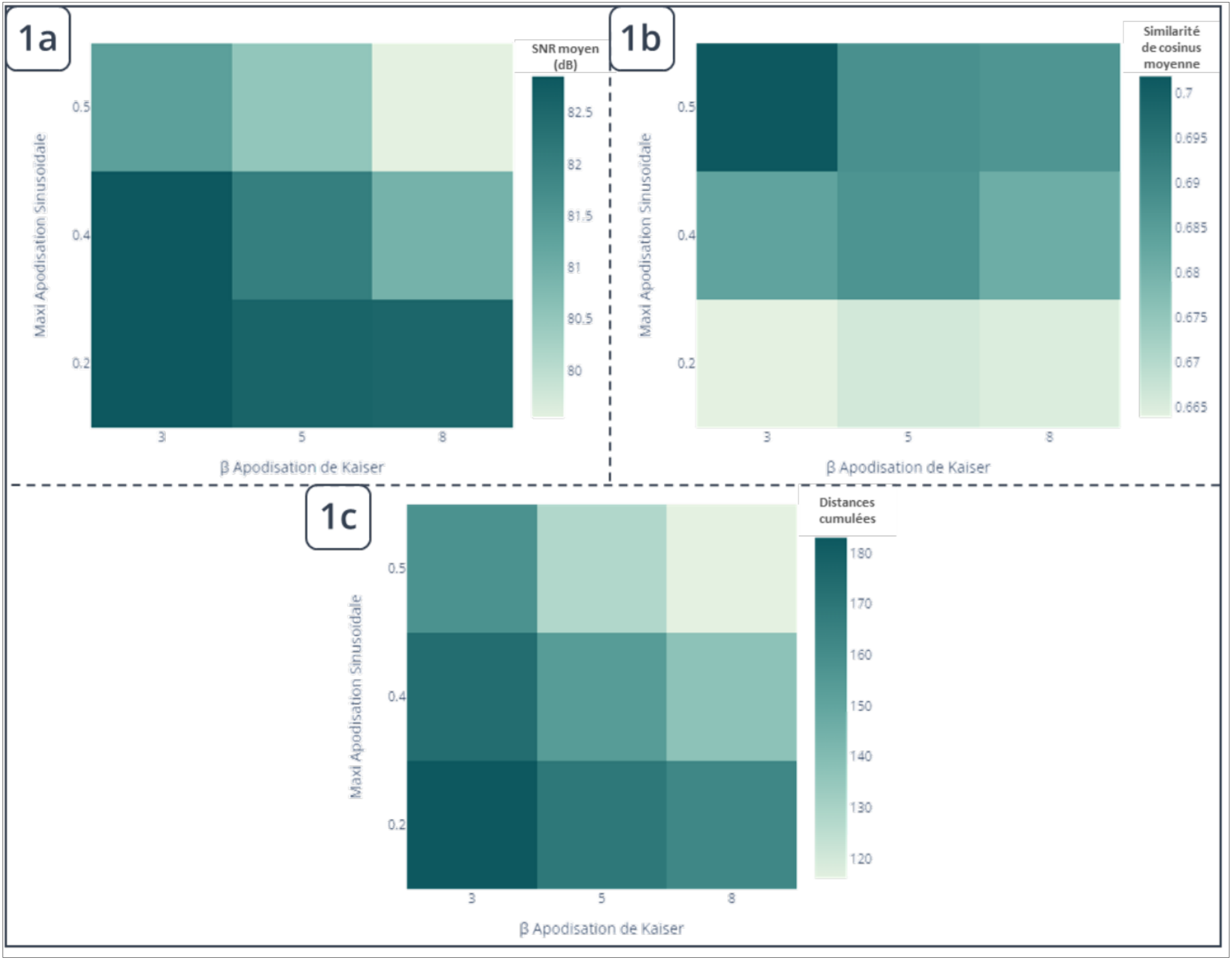
Laura Duciel

Comparison between the various processing parameters

Batch processing of several *different* MS experiment



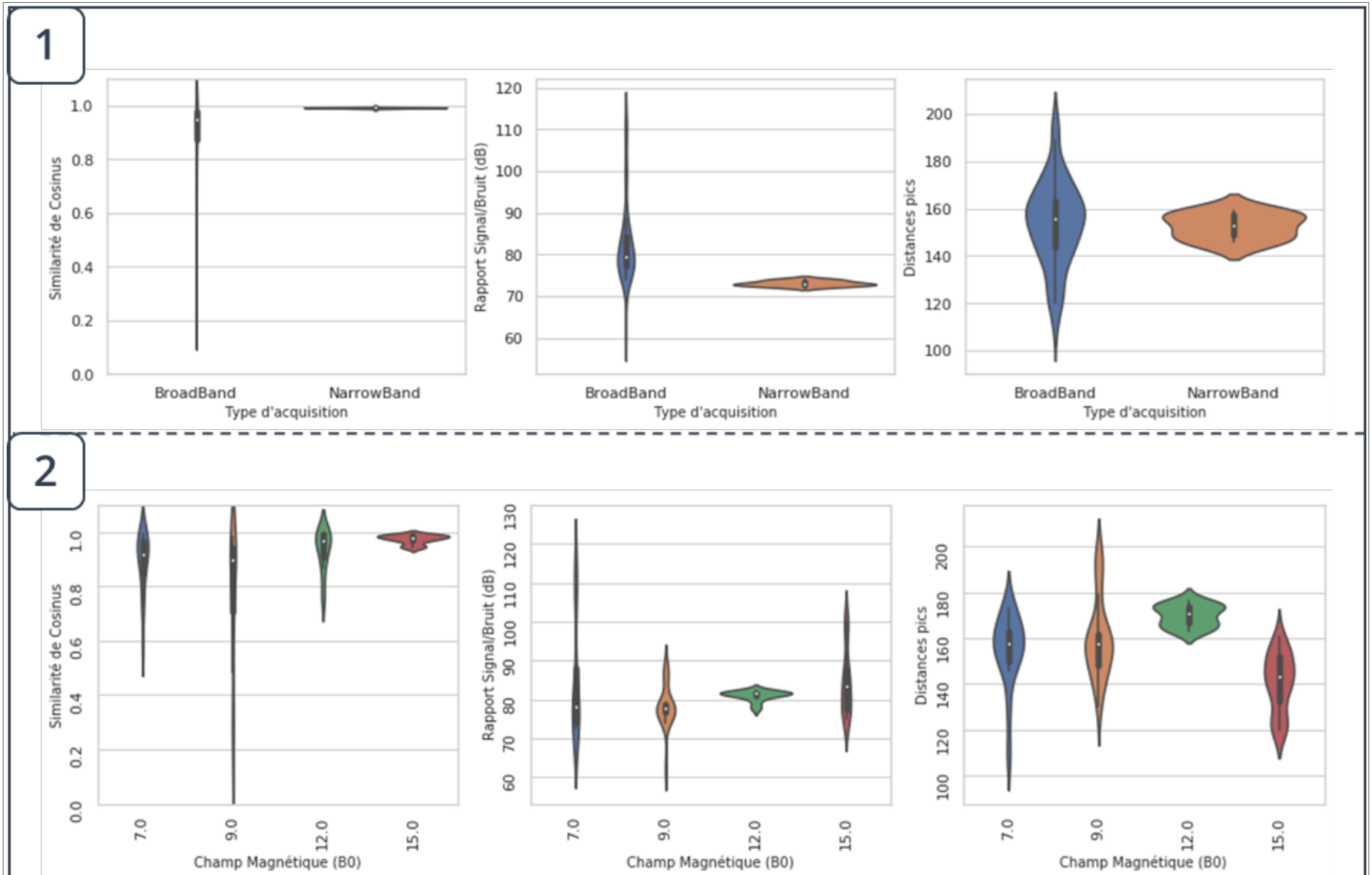
comparing zero-filling
with 2 methods:
(left) cosine difference
(right) SNR

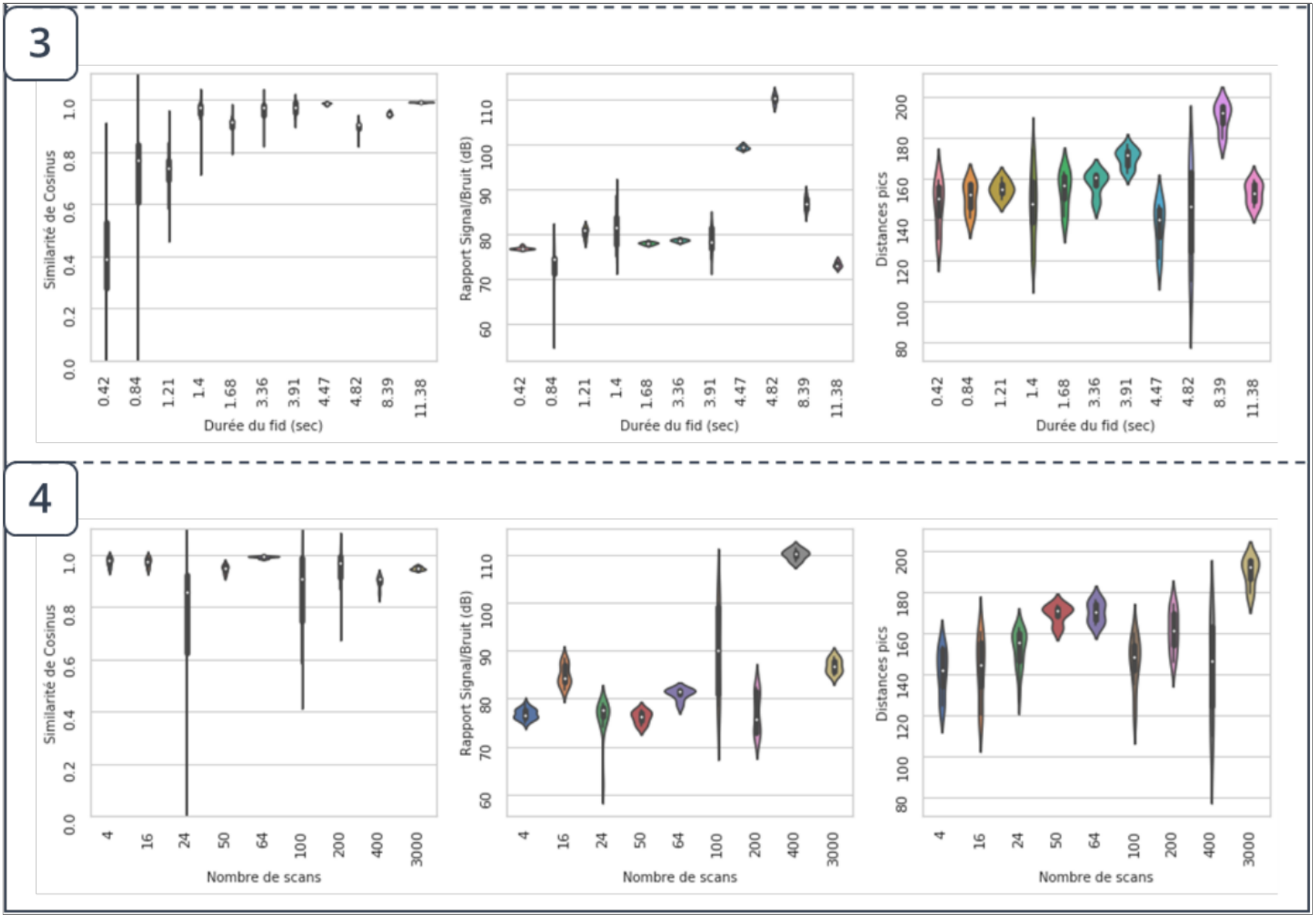

1a
SNR moyen (dB)
82.5
82
81.5
81
80.5
80
Maxi Apodisation Sinusoïdale
0.5
0.4
0.2
3
5
8
β Apodisation de Kaiser
1b
Similarité de cosinus moyenne
0.7
0.695
0.69
0.685
0.68
0.675
0.67
0.665
Maxi Apodisation Sinusoïdale
0.5
0.4
0.2
3
5
8
β Apodisation de Kaiser
1c
Distances cumulées
180
170
160
150
140
130
120
Maxi Apodisation Sinusoïdale
0.5
0.4
0.2
3
5
8
β Apodisation de Kaiser

comparing apodisation

DATA Mining

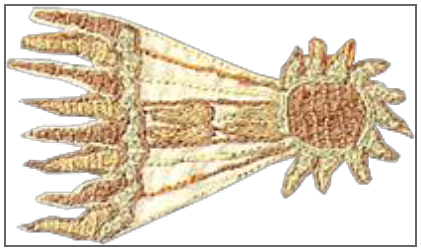
Comparison between the various acquisition optimisation





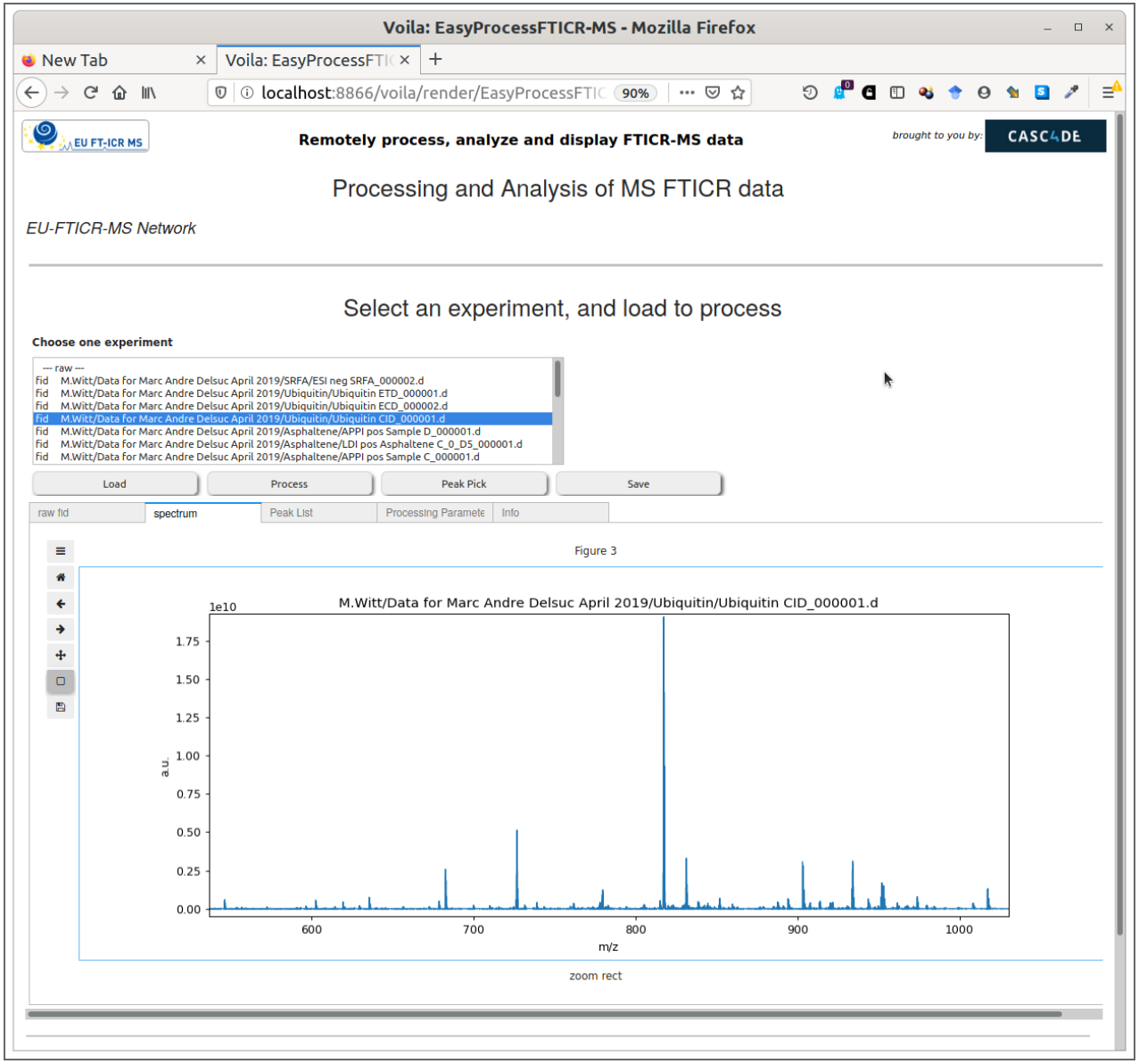
publication in progress

Interactive Program



developped by

presentation



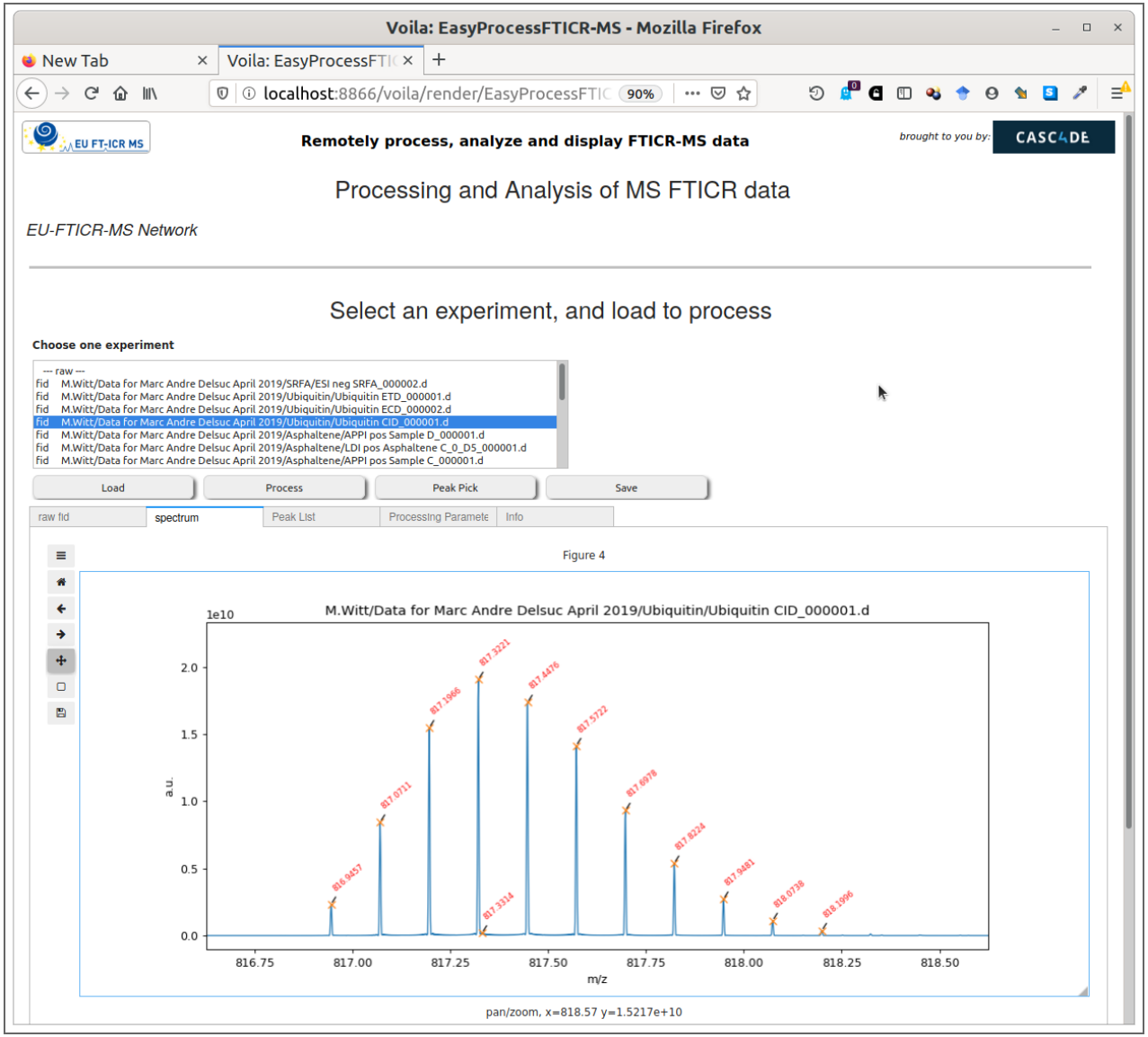
On-line tool

developed for
EU project
used internally
in the network so far

all data-sets accessible

integrated environment

presentation



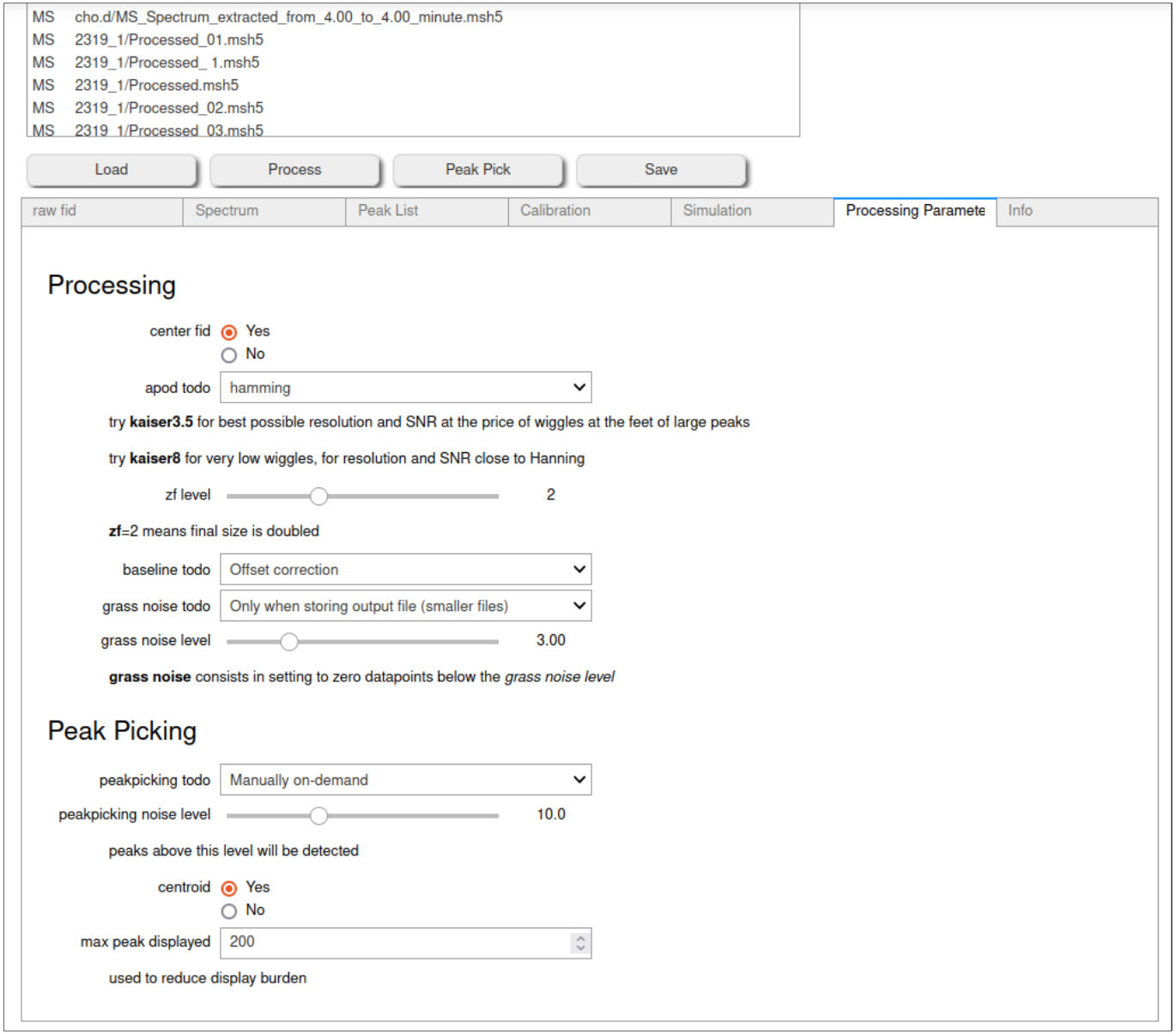
On-line tool developed for

EU project
used internally
in the network so far

all data-sets accessible

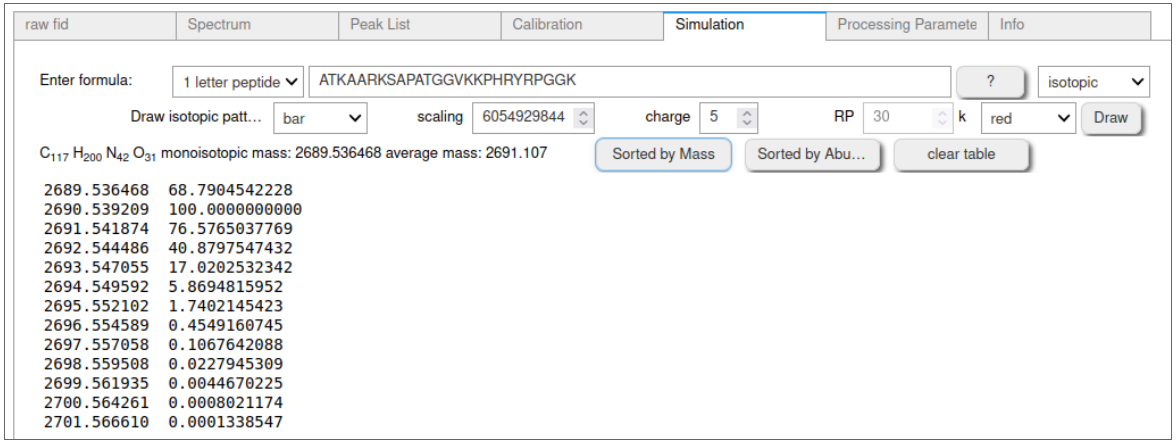
integrated environment

detailed parameters

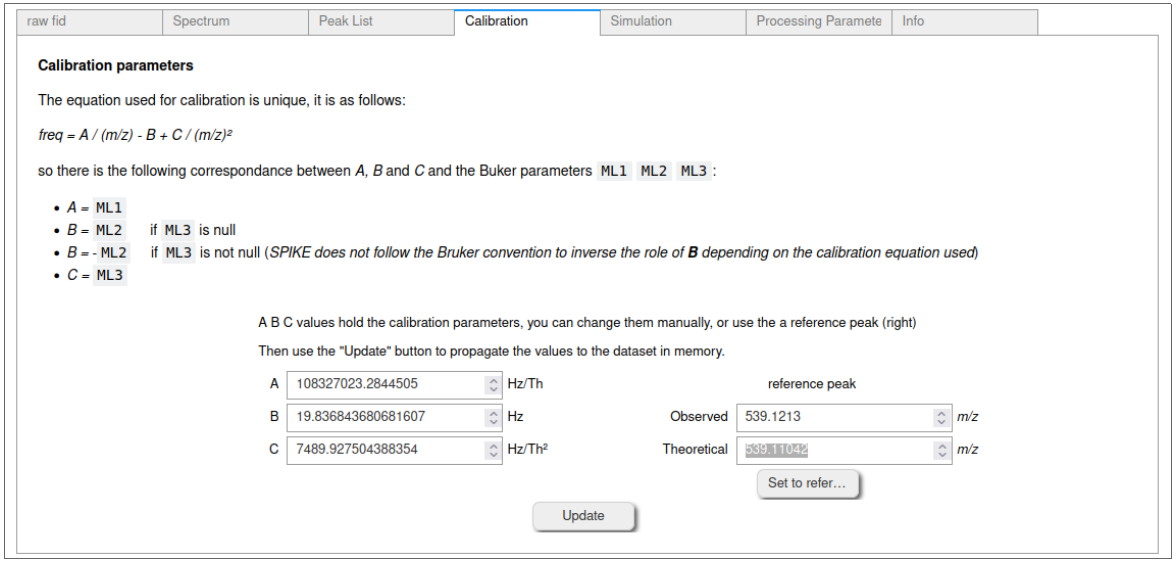


interface kept
simple
optimised
data processing
auto adapt
to dataset type

various possibilities



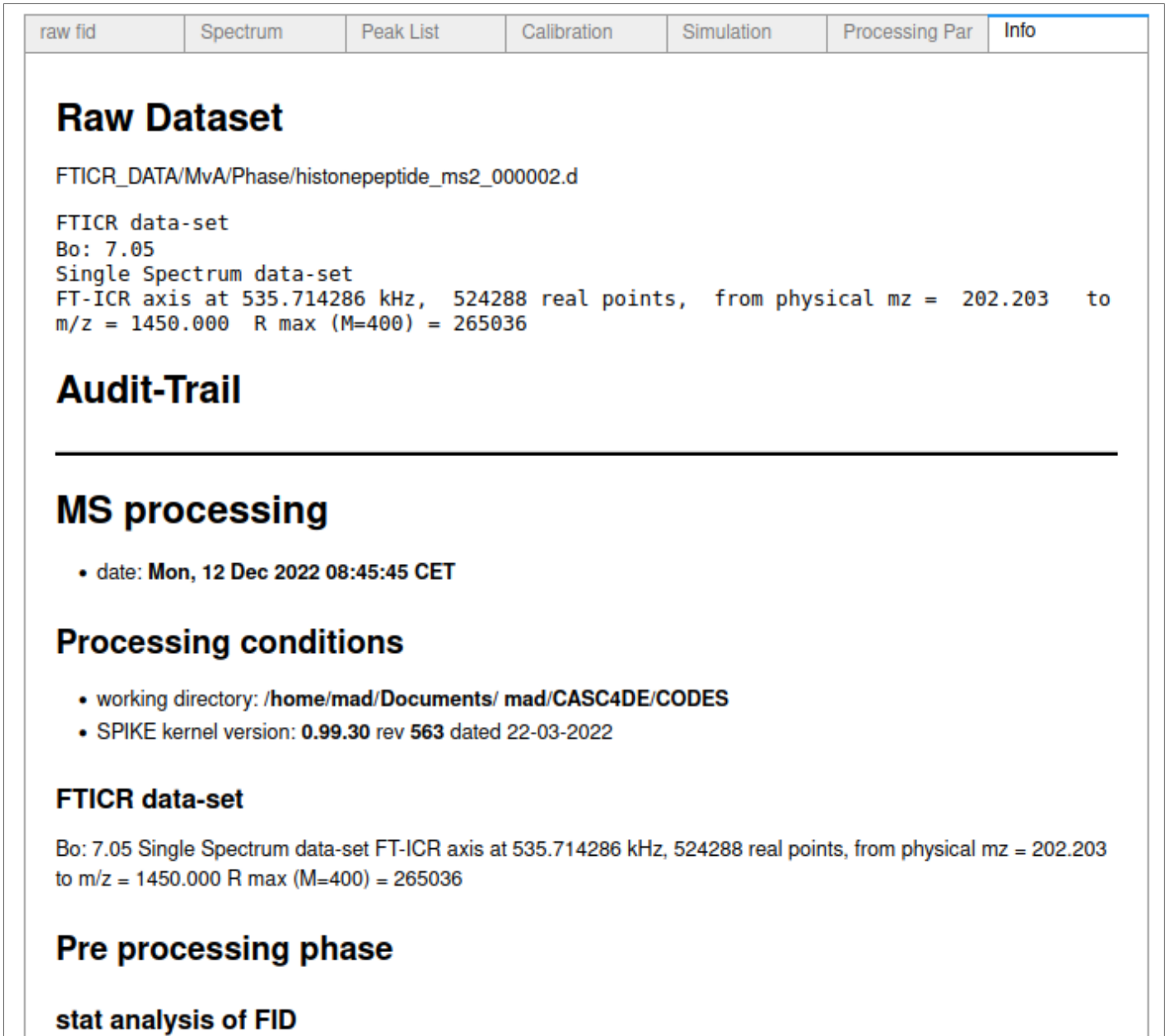
peak assignment



recalibration

audit-trail

All actions to the data are logged



- offset: -6.301655572297774
- noise: 47039.92696468851

vertical shift of FID

- center_fid: Yes
- offset: 6.301655572297774

Spectral Analysis phase

FID apodisation before FT

- apod_todo: kaiser, beta=3.5

Fourier Transform

- initial size: 524288
- zf_level: 2
- Fourier_algo: rfft
- final size: 1048576

Post processing phase

Baseline correction

- mean offset: 1077806.4713130128

final

- modulus: applied
- unit: m/z
- calibration - A: 108327023.2844505
- calibration - B: 19.836843680681607
- calibration - C: 7489.927504388354
- spectrum final size: 524288

  `SHA256: 0558b4161e58c31d2c928f2e9bfec418aa5868cc938272aa24bc22bfbd2580d6`

MS post-processing

- date: **Mon, 12 Dec 2022 08:45:57 CET**

Processing conditions

- working directory: **/home/mad/Documents/ mad/CASC4DE/CODES**
- SPIKE kernel version: **0.99.30** rev **563** dated 22-03-2022

FTICR data-set

Bo: 7.05 Single Spectrum data-set FT-ICR axis at 535.714286 kHz, 524288 real points, from physical mz = 202.203 to m/z = 1450.000 R max (M=400) = 265036

MS post-processing

- date: **Mon, 12 Dec 2022 08:46:07 CET**

Processing conditions

- working directory: **/home/mad/Documents/ mad/CASC4DE/CODES**
- SPIKE kernel version: **0.99.30** rev **563** dated 22-03-2022

FTICR data-set

Bo: 7.05 Single Spectrum data-set FT-ICR axis at 535.714286 kHz, 524288 real points, from physical mz = 202.203 to m/z = 1450.000 R max (M=400) = 265036

Peak-Picking phase

Peak-Picking

- threshold above noise level: 50.0
- zoom: None
- number of detected peaks: 890
- output file:: histonepeptide_ms2_000002.d/peaklist_04.csv

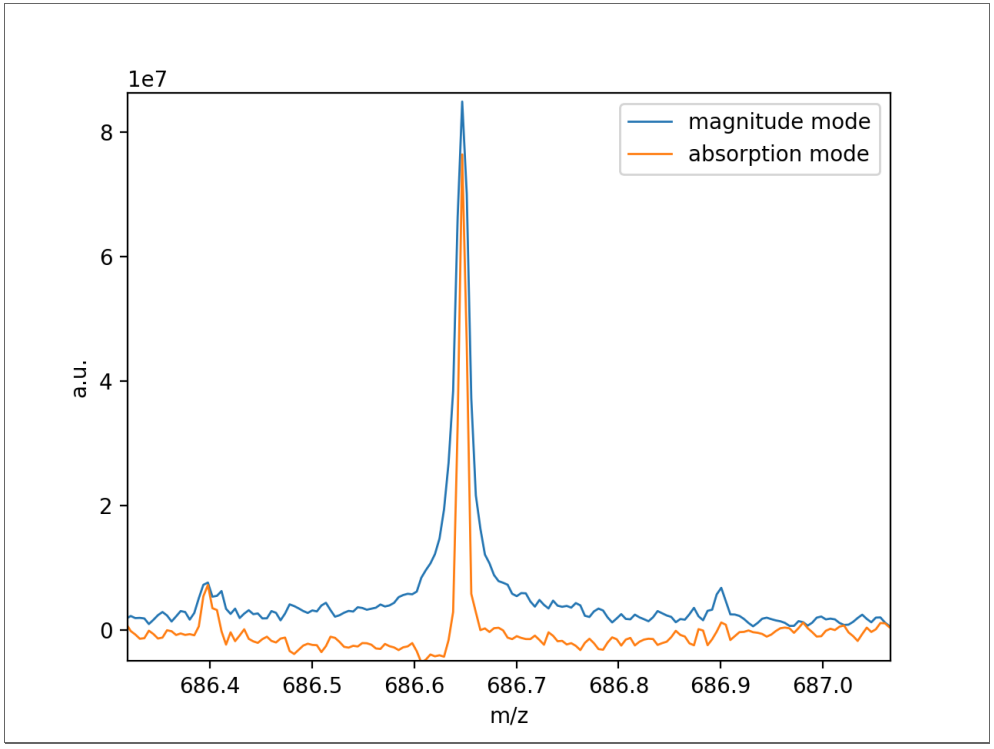
  `SHA256: 3fa94ee05bbcc02f2387de0b27ad7dd087220f254a3dd4f5e78580c4b054f5c7`

The log file is signed, and can be exported.

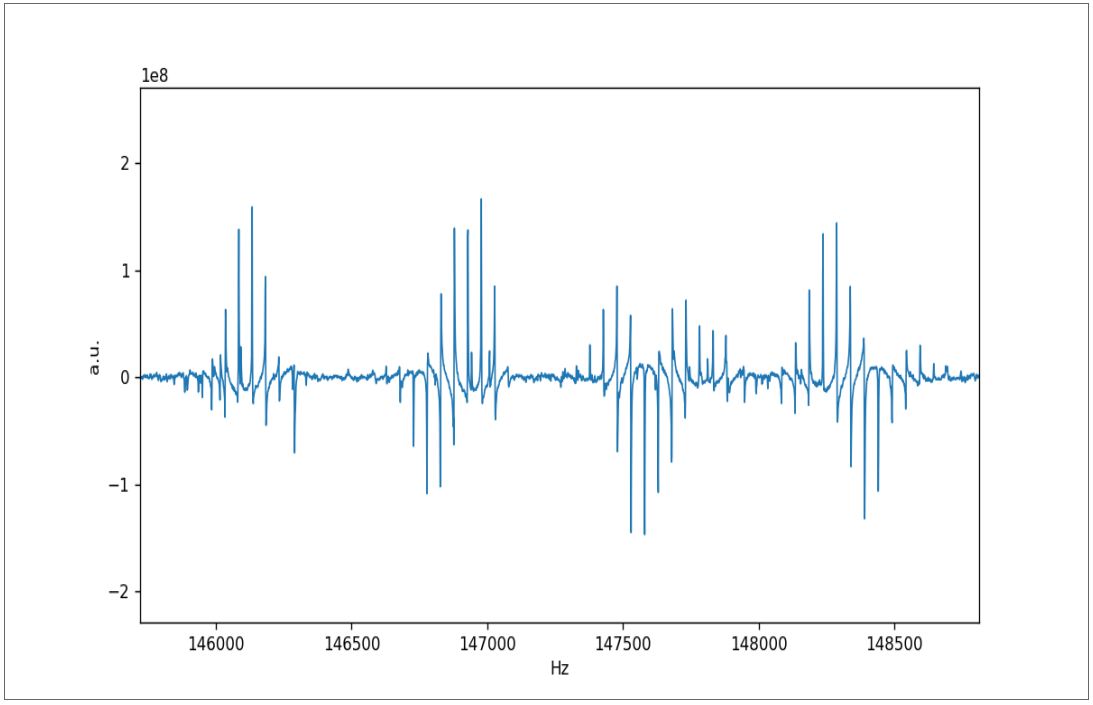
phase sensitive processing

Spectra, presented in “Absorption” mode bring:

- better precision
- better Signal-to-Noise

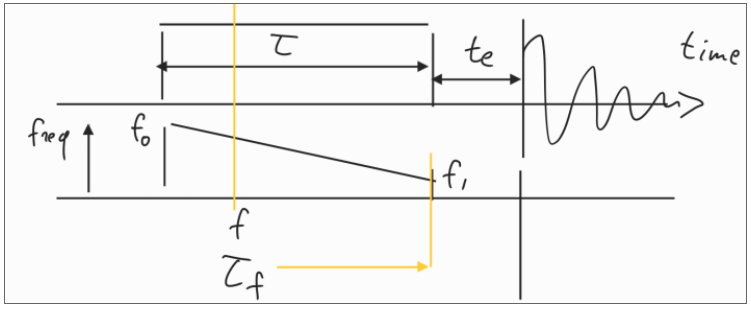


However, they need to be "Phased"



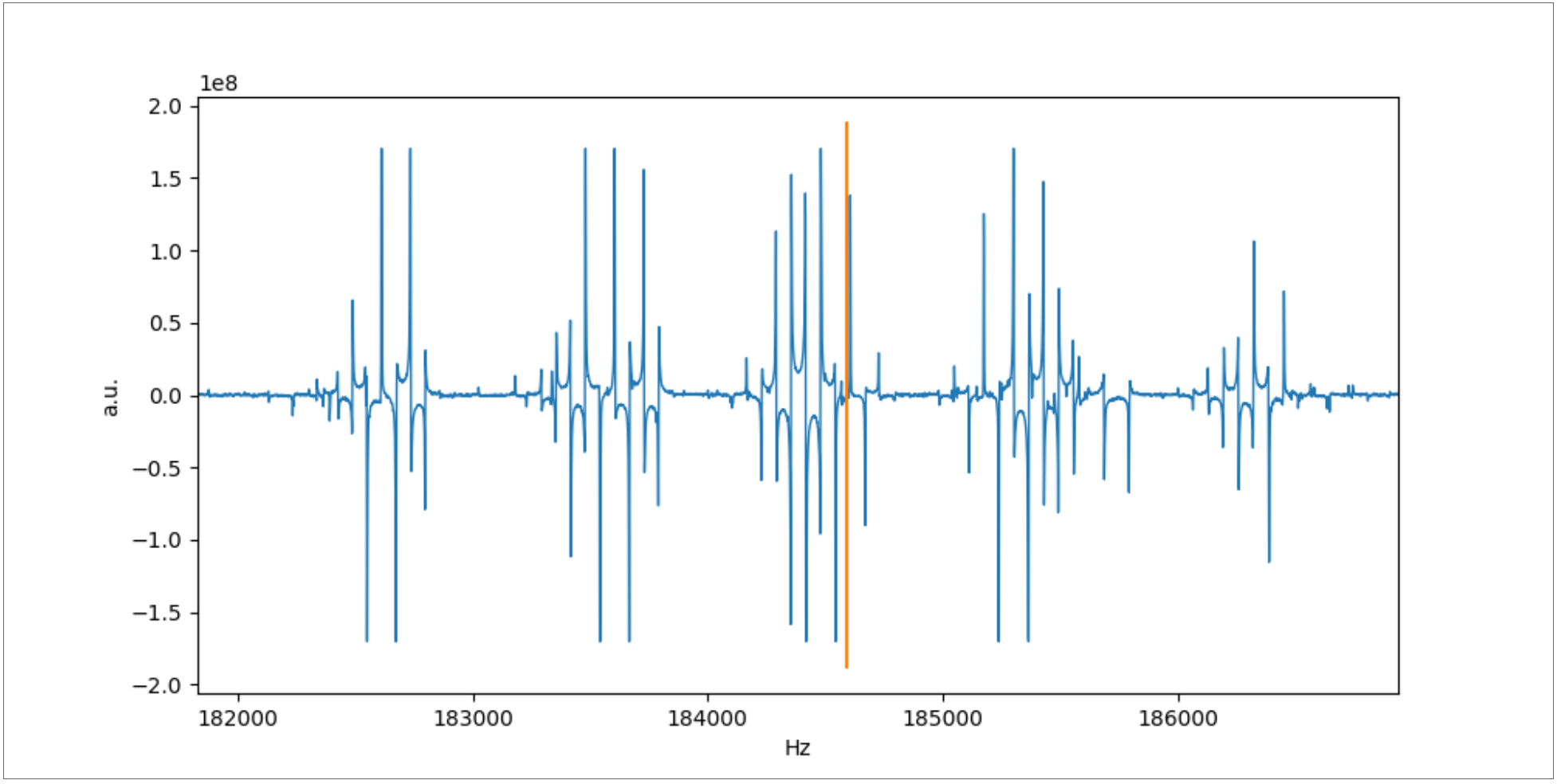
zoom on a small region

Note display in frequency

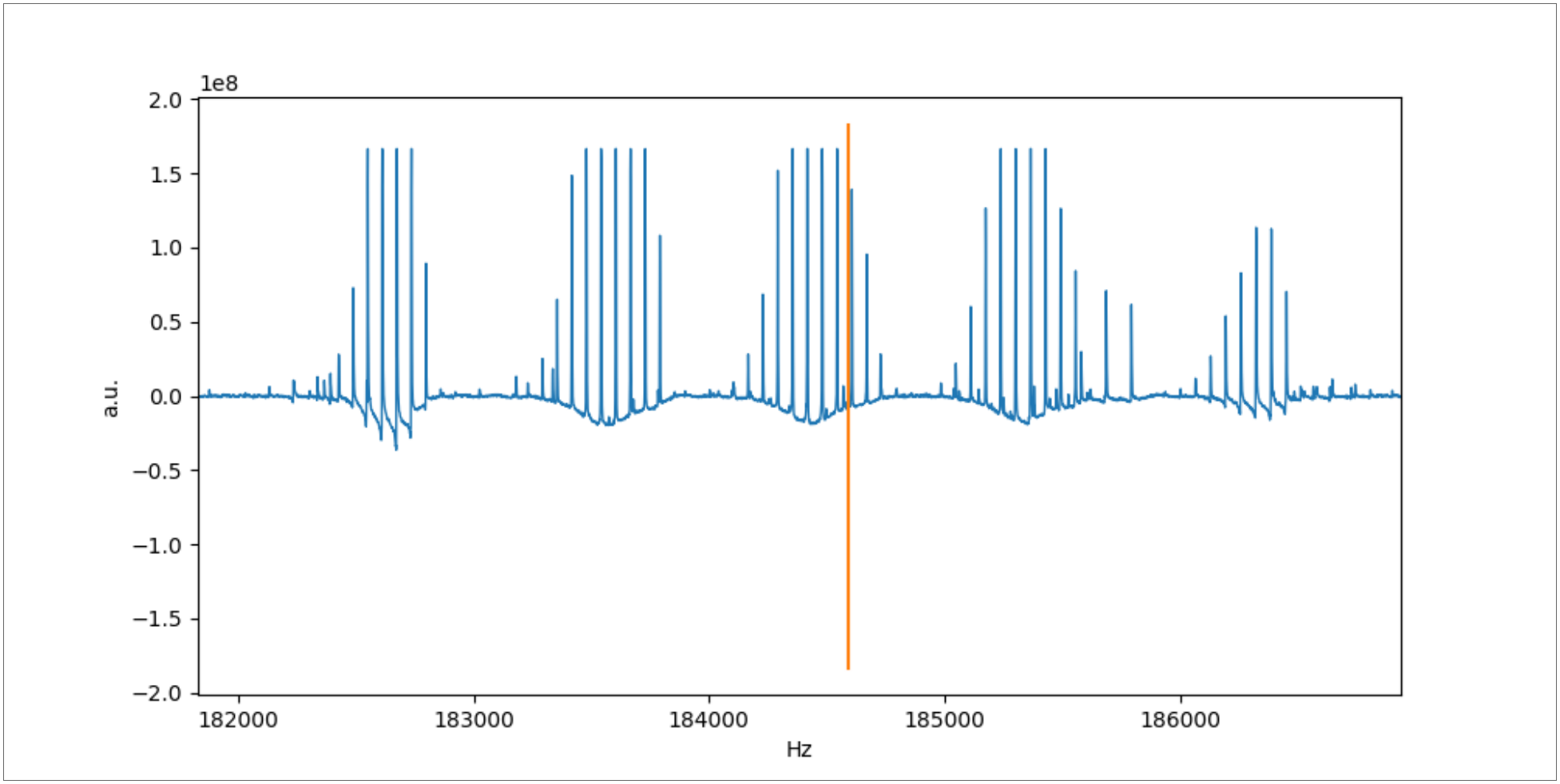


Schematic of the excitation pulse

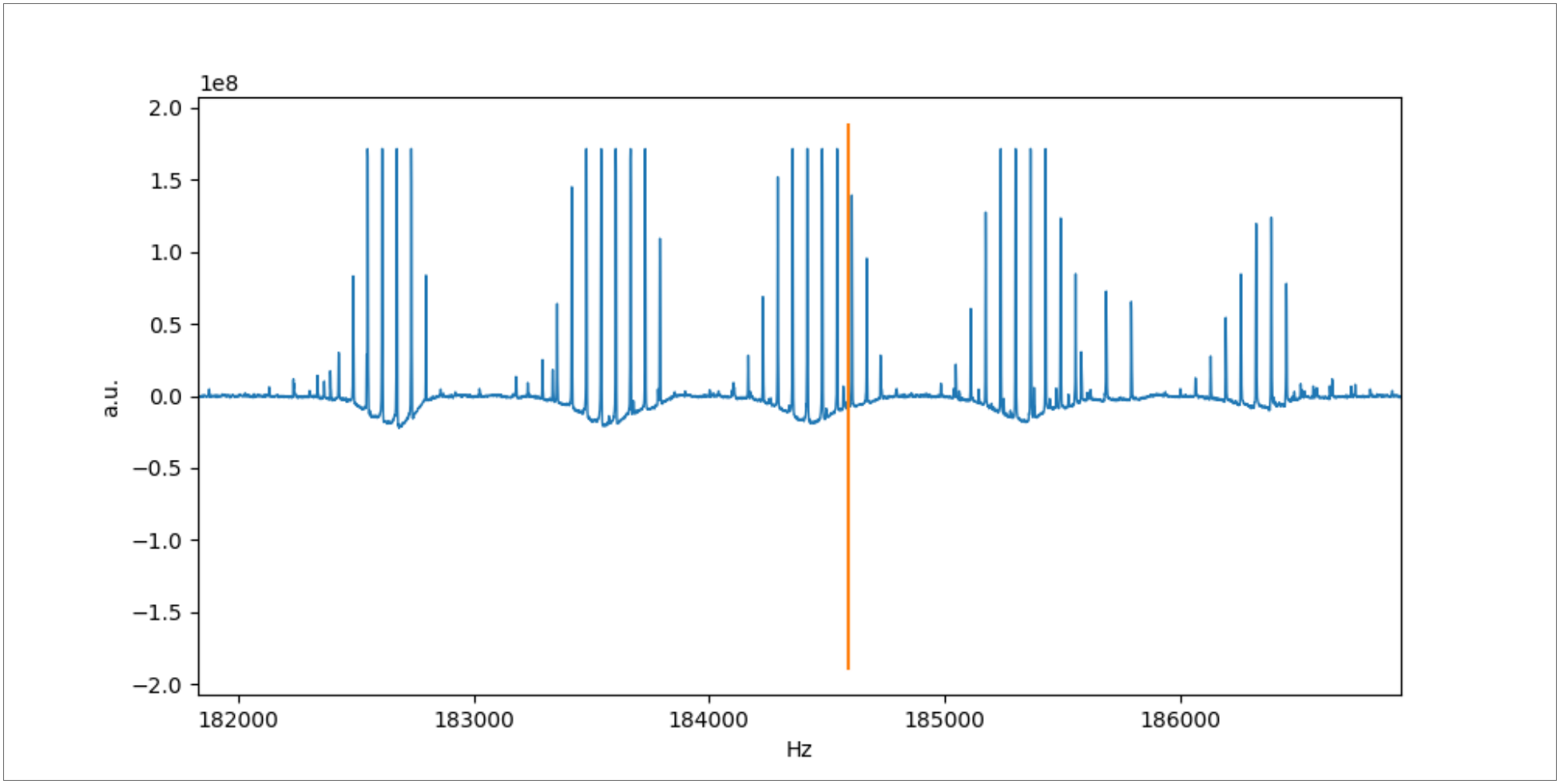
Phasing procedure



1/ first select a zone (here around 184kHz : m/z 588)

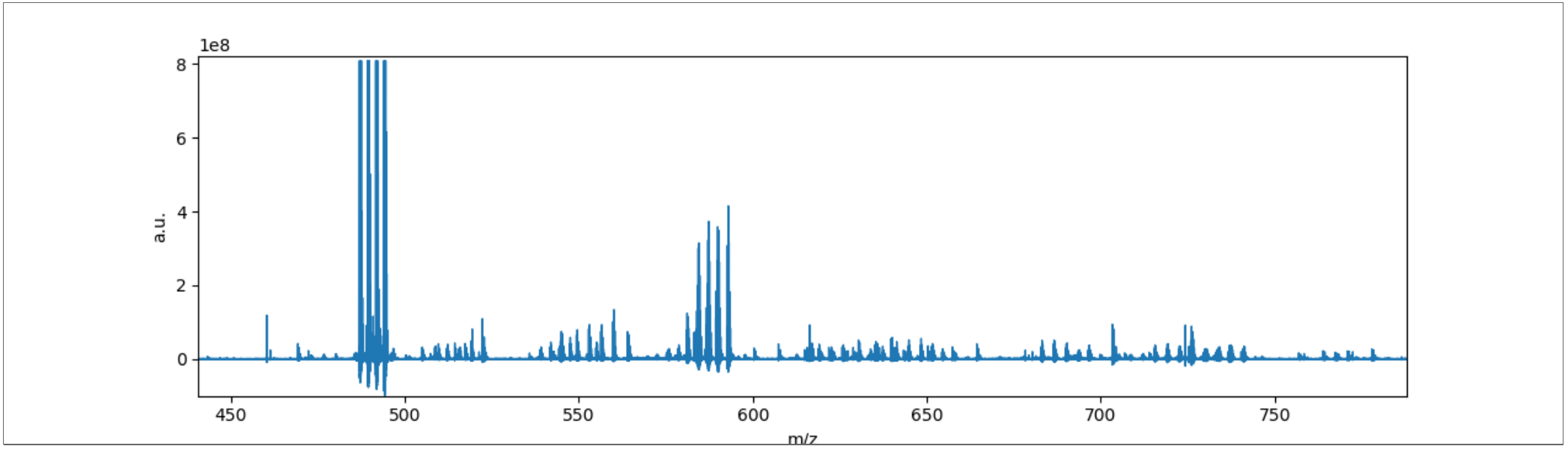


2/ unroll linear dependence



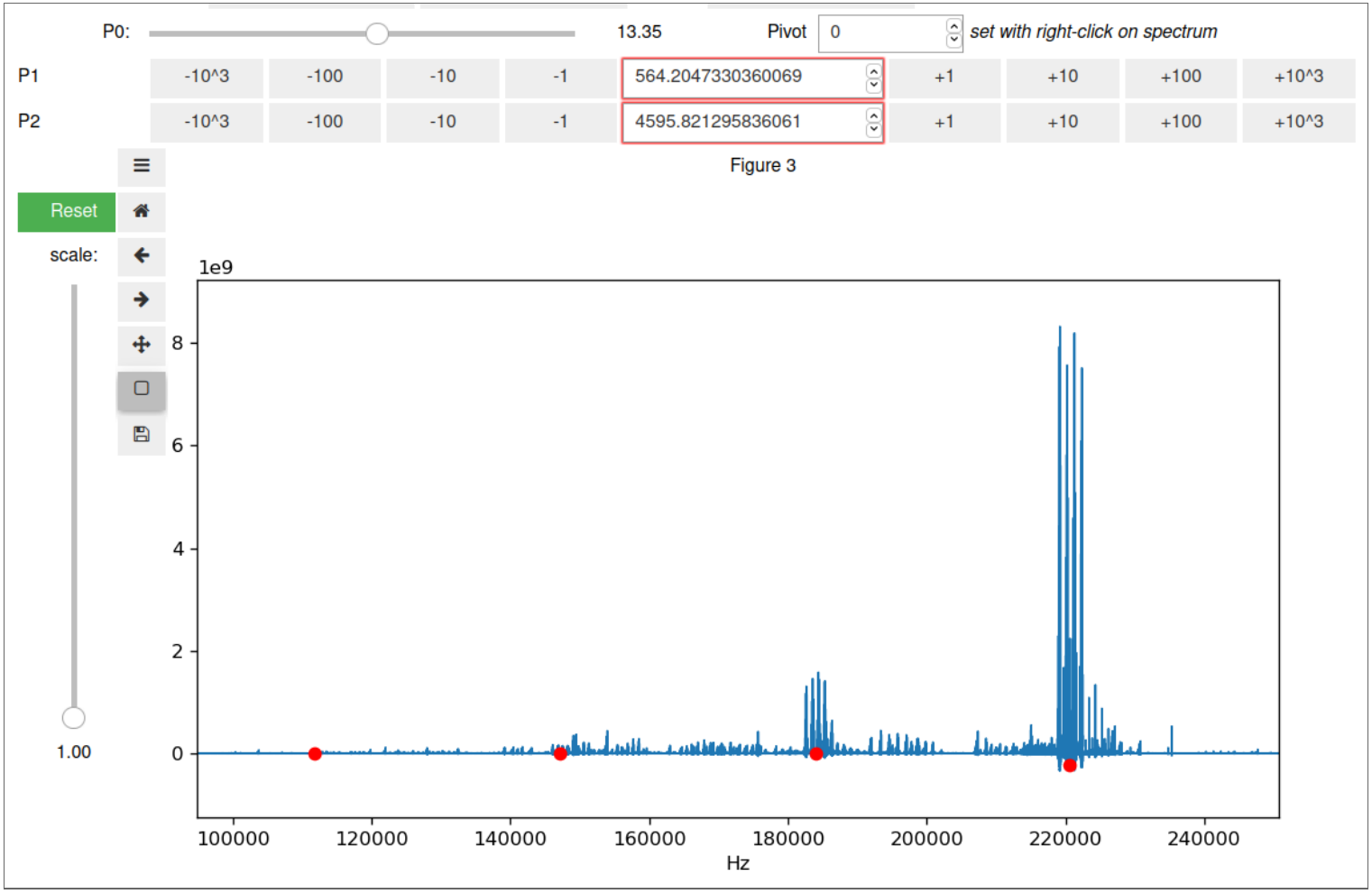
3/ unroll quadratic dependence

Phased Broadband spectrum

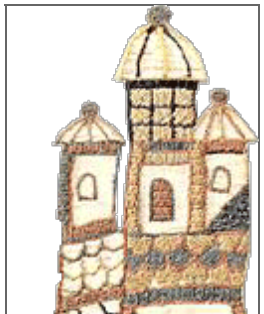


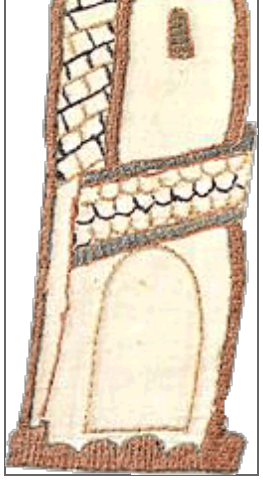
Phased Broadband spectrum

Phasing helper tool

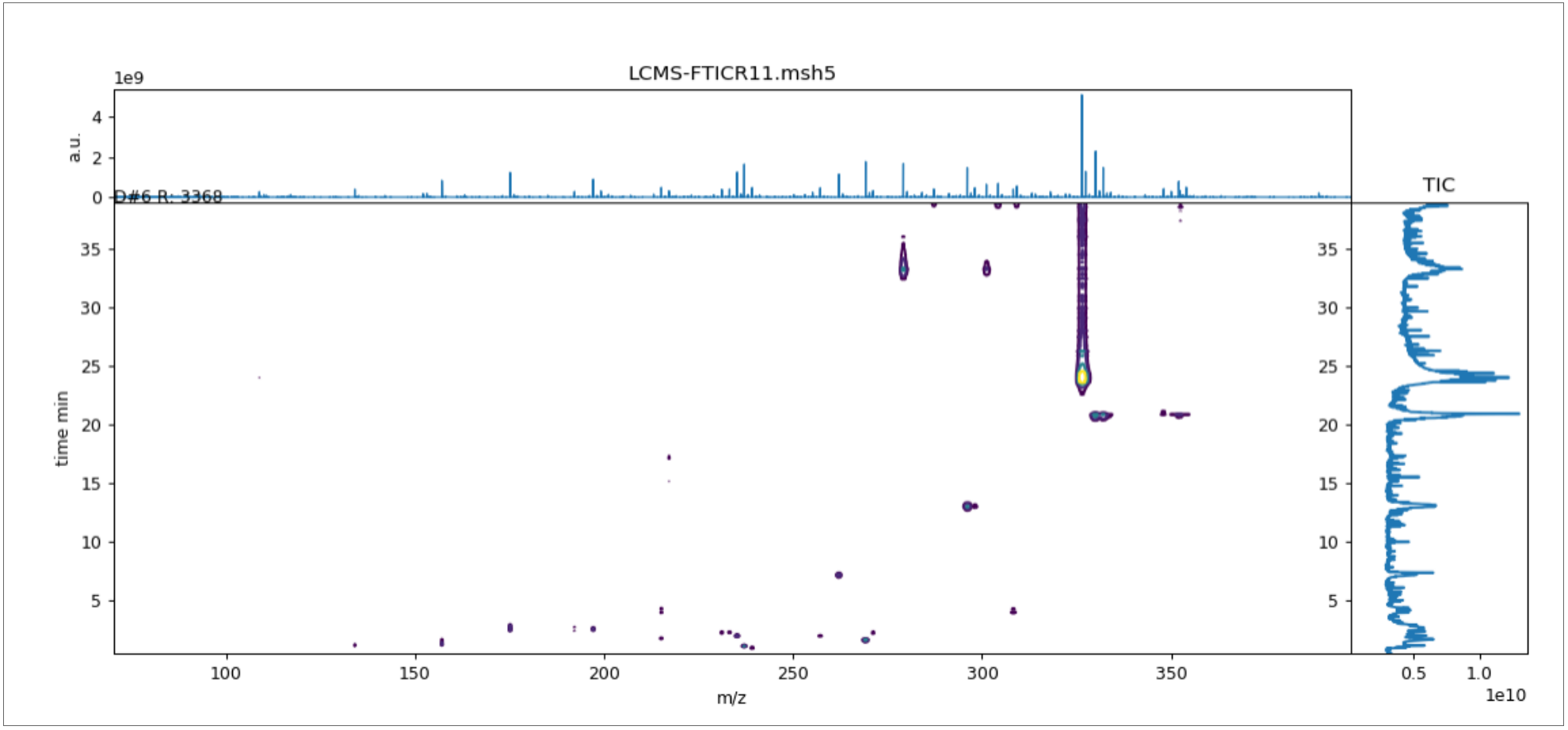


Larger datasets

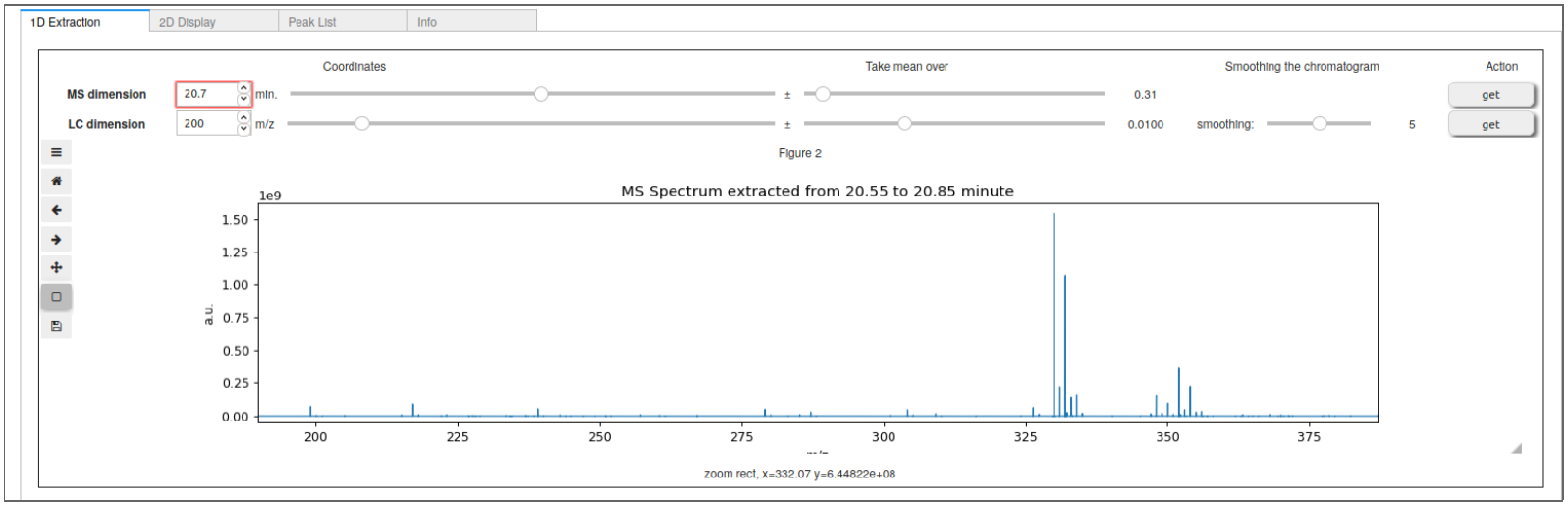




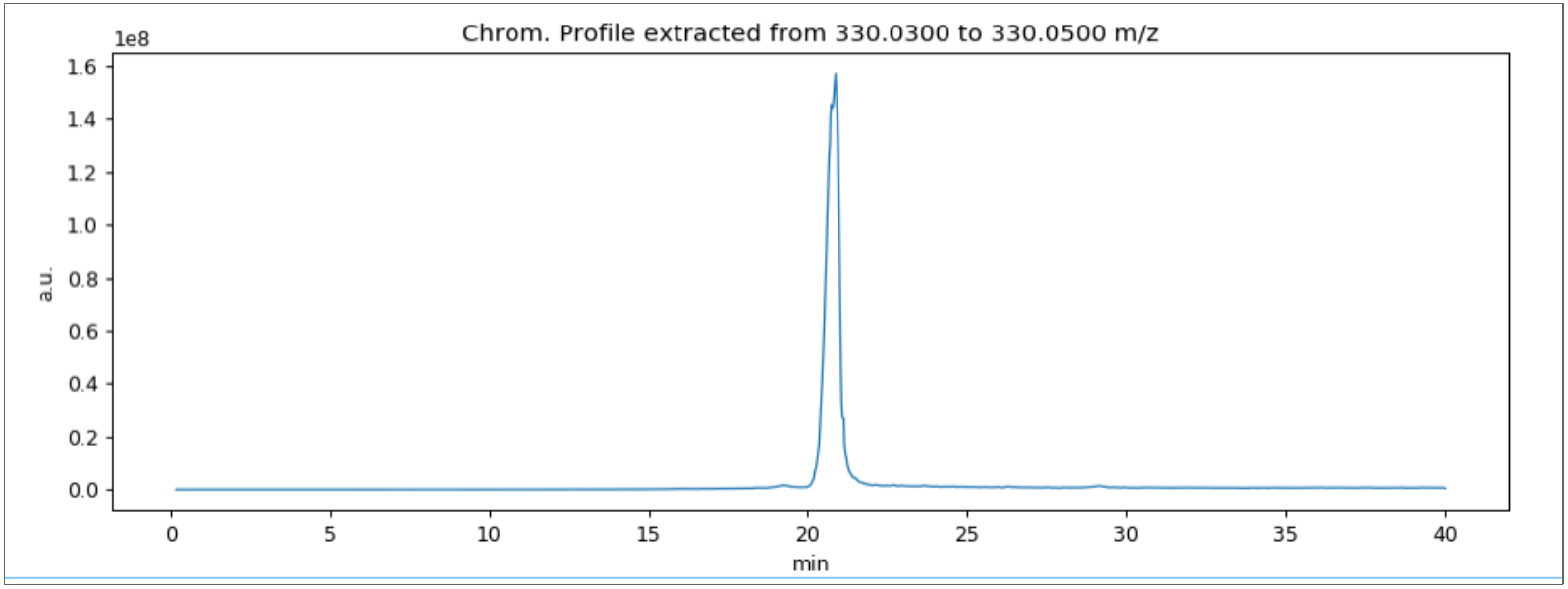
LC-MS



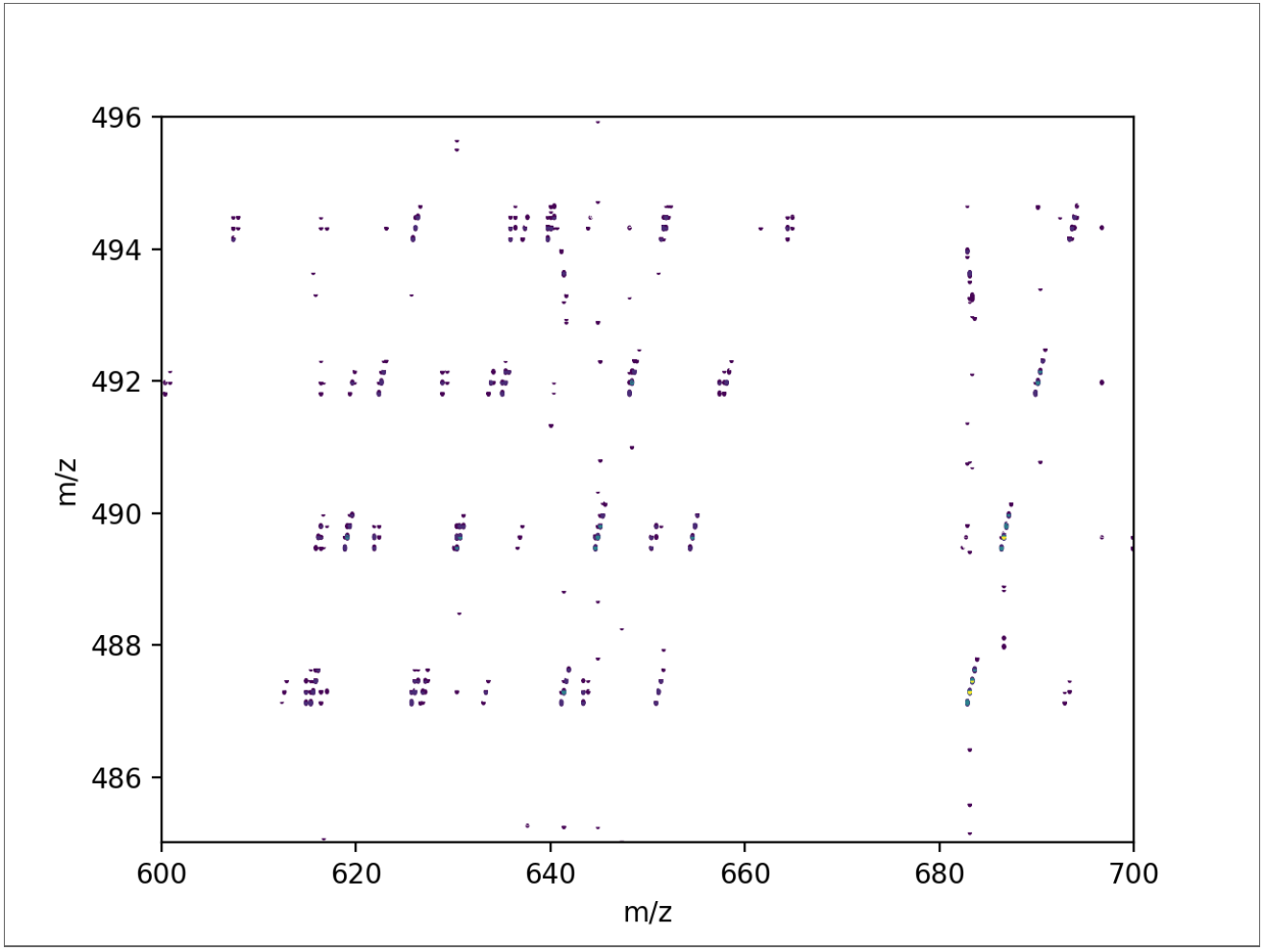
LC-FTICR-MS



extracting slices

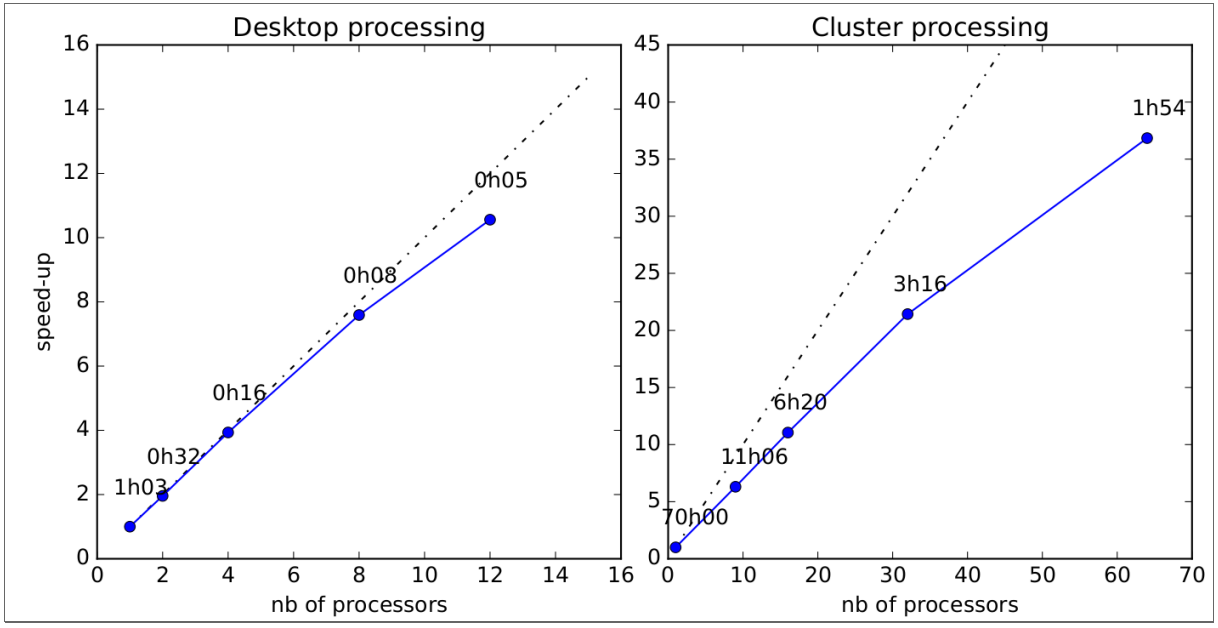


2D FTICR-MS



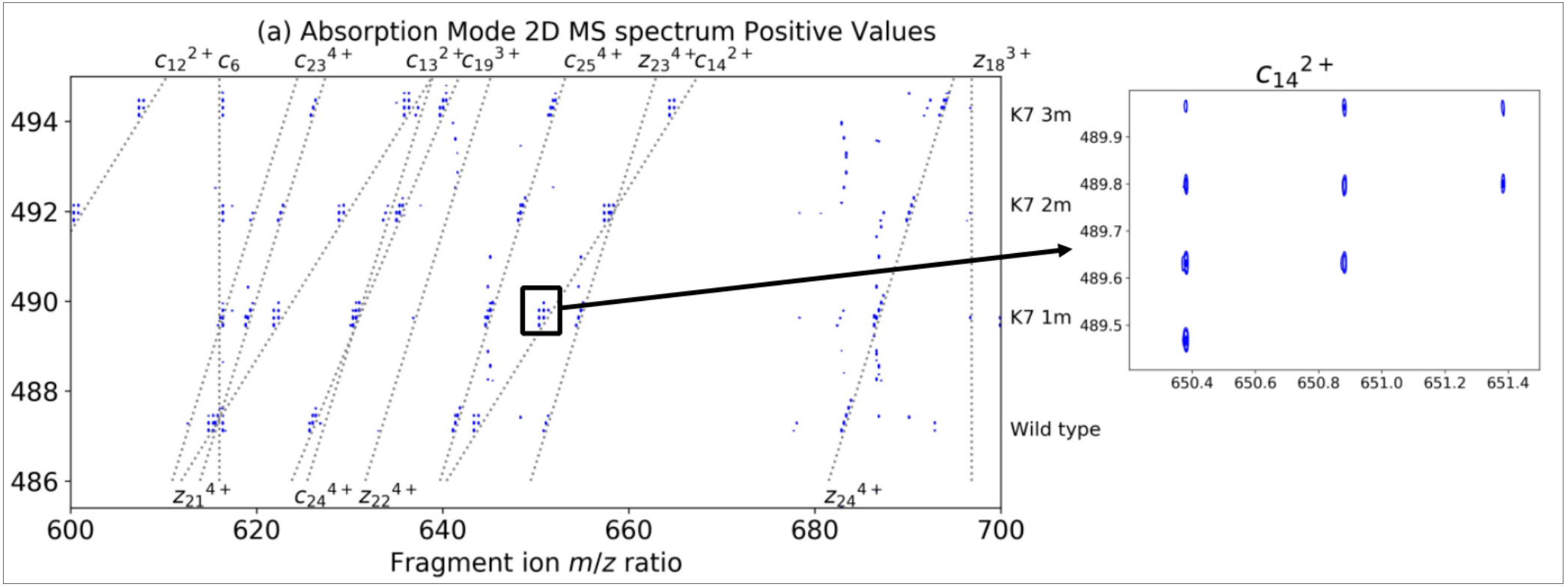
Several Challenges

- large File *(10 .. 100GB)*
- Large Processing



- DeNoising is usually required
  - urQRd : fast and efficient

Phasing 2D FTICR-MS

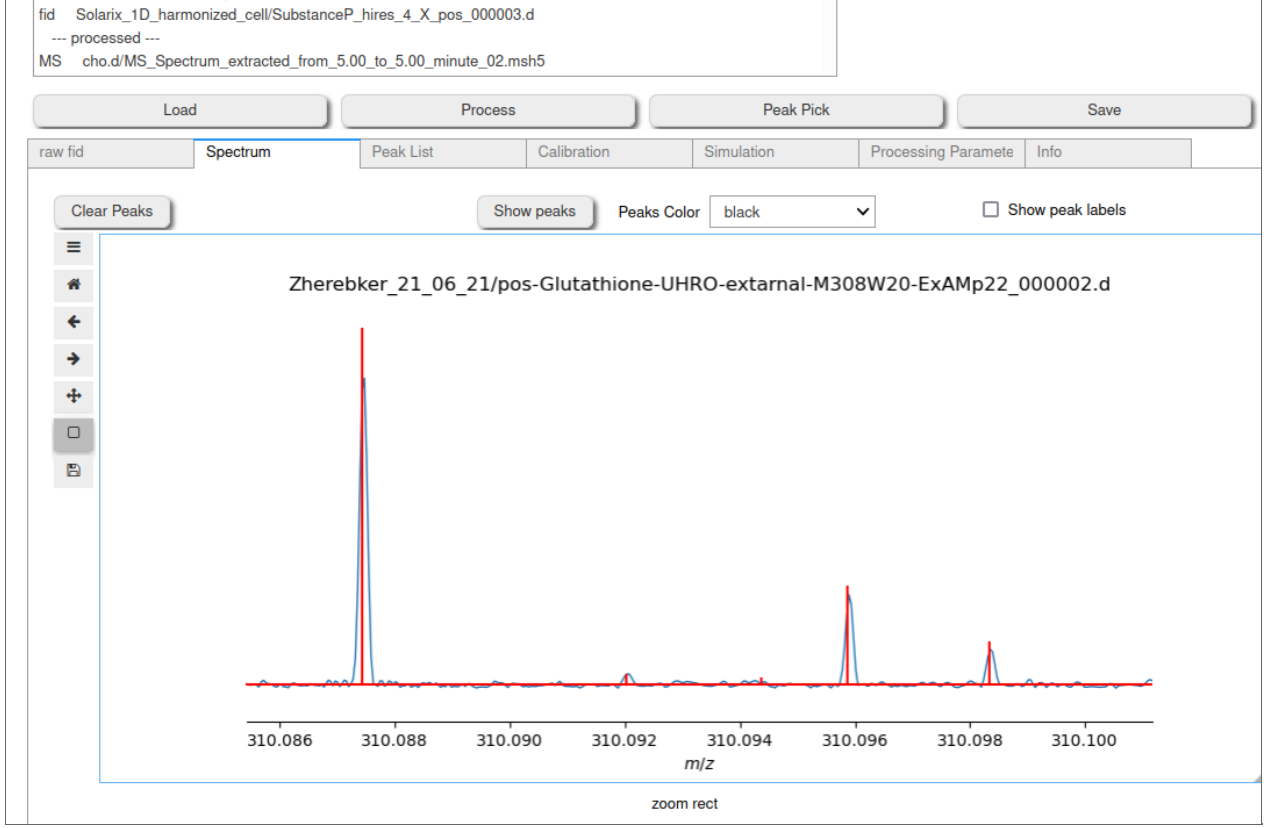


4MS

New Development

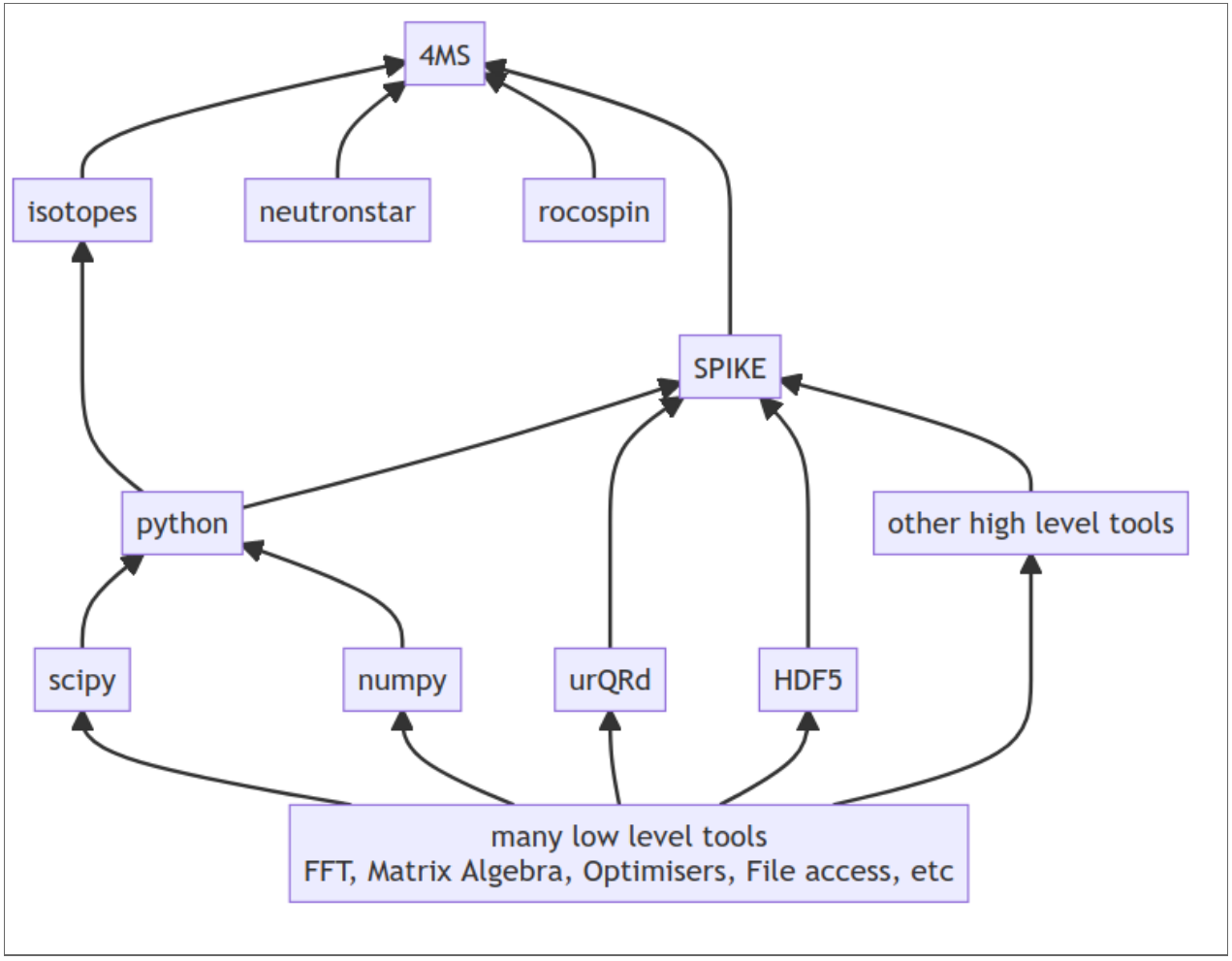
- by
- open-source scientific lib.
  - SPIKE
  - isotopes
  - neutronstar
  - rocospin
- technical lib.
  - tornado
  - ipywidget
  - voilà

⇒ Available 2023



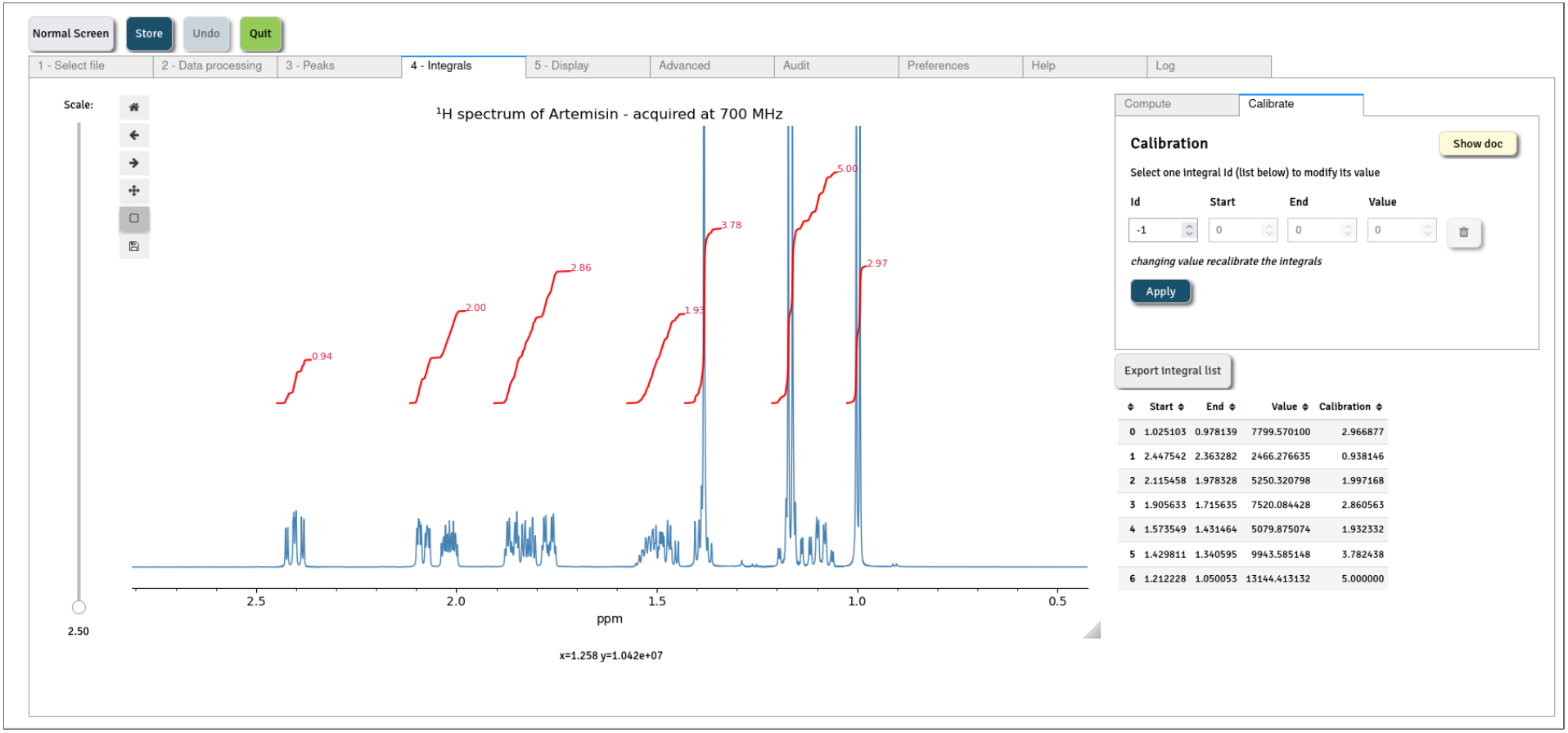
4MS with fine isotopic pattern analysis

stack of Open-Source software



4NMR

Same development, available soon !



Acknowledgments

work of many

Developments

- B.Kieffer
- *all open-source authors !*

2D FTICR-MS

- C.Rolando
- G.Bodenhausen
- P.O'Connor
- M.van Agthoven
- F.Bray
- …

CASC4DE

- Early
  - M-A.Coutouly
  - J-P Starck
  - L.Chiron
  - J.Asencio
  - C.Marin
- Data Mining / Software
  - L.Duciel
  - L.Baptista
  - A.Briot-Dietsch

source: ***"Historic Tale Construction Kit" github.com/htck***
presentation done with **Quarto / Revealjs**