Supporting Information

PyC2MC: an open-source software solution for visualization and treatment of high-resolution mass spectrometry data

Maxime SUEUR[1,3], Julien F. MAILLARD[1,3], Oscar LACROIX-ANDRIVET[1,3,4], Christopher P. RÜGER[3,5*], Pierre GIUSTI[1,2,3], Hélène LAVANANT[1], Carlos AFONSO[1,3*]

[1] Normandie Univ, UNIROUEN, INSA Rouen, CNRS, COBRA, 76000 Rouen, France.

[2] TotalEnergies OneTech R&D, TotalEnergies Research & Technology Gonfreville, BP 27, 76700 Harfleur, France

[3] International Joint Laboratory - iC2MC: Complex Matrices Molecular Characterization, TRTG, BP 27, 76700 Harfleur, France.

[4]TotalEnergies OneTech R&D, Centre de Recherche de Solaize (CRES), Chemin du canal, BP 22, 69360 Solaize, France

[5] Joint Mass Spectrometry Centre, Chair of Analytical Chemistry, University of Rostock, 18059 Rostock, Germany; Interdisciplinary Faculty, Department Life, Light & Matter (LL&M), University of Rostock, 18051 Rostock, Germany.

*** corresponding authors: christopher.rueger@uni-rostock.de, carlos.afonso@univ-rouen.fr**

Keywords: complex matrices, data visualization, statistical analysis, open access software, Python

Content

1. Example files

Even though the developed application is made to load result files produced by specific third-party software, it is also possible to load generic files. In this case, two options are presented to the users: create and load a file containing non-attributed data or create and load a file containing attributed data. For the first option, the user should create a .asc file with m/z ratio, intensity, and S/N ratio (optionally) as columns. This can be done by copy-pasting a table from spreadsheet software to a text editor and saving this file in the .asc format. No header has to be specified, only the column arrangement matters here. An example is displayed on the left side of Table S1 For the other option, a custom attributed data file should contain the m/z ratio, absolute intensity, attribution error (in ppm), and the molecular formula in a .csv file. A header must be specified for this type of file; however, it is not necessary to use the same names as only the columns arrangement matters. An example is displayed on the right side of table S1.

Table S1: Tables representing accepted architectures. On the left: The non-attributed data file contains m/z ratio, intensity, and S/N ratio in this order with only a blank space between each piece of information. On the right: The attributed data file is a proper table with a header for each piece of information: m/z ratio, intensity, error, and molecular formula.

Non-attributed data file	Attributed data file			
No headers	m/z ratio	Intensity	Error	Sum formula
98.84049 1287542 0.00008	335.182798	16687068	0.368	C23 H27 S
101.38376 834970 0.00008	335.211790	3733477	0.219	C22 H27 N2 O
102.51128 3366713 0.00009	335.224366	2366441	-0.131	C23 H29 N O
102.61279 1008359 0.00007	335.236942	7102445	0.082	C24 H31 O
102.95155 1034841 0.00008	335.240313	6684655	0.075	C21 H35 O S
103.40512 844062 0.00009	335.248175	5672945	0.163	C23 H31 N2
103.69160 1079009 0.00009	335.260751	55652340	0.424	C24 H33 N
103.70885 1023189 0.00007	335.273328	3583993	0.360	C25 H35
103.71896 1190111 0.00009	336.063709	2128115	-0.372	C20 H16 O S2
104.10701 5630360 0.00011	336.084147	3498233	0.267	C23 H14 N S
104.41281 869834 0.00011	336.087518	2377978	0.132	C20 H18 N S2
105.26436 837208 0.00007	336.096723	23154940	0.340	C24 H16 S
106.04990 2359553 0.00013	336.100094	14373118	0.408	C21 H20 S2
106.28434 1115756 0.00008	336.114481	1868035	0.268	C24 H16 O2
106.28457 1231404 0.00008	336.125715	1520390	-0.201	C23 H16 N2 O
106.40151 1157645 0.00011	336.138291	11501834	0.013	C24 H18 N O
106.95061 1028292 0.00009	336.141662	3336459	-0.011	C21 H22 N O S
107.01977 891559 0.00009	336.150867	12230926	0.062	C25 H20 O
107.07981 1231473 0.00009	336.154238	4158735	0.112	C22 H24 O S
109.04053 908808 0.00009	336.162100	20107538	0.220	C24 H20 N2

2. Datasets details

Hereafter is the tree structure of the data files used in the main body that will also be available on GitHub (https://github.com/iC2MC/PyC2MC_viewer) and Zenodo (https://doi.org/10.5281/zenodo.7602288) repositories of the software.

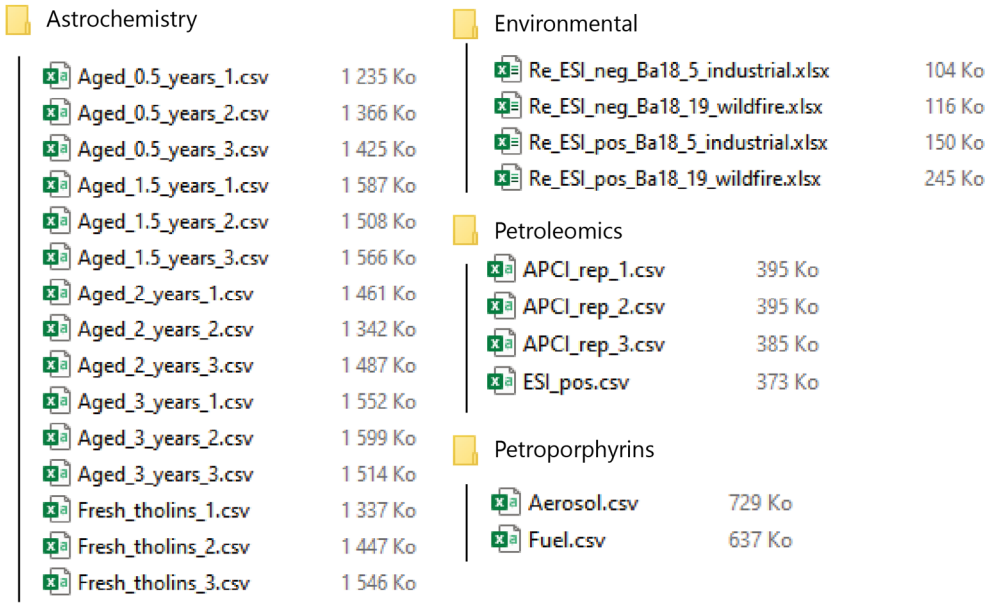

Figure S1: Tree view of the datasets used

Table S2: Details on the datasets used

Dataset	Petroleomics	Petroporphyrins	Environmental	Astrochemistry
Analysis type(s)	ESI+ / APCI +	ET-MALDI	ESI +/-	LDI +
Mean number of attributions	4942	7615	1988	9719
Molecular formula boundary	$C_xH_yN_2O_5S_2$	$C_xH_yN_4$ (-VO_3) or (-Ni)	$C_xH_yN_3O_{17}S_2$	$C_xH_yO_2N_{30}$

These datasets were respectively used by Mase et al.[1], Sueur et al. [2], Schneider et al[3] and Maillard et al [4] for their published studies.

3. Error plots

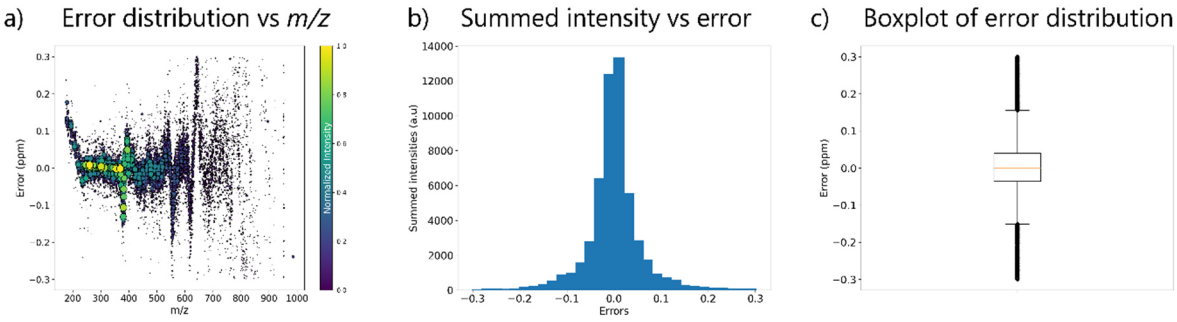

Figure S2: Three ways to represent the attribution error in PyC2MC: a) error distribution versus m/z ratio with intensity as color code. b) histogram representing the distribution of the intensity versus attribution error. c) Boxplot representation of the error distribution

This figure shows the different representations of the attribution error available in the software. The above diagrams were generated using the Astrochemistry dataset.

4. Aromaticity index and maximum carbonyl ratio as a color-coding variable for Van Krevelen plots

In addition to the average carbon oxidation state displayed in a Kroll plot, two other variables are commonly used in environmental science to evidence chemical properties of detected species: the maximum carbonyl ratio (MCR) and the aromaticity index (AI). MCR is calculated as follows [5]:

The resulting value gives information on the oxidation of a molecule : [0;0.2] : Very highly oxidized; [0.2;0.5] : highly oxidized; [0.5;0.9] : Intermediately oxidized and [0.9;1] : highly unsaturated. An example of a Van Krevelen diagram of an industrial wildfire sample using MCR as a color-coding variable is displayed in Figure S2.a.

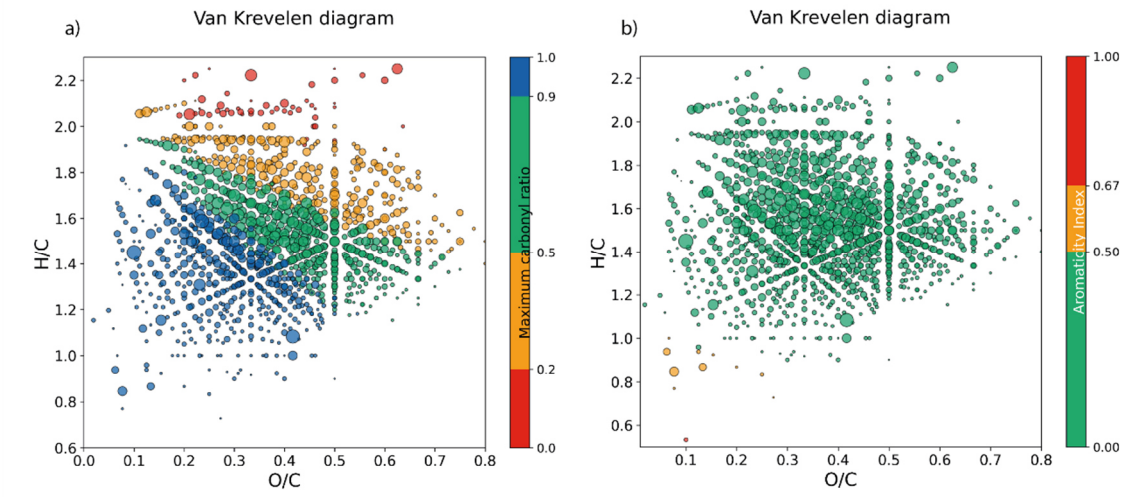

Figure S3: Van Krevelen diagrams using a) MCR and b) AI as a color-coding variable.

Aromaticity index is a measure of the carbon-carbon double bond density[6]. It is calculated as follows:

$$AI = \frac{DBE_{AI}}{C_{AI}} = \frac{1 + C - O - S - 0.5(N + P + H)}{C - O - S - N - P}$$

$$If\ DBE_{AI} \leq 0\ or\ C_{AI} \leq 0, then\ AI = 0$$

AI thus gives information on the saturation and aromaticity of detected species: AI > 0.5 : aromatic species; AI ≥ 0.67 : condensed aromatics. An example of a Van Krevelen diagram using AI as a color-coding variable is displayed in Figure S4.b2. A modified version of the AI is used to characterize dissolved organic matter (DOM) as in this kind of sample, approximately half of the oxygen is bound using σ-bounds rather than π-bounds:

$$AI_{mod} = \frac{1 + C - 0.5\,O - S - 0.5\,(N + P + H)}{C - 0.5\,O - S - N - P}$$

5. Inter sample comparison

The following figure exemplifies the primary comparative figures of the software. Figure S2.a shows the abundance of the two main compound families in Tholins samples at different ageing steps, exhibiting the oxidation process. However, Figure S2.b shows no variation of the DBE pattern correlated to the ageing process.

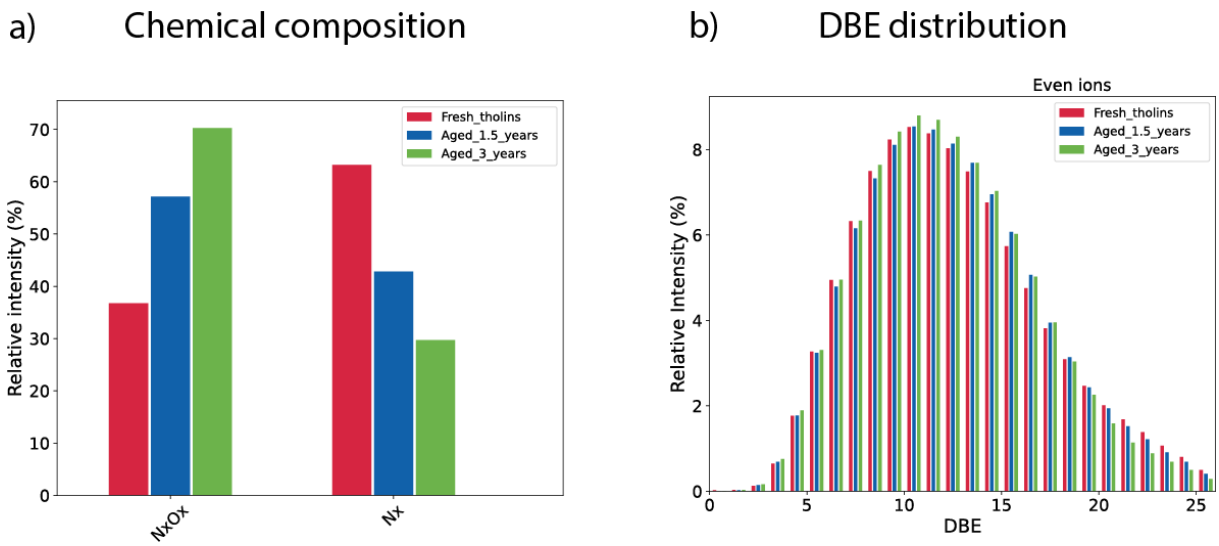

Figure S4: Demonstration of the inter sample comparison functionalities.

6. Statistical analysis features

In addition to the volcano plot featured in the main body of this article, PyC2MC encompasses other statistical analysis tools such as PCA and HCA (See figure S3). The following diagrams were generated using the astrochemistry dataset and illustrate the separation of the samples according to their age.

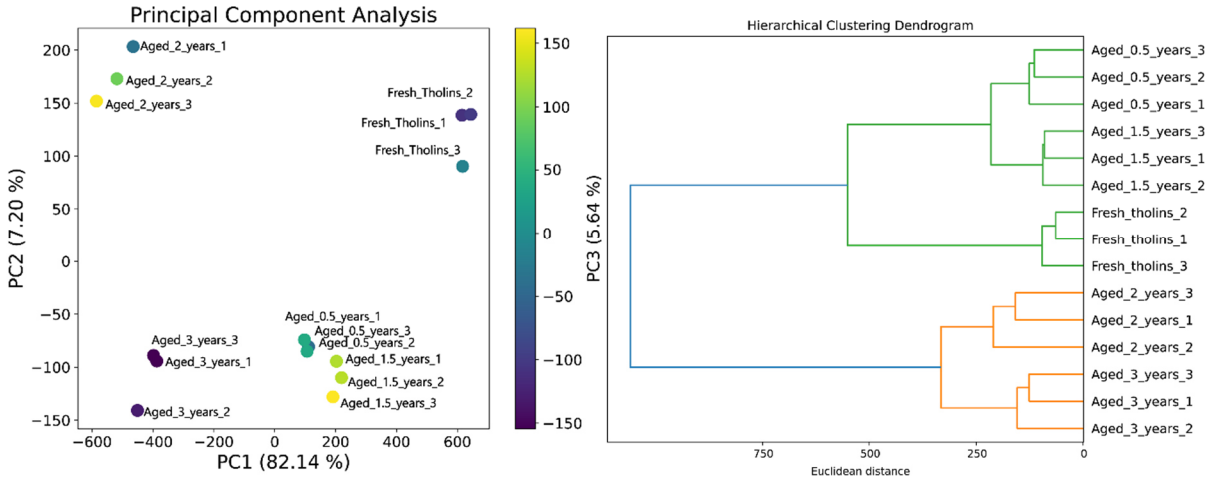

Figure S 5: Statistical analysis features. On the left: Principal Component Analysis (PCA). On the right: Hierarchical Clustering Analysis (HCA).

7. Split finder

The additional feature named “Split Finder” is provided with its own GUI which is shown in Figure S 6 below. On this interface, the user can input the value of the observed split and the tolerance, both in Daltons. Then when the “Find my split” is clicked, a list of the plausible isotope couples and their corresponding error is shown. The solution presenting the smallest error (in absolute value) is highlighted; however, the users should acknowledge that this solution might not be the most plausible one.

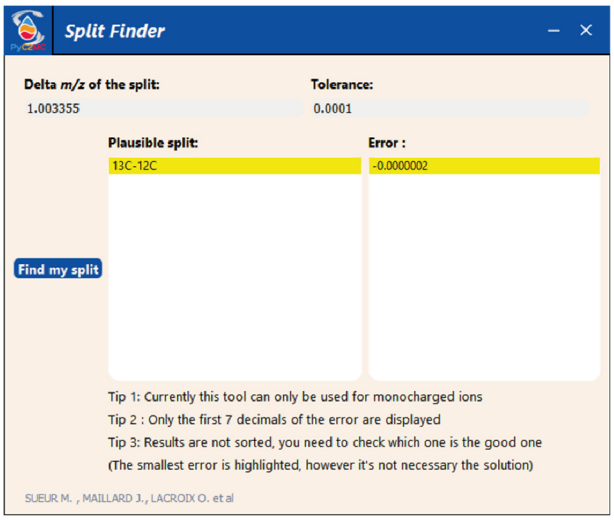

Figure S6: Split Finder GUI developed under PyQT5

References

(1) Mase, C.; Maillard, J. F.; Paupy, B.; Farenc, M.; Adam, C.; Hubert-Roux, M.; Afonso, C.; Giusti, P. Molecular Characterization of a Mixed Plastic Pyrolysis Oil from Municipal Wastes by Direct Infusion Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Energy Fuels* **2021**, *35* (18), 14828-14837. DOI: 10.1021/acs.energyfuels.1c01678.
(2) Sueur, M.; Rüger, C. P.; Maillard, J. F.; Lavanant, H.; Zimmermann, R.; Afonso, C. Selective characterization of petroporphyrins in shipping fuels and their corresponding emissions using electron-transfer matrix-assisted laser desorption/ionization Fourier transform ion cyclotron resonance mass spectrometry. *Fuel* **2023**, *332*. DOI: 10.1016/j.fuel.2022.126283.
(3) Schneider, E.; Czech, H.; Popovicheva, O.; Lütdke, H.; Schnelle-Kreis, J.; Khodzher, T.; Rüger, C. P.; Zimmermann, R. Molecular Characterization of Water-Soluble Aerosol Particle Extracts by Ultrahigh-Resolution Mass Spectrometry: Observation of Industrial Emissions and an Atmospherically Aged Wildfire Plume at Lake Baikal. *ACS Earth and Space Chemistry* **2022**, *6* (4), 1095-1107. DOI: 10.1021/acsearthspacechem.2c00017.
(4) Maillard, J.; Carrasco, N.; Schmitz-Afonso, I.; Gautier, T.; Afonso, C. Comparison of soluble and insoluble organic matter in analogues of Titan's aerosols. *Earth. Planet. Sci. Lett.* **2018**, *495*, 185-191. DOI: 10.1016/j.epsl.2018.05.014.
(5) Zhang, Y.; Wang, K.; Tong, H.; Huang, R. J.; Hoffmann, T. The maximum carbonyl ratio (MCR) as a new index for the structural classification of secondary organic aerosol components. *Rapid Commun. Mass Spectrom.* **2021**, *35* (14). DOI: 10.1002/rcm.9113.
(6) Koch, B. P.; Dittmar, T. From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. *Rapid Commun. Mass Spectrom.* **2006**, *20* (5), 926-932. DOI: 10.1002/rcm.2386.