

PyC2MC: An Open-Source Software Solution for Visualization and Treatment of High-Resolution Mass Spectrometry Data

Maxime Sueur, Julien Maillard, Oscar Lacroix-Andrivet, Christopher Rüger, Pierre Giusti, Hélène Lavanant, Carlos Afonso

▶ **To cite this version:**

Maxime Sueur, Julien Maillard, Oscar Lacroix-Andrivet, Christopher Rüger, Pierre Giusti, et al.. PyC2MC: An Open-Source Software Solution for Visualization and Treatment of High-Resolution Mass Spectrometry Data. Journal of The American Society for Mass Spectrometry, 2023, 10.1021/jasms.2c00323 . hal-04018590

HAL Id: hal-04018590

https://normandie-univ.hal.science/hal-04018590

Submitted on 8 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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[1,2,3], Christopher P. RÜGER[3,4*], 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] 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 author: christopher.rueger@uni-rostock.de**

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

ABSTRACT

Complex molecular mixtures are encountered in almost all research disciplines, such as biomedical ‘omics, petroleomics, and environmental sciences. State-of-the-art characterization of sample materials related to these fields, deploying high-end instrumentation, allow for gathering humongous quantity of molecular composition data. One established technological platform is ultrahigh-resolution mass spectrometry, *e.g.*, Fourier-transform mass spectrometry (FT-MS). However, the huge amounts of data acquired in FT-MS often result in tedious data treatment and visualization. FT-MS analysis of complex matrices can easily lead to single mass spectra with more than 10,000 attributed unique molecular formulae. Sophisticated software solutions to conduct these treatment and visualization attempts from commercial and non-commercial origins exist. However, existing applications have distinct drawbacks, such as focusing on only one type of graphic representation, being unable to handle large datasets, or not being publicly available. In this respect, we developed a software, within the international complex matrices molecular characterization joint lab (IC2MC), named “python tools for complex matrices molecular characterization” (PyC2MC). This piece of software will be open-source and free to use. PyC2MC is written under python 3.9.7 and relies on well-known libraries such as pandas, NumPy, or SciPy. It is provided with a graphical user interface developed under PyQt5. The two options for execution, 1) user-friendly route with pre-packed executable file or 2) running the main python script through a Python interpreter, ensure a high applicability but also an open characteristic for further development by the community. Both are available on the GitHub platform (https://github.com/iC2MC/PyC2MC_viewer).

INTRODUCTION

Complex molecular mixtures are encountered in almost all research domains. Their comprehensive chemical description is referred to as 'omics sciences, such as in proteomics, lipidomics, and metabolomics [1], petroleomics [2, 3, 4] or in environmental sciences [2, 5], to name a few. Omics approaches allow for a massive richness of in-depth information on these complex samples. Consequently, transformations and processes can be studied at the molecular level. Naturally, omics experiments are associated with large numbers of molecular data. In practice, the chemical characterization of complex mixtures requires state-of-the-art analytical instrumentation. One commonly applied technique is high-resolution mass spectrometry. In this category, Fourier transform ion cyclotron resonance mass spectrometry (FTICR MS) and Orbitrap MS are the two primary platforms with adequate performance, *i.e.*, resolving power above 10^5 and mass accuracy below 1 ppm [6]. Their superior resolving power allows the separation of isobaric constituents, whereas the high mass accuracy enables unique molecular formula attribution to each ion. The attribution of molecular formulas can be performed either using vendor proprietary software, such as DataAnalysis (Bruker Daltonics, US), Xcalibur (Thermo Fisher Scientific, US), and Composer (Sierra Analytics, US) or using specific workflows, such as PetroOrg [7], Peak-by-Peak (Spectroswiss, Switzerland), Attributor [8], OpenMS [9] or mMass [10]. Moreover, numerous specific workflows for given research domains or single laboratory solutions exist. An example is ICBM-OCEAN [11], developed at the University of Oldenburg, which is specialized in dissolved organic matter (DOM) analysis. Results can then usually be recovered in large spreadsheets or software-specific formats. Even though these software workflows allow for robust molecular attribution, most do not permit advanced visualization or comparison capabilities. Hence, in daily routine as a rapid and easily accessible solution, data treatment and visualization are performed based on simple scripts (*e.g.*, R, Python, MATLAB) or even utilizing spreadsheet software like Microsoft Excel or OriginLab Origin. However, when the number of analyses to be processed becomes high, this becomes tedious or even impossible. Complex matrices analyses require the use of fingerprint visualizations that are essential to understand the underlying chemistry. Diagrams such as double bond equivalent [12] (DBE) maps, van Krevelen [13], or Kendrick [14] plots are of common use in petroleomics due to the clear representation of a complex organic sample. In the same way, average carbon oxidation state [15], modified aromatic index [16], and maximum carbonyl ratio [17] representations are often used in environmental sciences. Manually generating these graphs is extremely time-consuming and requires the use of individual coding solutions.

Several open-source software packages are publicly available for the visualization of complex molecular data. Exemplary, DEIMoS (Data Extraction for Integrated Multidimensional Spectrometry) [18], which is a Python package for treating data from hyphenated analytical techniques, such as liquid or gas chromatography coupled to mass spectrometry, can be named. In brief, DEIMoS allows an alignment of the *m/z* information and to visualize the extracted compounds and molecular features. Although it provides two-dimensional visualization options, often missed in vendor software, it does not give any other type of data representation like the aforementioned DBE maps and other utilization of the attributed molecular information. Furthermore, no graphical user interface (GUI) is available, and command-line programming is required. Another recently published software package is Constellation [19]. This program focuses explicitly on systematic trend detection using Kendrick mass defect (KMD) plots. It aims at finding repetitive patterns in a KMD plot obtained from non-attributed data to exhibit homologue rows with the goal of supporting and/or validate molecular formula attribution [20]. Unlike other software solutions, Constellation has the advantage of being a web application where computing takes place on a remote server. Thus, beneficially, it does not require any installation nor extended local computational resources. However, a limit is imposed on the uploaded data size, with a maximum of 5,000 peaks that can be treated at once, and permanent network connection is needed. Even though a number of 5,000 peaks seems high at first, in most FTMS data on complex mixtures, it is not sufficient as it is widespread to recover more than 10,000 attributed species per mass spectra. A recent application is PyKrev [21], which provides several useful visualization options, such as van Krevelen plots (VK) and violin representations, of intersections between samples, *i.e.*, the molecular formulas unique to a sample or common between the selected samples. In addition to these very recent software solutions, established workflows might often use older approaches, such as it is the case for OpenMS [9], published in 2016, which proposes a flexible workflow designer for data treatment and basic visualization. It has been optimized for proteomics and metabolomics and thus is not necessarily adapted to other fields like environmental sciences. Last, the work of the Barrow group needs to be mentioned here, *e.g.*, KairosMS, published in 2020 [22]. KairosMS is specialized in handling hyphenated mass spectrometric data, including scan-by-scan recalibration, a suite of visualization tools, including DBE, VK, evolution of class intensity over time, and principal component analysis. Even though KairosMS efficiently addresses most of the desired features, it is neither open-source nor available in a public repository.

In this context, we have developed a user-friendly and open-source solution for attributed high-resolution mass spectrometric data visualization. We aimed to process and handle complex organic mixture data with ten-thousands

of molecular features. This package, called PyC2MC, is based on Python as a high-level, interpreted language with an intrinsic comprehensive library and compiled with a graphical user interface. The software should be easily utilizable and improvable even by people outside of the initial project. Thus, the primary goal of this work is to provide a robust data visualization tool producing numerous plots as well as statistical analysis. Indeed, inspired by tools such as InteractiVenn [23], we implemented features allowing to print Venn diagrams or to perform clustering or principal component analysis (PCA). Intending to deliver an easy-to-use and extendable application, we use a PyQT-based GUI designed under QtDesigner for its cross-platform functionality and broad usage in the scientific community. For broad applicability to various research areas and types of complex mixtures, numerous parameters derived from the molecular formula have been considered, such as DBE, heteroatoms ratio (O/C, N/C, and others), average carbon oxidation state, aromaticity index or maximum carbonyl ratio.

Here we present our work on PyC2MC by highlighting its main features and capabilities. We will first describe the choices of programming language and libraries in the software construction, as well as the file architecture that input files should follow, exemplified in the used datasets. Then, we will present the workflow and data processing, that allow rapid plotting of basic functions, how specific KMD plots may be built and used or how environmental science variables may be represented. Finally, we will present the basic comparative features and the statistical tools available within PyC2MC.

METHODS

Environment. This code has been written in Python release 3.9.7 (August 2021). It relies on robust and broadly deployed Python libraries, the most important ones being: PyQT5, pandas [24], NumPy [25], SciPy [26], sklearn [26], matplotlib [27] and chemparse [28]. It was developed and primarily tested on a computer embedding Intel Core i5-9500 CPU at 3 GHz, 16 GB RAM, and Intel UHD Graphics 630 GPU, running under Microsoft Windows 10 as the operating system. The software was also beta-tested on various machines ranging from desktop working stations to laptops. The application can be launched directly from the main script in a Python interpreter, like Spyder 5.1.5 [29], or using the command line option. Consequently, individual changes from the programming user base can easily be made and directly tested. For classical usage cases, an executable file compiled using pyinstaller 4.8 [30] enables easy, straightforward exploitation without any programming knowledge required.

Input file architecture. Currently, four file extensions are supported: American standard code (*.asc*), comma-separated values (*.csv*), and in specific cases binary interchange file formats, such as Microsoft Office Excel sheets (*.xls* and *.xlsx)*. Examples of files are given in the supplementary information (SI 1). New formats can also be implemented easily by modifying the "loading_function.py" file.

PyC2MC does not provide a built-in molecular attribution feature, so molecular attribution should be performed using another software, exemplary workflows described below. Nevertheless, the input can be either attributed or non-attributed mass lists, as both data types may be processed to build relevant plots as described afterward. The *.asc* files are used to load non-attributed mass lists and should follow a simple structure with *m/z* values and intensities. The *.csv* files, if not exported from the already compatible vendor software, such as Bruker DataAnalysis or Thermo Fisher Scientific Xcalibur, should only be used for importing attributed data with the hereafter structure: *m/z* ratio, absolute intensity, attribution error (in ppm), molecular formula. Concerning Excel sheets, they are respectively the output format containing attributed peak lists of PetroOrg [14] and a MATLAB user-interface suite called CERES [42] (MATLAB R2022a) used in previous research [4, 31, 32]. All other variables and parameters are calculated further on. Attention should be paid to the order of columns, and a header should be included; however, not necessarily using the names mentioned above.

Datasets. In this study, several datasets are used to demonstrate the features of the PyC2MC application. Classical fingerprint visualization functions will be illustrated with the help of recently published data obtained from analyzing plastic pyrolysis oil by direct infusion ESI(+) and APCI(+) FTICR MS [33]. The representation of

parameters more common to environmental sciences will be performed using a dataset obtained with water-soluble ambient aerosol particles from emissions affected by anthropogenic industrial and wildfire sources analyzed by ESI(+/−) FTICR MS [31]. Specific use of the Kendrick mass defect plot will be shown using data gathered by the selective characterization of petroporphyrins in shipping fuels and their corresponding emission by electron-transfer MALDI FTICR MS [20]. This example highlights the usage of data on complex mixtures containing organic (C, H, N, O, S) and metal-organic (V, Ni) compounds. Finally, aerosol samples obtained under the mimicked atmospheric conditions of an exoplanet will constitute the last dataset to demonstrate the statistical analysis and inter-sample comparison. [34] With compounds having a very high N/C ratio – commonly not detected for natural earth mixtures – this dataset exemplarily represents the utilization of the PyC2MC software for less common complex matrices. Additional information on the selected datasets is found in the supplementary information (SI 2).

RESULTS AND DISCUSSION

Software workflow and data processing functionalities. Figure 1 gives a simplified graphical representation of the PyC2MC workflow. A view of the graphical user interface (GUI) is also presented in Figure 2. A primary input file can on the one hand contain a peak list (*m/z* and intensities without molecular formulae), it will then be identified by the import function as a *mass list*. On the other hand, the input file may be an *attributed* list (with molecular attribution performed with another software), it will then be identified as *attributed*. The loading of files is typically carried out from the "File" menu, and when an input file meets the specific data architecture of *attributed* or non-attributed *mass list* described in the data architecture paragraph (and SI), the import function will automatically recognize and identify it (as *attributed* or *mass list*) and select the appropriate loading and formatting method. For a user-friendly and easy recognition, the data type is transcribed in the *loaded files* section of the GUI using a color code on the file name (blue for non-attributed *mass lists* and pale yellow for *attributed* mass lists) (Figure 2). Peak lists are directly obtained through the user's routine mass spectrometric data treatment software (typically after noise detection, peak picking, creation of centroid *m/z* peak position lists with intensities).

However, for raw peak lists (in PyC2MC referred to as *mass lists*), only several basic functions of PyC2MC are available: Kendrick's mass defect plots and statistical analysis features such as volcano plot and Venn diagram. Indeed, PyC2MC, for now, does not support the attribution of peak lists to molecular formulae. Lists of attributed molecular formula (referred in PyC2MC as *attributed*) are obtained from vendor software or other software workflows, as outlined above. Instead of directly importing them, the data files may be merged by different processes to create a new data file. Most of the comparison and statistical analysis functions rely on input files created by the PyC2MC functionality of merging files, thus creating implicit 3D matrices, also called hypercubes as input [35]. Input files can be merged using one of the merge functions available. In the GUI, the function "Fuse replicates" is used to create one data file from several *attributed* mass lists obtained from analytical replicates (n) of the same sample. It results in a .csv file composed of the list of every ion observed in at least X replicate(s), X being an integer number chosen by the user ($1 \leq X \leq n$). Compounds not found in a sufficient number of replicates are discarded. The resulting file contains the molecular formulas, exact masses (*m/z*), arithmetic mean intensities, and individual intensity values of each replicate. This type of file is subsequently identified as *Fused*. This function is meant for visualizing and exporting a series of replicates as one single averaged data set file and not for comparing different samples. Rather than choosing one replicate for a fingerprint plot, the *Fused* file allows the selection and illustration of robust data from a number of technical replicates.

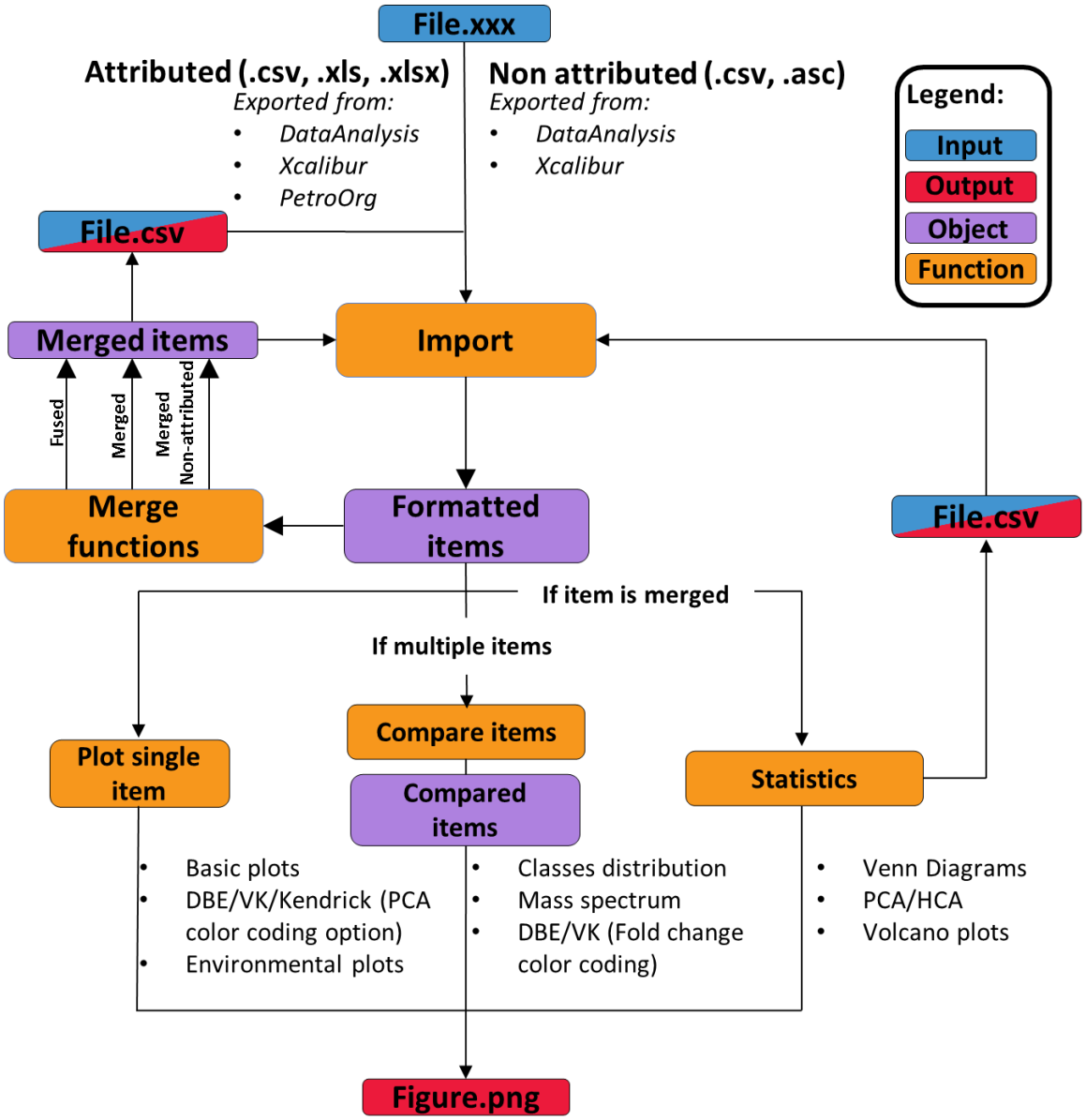


Figure 1: Workflow diagram of the PyC2MC software. Attributed – list of elemental composition resulting from molecular formula assignment, non-attributed – peak-picked mass spectrometric data containing position as m/z and abundance (I), DBE – double bond equivalents, VK – van Krevelen.

In the "Process" menu, two other merging functions called "Merge files (with attributions)" or "Merge files (without attributions)" were designed for inter-sample comparison. These functions create one file encompassing all entries of the selected files, their intensities in each file, and the other data, such as molecular formula attribution, in the case of attributed data. The resulting files are identified either as *Merged* or as *Merge non-attributed* appearing respectively with a green or grey/violet color.

Then, depending on the type and quantity of loaded items (*Mass list*, *Attributed*, *Fused*, *Merged,* or *Merge non-attributed,* each identified in the Loaded files section of the GUI by a specific color), specific functionalities are enabled, as seen on Figure 1. For example, merged items allow for statistical analysis such as PCA and HCA.

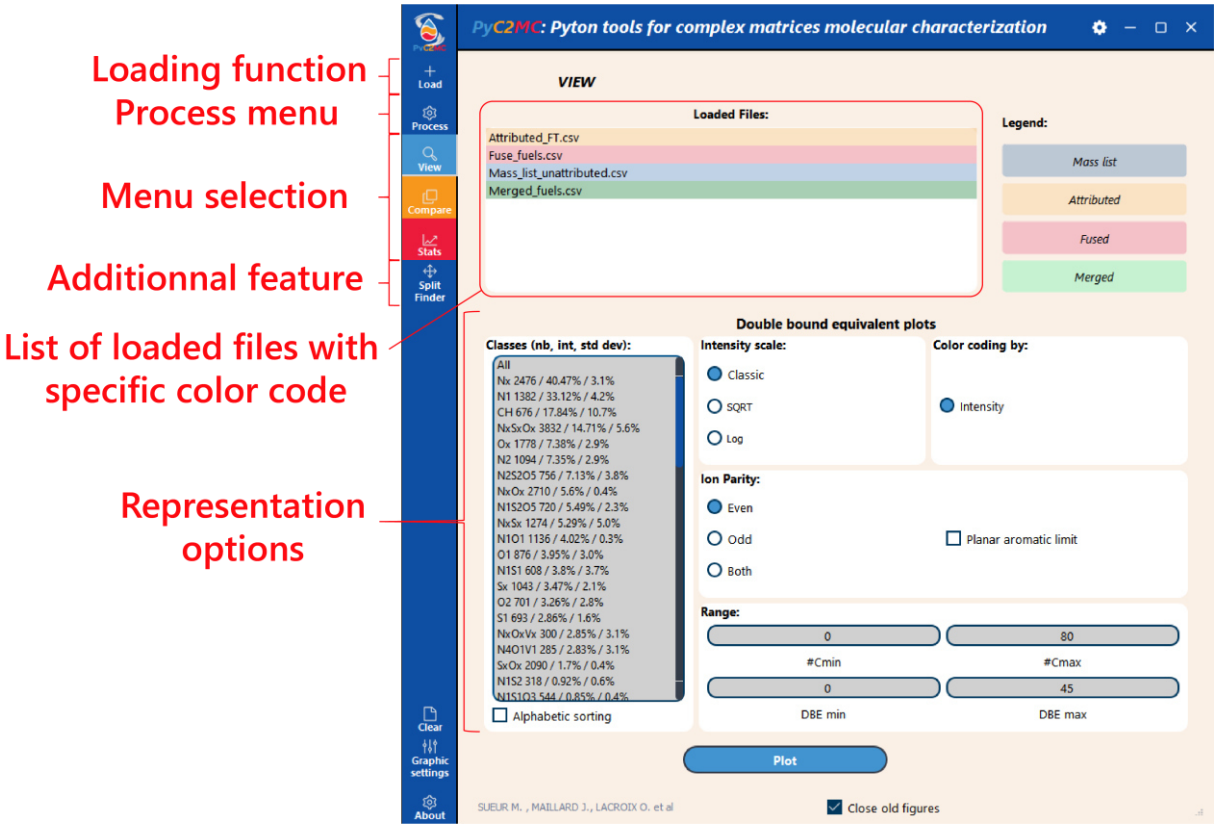


Figure 2: Graphical user interface (GUI) of the PyC2MC viewer software package.

As stated before, a minimalist GUI is provided within the software and is shown in Figure 2. Aside from combining all functionalities, this user-friendly interface requires no dedicated and lengthy software training. The program functionalities are divided into three parts: “View”, “Compare” and “Stats”, each one corresponding to a menu accessible with the corresponding button on the left side of the GUI. The interface allows the users to directly select the part of the application they want to use, the file(s) they want to visualize among the loaded ones, the type of visualization, and finally, the options of the selected representation, if any.

Fundamental Visualization. Molecular attribution lists from the analysis of complex mixtures are classically visualized by common fingerprint approaches. Different plots may be built utilizing general parameters of the molecular formulae, such as carbon number (#C), unsaturation degree or double bond equivalents (DBE), or elemental ratios (H/C, O/C). Selected examples of these basic representations are represented in Figure 3 and

available in the GUI under the "View" menu. This menu encompasses all plots that can be created using only one file containing attributions either from one sample/replicate, or more in the case of a *fused* or *merged* input file. Moreover, PyC2MC allows to create data evaluation representations like error plots: relative error (in ppm) between the theoretical *m/z* ratio calculated with the molecular formula and the observed *m/z* ratio against the observed *m/z* ratio. Using the same information, three variants of the error plot are feasible: the error distribution in a histogram or a boxplot (Figure S2), as well as error distribution versus *m/z* and abundance as the color coding variable, to evidence *m/z*-dependencies. The application can also print the mass spectrum of a selected file. Here, color-coding of signals belonging to a specific compound class, as seen in Figure 3a with the peaks of the O_x and N_xO_y colored respectively in blue and red, allow for accessible insights into complex spectral information. For first insights into the chemical composition and a rapid overview, this functionality can be beneficial, *e.g.*, this can be used to demonstrate the predominance of a compound class of interest or highlight low abundant classes within dominating molecular series.

Another approach to data reduction is to visualize the compound class distribution bar plot (summed relative or absolute abundance of all attributed signals belonging to the same compound class) shown in Figure 3b. In these plots, the data is reduced by summing the intensities of molecular formulae belonging to the same class. Other measures are also used to present the distribution of other characteristic values such as the DBE or the number of carbon atoms in the molecular formulae (Figure 3c).

PyC2MC can also build diagrams that have become typical for complex matrices analysis such as the DBE versus carbon number maps, van Krevelen plots and Kendrick mass defect plots [36]. In such representations, a color code is usually used to represent the intensity of each species. DBE versus carbon number maps (Figure 3d) allow to bring out repetitive moieties and are useful to highlight alkylated series. The planar limit line (red line in Figure 3d) can be utilized for planar limit-assisted structural interpretation [36]. For these maps, the user can apply a filter to only display data relative to one compound class of interest. The van Krevelen plot consists in plotting the H/C ratio, or the ratio of a heteroelement to carbon, against the ratio of another heteroelement to carbon. Usually, the selected axes are H/C and O/C, giving a representation related to the degree of saturation versus oxygen content of the compounds, exemplarily given in Figure 3e. However, it is also possible to plot an atom number ratio against other parameters, such as *m/z*.

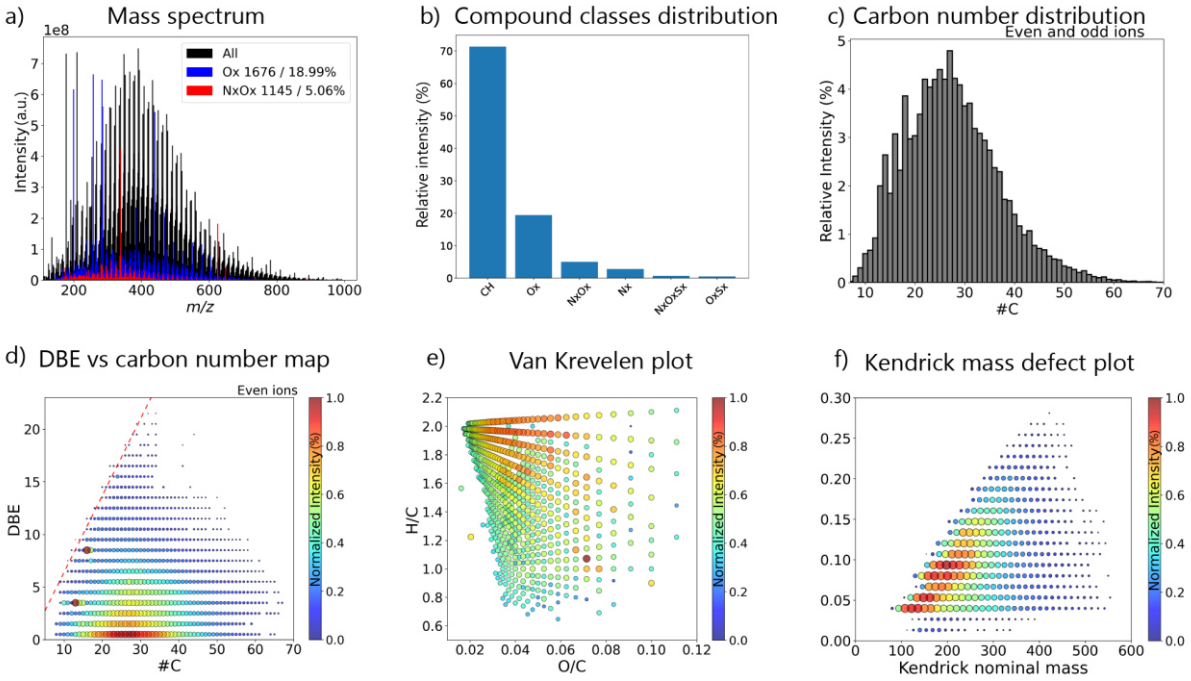


Figure 3: Basic visualization plotted using the Petroleomics dataset: a) Mass spectrum with overlaid compound classes (Compound class, number of peaks / Relative intensity). b) Compound class distribution. c) Distribution of the number of carbon atoms. d) DBE versus C# map with the aromatic planar limit (red dotted line) showing the maximum DBE value at given C# for aromatic compounds. e) Van Krevelen plot displaying alkylation (H/C) versus oxidation (O/C). f) Kendrick plot with -CH_2 as the repetition unit highlighting alkylated series without needing molecular formulas.

Finally, PyC2MC computes for Kendrick mass defect plots as displayed in Figure 3f, based on the calculation of the Kendrick mass defect (KMD) with Equation (1):

$$KMD = Nominal\ Kendrick\ mass - Kendrick\ mass \tag{1}$$

with:

$$Kendrick\ mass\ = observed\ mass\ \times \frac{Repetition\ pattern\ nominal\ mass}{Repetition\ pattern\ exact\ mass} \tag{2}$$

and:

$$Kendrick\ nominal\ mass = rounded\ Kendrick\ mass \tag{3}$$

The Kendrick nominal mass can be obtained from equation (3) using two rounding methods: either round to the closest integer or round to the upper integer (equivalent to rounding the nominal mass to the lower integer). Both methods have their limits [37] and are available within the software. Kendrick mass defect plots consist in plotting KMD versus Kendrick nominal masses, which allows species differing only by *n* repetition pattern(s) to be displayed on the same horizontal line. Historically and still often used, the repetition pattern is CH_2; in this case, KMD plots highlight alkylated series, just as DBE versus carbon number maps. However, the repeating unit can be changed to exhibit any other interesting particularity of the sample, such as methoxy moieties (CHO), oxidation (O), or any polymeric building units (*e.g.*, polystyrene C_8H_8) [38]. Contrary to most other visualization concepts, KMD plots do not require prior molecular formula attribution and thus can be used with non-attributed mass lists.

KMD extraction for molecular formula validation. KMD plots built from non-attributed mass lists, using only the *m/z* ratio (and the intensity as a color code), may be used with PyC2MC to assist the molecular formula attribution. Indeed, the KMD value of a compound of interest can be computed and highlighted on a KMD plot, allowing one to point out homolog series with the same KMD value (same repetitive feature, such as alkylation CH_2). Non-attributed peaks and compounds sharing the same KMD can be extracted and exported into a *.csv* file and used as input in a molecular attribution software or for further cross-software usage. Classically in KMD workflows, one ion for which the molecular formula has been confirmed, may be used as starting point to deduce the molecular formulae of neighboring compounds in the KMD plots. As the KMD is calculated using a chosen repetition pattern, the user can easily find the signals corresponding to their starting compound plus or minus the chosen repetition unit and manually attribute the molecular formula in their molecular attribution software with this guidance. Figure 4 illustrates this functionality with metal-organic petroporphyrins contained in a complex organic matrix. Knowing the *m/z* ratio of the most intense porphyrin signal in this dataset ($C_{28}H_{28}N_4VO$: *m/z* 487.16972), we were able to identify not only the other porphyrins belonging to the same alkylated series (red dots) but also porphyrins of higher DBE, *i.e.*, with one or more additional carbon atoms. This functionality was particularly useful in this case, as the attribution of petroporphyrins requires validations using the isotopic fine structure [39]. This is impossible in the case of low-intensity signals but easily achievable with this Kendrick homologue row approach.

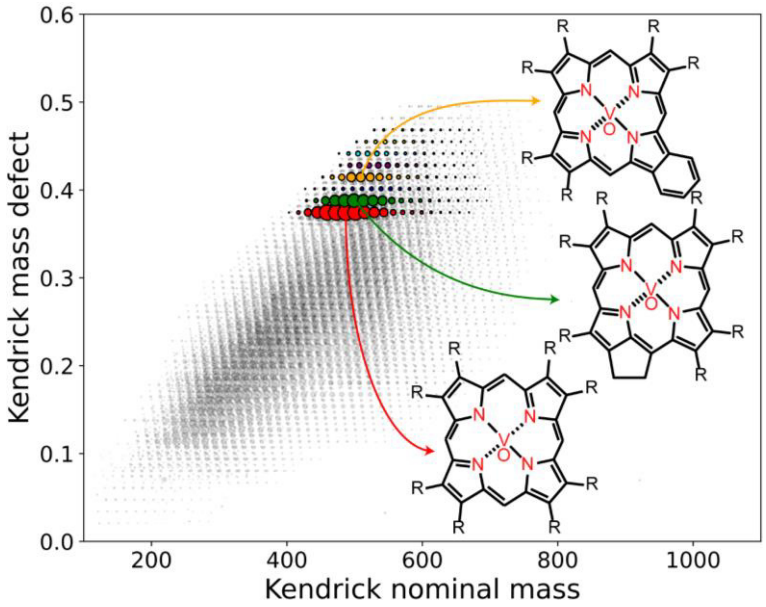


Figure 4: PyC2MC allows to extract homologue rows on non-attributed data via visualization of KMD (CH_2 – alkylation) series. Here, the challenge of metal-organic compounds (petroporphyrins) within a complex organic matrix (bunker fuel combustion aerosol). This graphical approach facilitates molecular formula attribution or can be used as additional validation and evaluation tool.

Variables related to environmental sciences. Another feature of PyC2MC is the "Environmental science variables" section of the 'View' menu, which computes parameters and variables retrieved from the molecular formula attribution commonly utilized in environmental sciences, such as dissolved organic matter (DOM) or particulate matter (aerosol, PM) research. Namely, the average carbon oxidation state (OSC) [15], the modified aromaticity index (MAI) [16], or the maximum carbonyl ratio (MCR) [17]. Figure 5 is an example of an often-used representation (average carbon oxidation state as function of the carbon atom number or Kroll plot) from which organic aerosol classes can be easily identified and information on the chemical nature of the sample material retrieved. Here, the diagram was plotted using data from an environmental dataset, specifically extracts of ambient particulate matter sampled in areas strongly affected by wildfires [31]. Other helpful visualizations proposed in this context and given by the software are Van Krevelen-type diagrams with a color code corresponding to limit values of either MAI or MCR, see Figure S3.

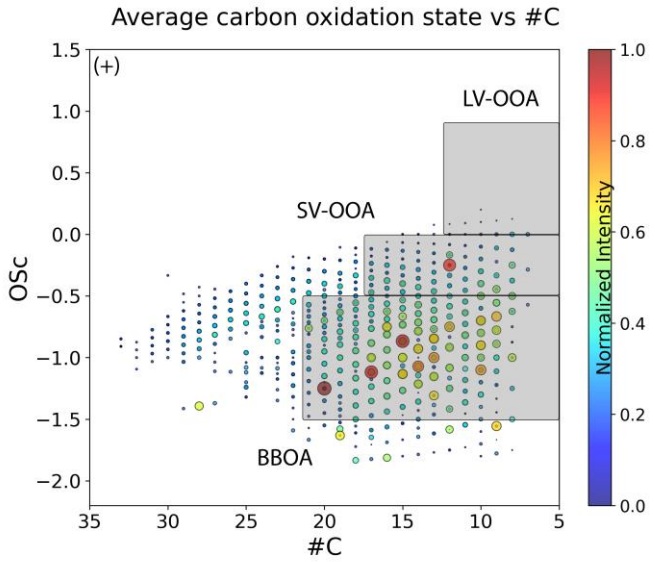


Figure 5: Determining the compound classes within an ambient aerosol sample using the average carbon oxidation state versus carbon number plot (so-called Kroll plot). LV-OOA: Low-volatility oxidized organic aerosol; SV-OOA: Semi-volatile oxidized organic aerosol and BBOA: Biomass burning organic aerosol.

Primary comparative features. For the features in the compare menu, the multiple files to be compared should be loaded in the tabular section at the left center of the interface (Figure 2), and the user should select the files they want to compare among the loaded ones. A merging is then performed automatically using the same principle as the previously mentioned "Merge files (with attributions)" function but no new csv files are created, which lightly reduces computing time, thus allowing a faster comparison. Once these import and preprocessing operations are completed, the user is notified by an indicator and can plot several representations such as chemical class distribution or the DBE distribution (Figure S4) as well as, DBE versus #C maps and van Krevelen plots using a calculated fold change (FC) as color coding. The fold change is derived from equation (4), from the intensity ratio of each peak between two analyses among the merged. According to the petroinformatics principles nicely described by Hur et al. [40], the binary logarithm of the FC (log2[FC]) can be used as color coding to emphasize similarities and differences in the common chemical space based on intensity/abundance.

$$FC = \frac{Peak\ intensity\ in\ sample\ 2}{Peak\ intensity\ in\ sample\ 1} \quad (4)$$

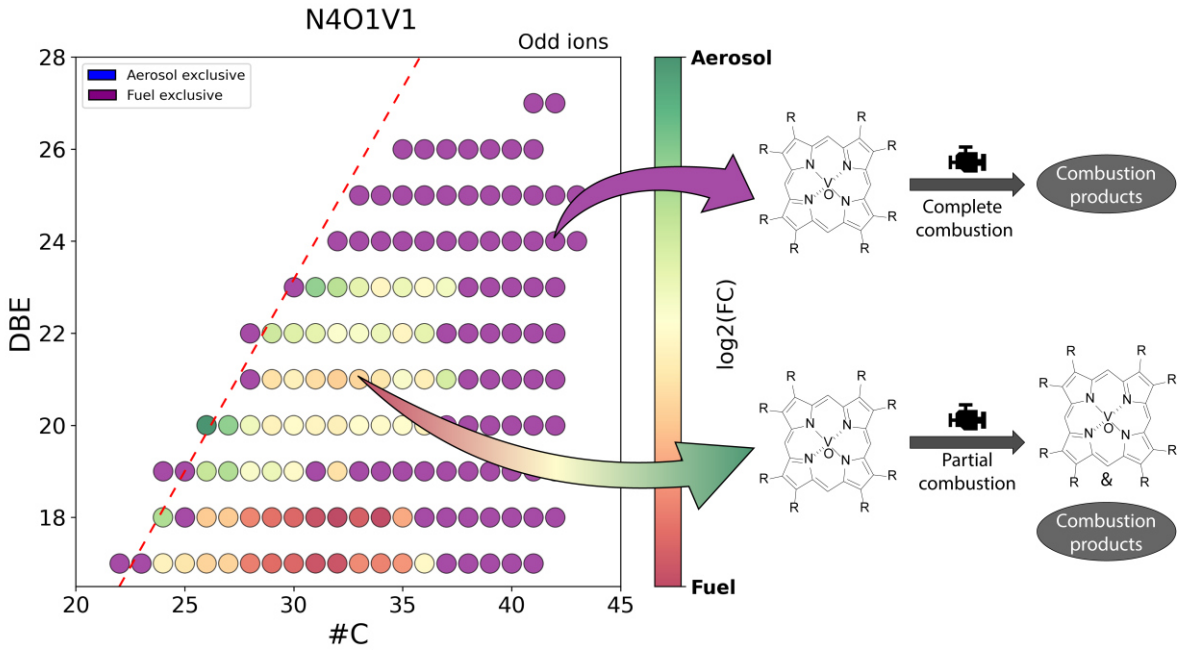


Figure 6: Exemplary utilization of DBE versus #C fingerprint diagram with color-coded fold change information for insights into the molecular fate of petroporphyrins through the combustion of bunker fuels in a ship diesel engine (feed fuel versus primary combustion aerosol).

This kind of representation graphically summarizes in which sample each molecular feature is given with a higher relative/absolute abundance and to which extent. The petroporphyrins dataset, used in the previous section for extraction of KMD series and improved molecular attribution, was used to plot the DBE vs #C map of extracted metal-organic constituents presented in Figure 6. For this figure, the data from two samples were selected to be compared: a feed bunker ship fuel (heavy fuel oil, fuel sulphur content > 0.5 w-%) and its primary combustion aerosol gathered at the engine exhaust of a research ship diesel engine.[41] The FC color coding in Figure 6, shows that most porphyrins were solely detected in the feed fuel (purple dots) or detected with lower intensity (red dots) in the corresponding aerosol emission, suggesting an overall total consumption through combustion.

Statistical tools and analysis. Contrary to the "Compare" menu, which allows fast and intuitive comparison, from multiple loaded files, the "Stats" menu allows more complex comparison but from one single file that should result from one of the merge processes: *merged* or *merged non attributed*. Note that the single merged file includes information on each replicate. Thus, the user is able to plot Venn diagrams and volcano plots and compute

unsupervised multivariate statistics for dimensionality reduction, such as principal component analysis (PCA) or hierarchical cluster analysis (HCA). PCA and HCA are performed using the Scikit-learn [26] algorithm on the relative intensity (normalized to the summed intensity of each peak) values of each peak in each sample. The user can choose to perform the analysis on every peak in the merged file or only on those common to every sample. In the first case, when a peak is not in a sample, its intensity is set to zero. The calculated PCA loadings of each component can then be exported in a csv file to be used in basic representations under the "View" tab (namely DBE vs #C maps, van Krevelen plots, and average carbon oxidation state vs #C) where the PCA loadings can be used as the color-coding variable [32]. The results of HCA are represented in a dendrogram plot with Euclidian distance as the variable characterizing the dissimilarity between samples (Figure S5). Concerning the Venn diagrams, an .xlsx file can be exported containing the species found in each region of the Venn diagram. Finally, echoing the fold change color-coded plot in the "Compare" section, volcano plots [42] can be built using various color-coding options such as the compound class as displayed in Figure 7a where the oxygen-containing compounds of the aged tholin sample [34] are color-coded in yellow. It is also possible to select another parameter as the color-coding variable *e.g.*, *m/z* ratio, DBE or O/C ratio (Figure 7b).

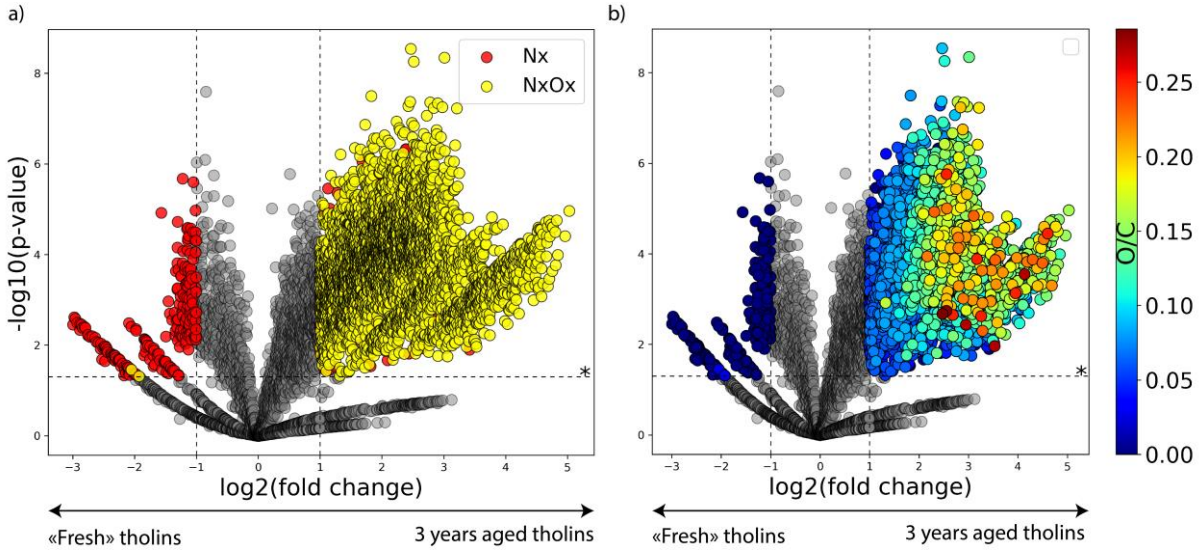


Figure 7: Volcano plot created via PyC2MC exhibiting the oxidation process of an astrochemical mimicked aerosol sample, so-called Tholins. On the left: "Fresh" tholin sample. On the right: 3 years-old tholin sample. Each sample here consists in the arithmetic mean of 3 technical replicates. Red dots: Species containing only nitrogen. Yellow dots: Species containing nitrogen and oxygen.

Additional feature. Alongside the visualization features, PyC2MC proposes an additional feature named Split Finder, which searches for a given pair of atoms (isotopes, for example) corresponding to a given $\Delta m/z$ within a tolerance. This feature has a separate GUI on which the user can input a $\Delta m/z$ observed on a mass spectrum and a tolerance value (10^{-4} Da by default). This feature will then search in a database [43] comprising the molecular mass with atomic masses and isotopic compositions of each element and return the user a list of the possible atom couples, highlighting the solution with less error. A screenshot of the GUI and the results for a $\Delta m/z$ of 1.003355 and a tolerance of 5.10^{-4} Da, *i.e.,* $^{12}C/^{13}C$ split, is shown in the supplementary information (Figure S6).

CONCLUSION

We developed an open-access program dedicated to the visualization and processing of highly complex high-resolution mass spectrometry data. On the one hand, the program can be handled by users with little to no programming experience, but on the other hand, it is improvable and/or adaptable for the community with knowledge of Python. The capabilities of this application have been demonstrated over a wide range of samples (plastic pyrolysis oil, fuel, ambient air, and Tholins simulating Titan's atmosphere). The intention of this software is to be useful for anyone treating high-resolution mass spectrometry data of complex matrices and, to do so, both standalone application and source code will be available at the following GitHub repository: (https://github.com/iC2MC/PyC2MC_viewer). The software will be continuously updated; thus, we also welcome and encourage any contribution to the repository.

ACKNOWLEDGMENTS

This work has been partially supported by the European Regional Development Fund (ERDF, HN0001343), Labex SynOrg (Grant ANR-11- LABX-0029), Carnot Institute I2C, the Graduate School for Research XL-Chem (Grant ANR-18EURE-0020), the European Union's Horizon 2020 Research Infrastructures program (Grant Agreement 731077), and Région Normandie. Access to the CNRS research infrastructure Infranalytics (FR2054) is gratefully acknowledged. We thank the DFG (ZI 764/28-1) and ANR (ANR-20-CE92-0036) for funding the research project TIMSAC.

References

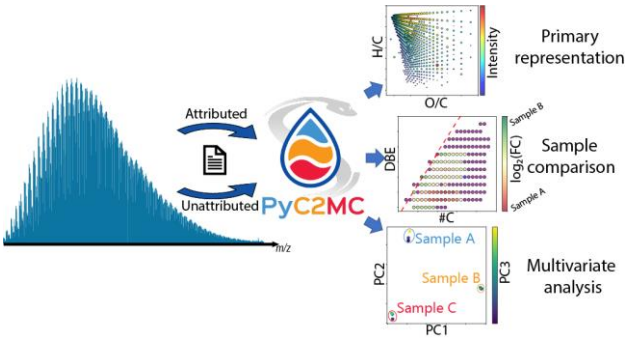
(1) Stopka, S. A.; Samarah, L. Z.; Shaw, J. B.; Liyu, A. V.; Veličković, D.; Agtuca, B. J.; Kukolj, C.; Koppenaal, D. W.; Stacey, G.; Paša-Tolić, L.; et al. Ambient Metabolic Profiling and Imaging of Biological Samples with Ultrahigh Molecular Resolution Using Laser Ablation Electrospray Ionization 21 Tesla FTICR Mass Spectrometry. *Analytical Chemistry* **2019**, *91* (8), 5028-5035. DOI: 10.1021/acs.analchem.8b05084.
(2) Rüger, C. P.; Sklorz, M.; Schwemer, T.; Zimmermann, R. Characterisation of ship diesel primary particulate matter at the molecular level by means of ultra-high-resolution mass spectrometry coupled to laser desorption ionisation--comparison of feed fuel, filter extracts and direct particle measurements. *Anal Bioanal Chem* **2015**, *407* (20), 5923-5937. DOI: 10.1007/s00216-014-8408-1.
(3) Marshall, A. G.; Rodgers, R. P. Petroleomics: Chemistry of the underworld. **2008**, *105* (47), 18090-18095. DOI: doi:10.1073/pnas.0805069105. Cho, Y.; Ahmed, A.; Islam, A.; Kim, S. Developments in FT-ICR MS instrumentation, ionization techniques, and data interpretation methods for petroleomics. *Mass Spectrometry Reviews* **2015**, *34* (2), 248-263. DOI: 10.1002/mas.21438.
(4) Lacroix-Andrivet, O.; Maillard, J.; Mendes Siqueira, A. L.; Hubert-Roux, M.; Loutelier-Bourhis, C.; Afonso, C. Molecular Characterization of Aged Bitumen with Selective and Nonselective Ionization Methods by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. 2. Statistical Approach on Multiple-Origin Samples. *Energy & Fuels* **2021**, *35* (20), 16442-16451. DOI: 10.1021/acs.energyfuels.1c02503.
(5) Rüger, C. P.; Schwemer, T.; Sklorz, M.; O'Connor, P. B.; Barrow, M. P.; Zimmermann, R. Comprehensive chemical comparison of fuel composition and aerosol particles emitted from a ship diesel engine by gas chromatography atmospheric pressure chemical ionisation ultra-high resolution mass spectrometry with improved data processing routines. *European journal of mass spectrometry (Chichester, England)* **2017**, *23* (1), 28-39. DOI: 10.1177/1469066717694286 From NLM. He, C.; Fang, Z.; Li, Y.; Jiang, C.; Zhao, S.; Xu, C.; Zhang, Y.; Shi, Q. Ionization selectivity of electrospray and atmospheric pressure photoionization FT-ICR MS for petroleum refinery wastewater dissolved organic matter. *Environmental Science: Processes & Impacts* **2021**, *23* (10), 1466-1475, 10.1039/D1EM00248A. DOI: 10.1039/D1EM00248A. Bianco, A.; Riva, M.; Baray, J.-L.; Ribeiro, M.; Chaumerliac, N.; George, C.; Bridoux, M.; Deguillaume, L. Chemical Characterization of Cloudwater Collected at Puy de Dôme by FT-ICR MS Reveals the Presence of SOA Components. *ACS Earth and Space Chemistry* **2019**, *3* (10), 2076-2087. DOI: 10.1021/acsearthspacechem.9b00153.
(6) Marshall, A. G. Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Acc. Chem. Res.* **1985**, *18*, 316-322. Marshall, A. G.; Chen, T. 40 years of Fourier transform ion cyclotron resonance mass spectrometry. *International Journal of Mass Spectrometry* **2015**, *377*, 410-420. DOI: 10.1016/j.ijms.2014.06.034. Nikolaev, E. N.; Boldin, I. A.; Jertz, R.; Baykut, G. Initial Experimental Characterization of a New Ultra-High Resolution FTICR Cell with Dynamic Harmonization. *Journal of The American Society for Mass Spectrometry* **2011**, *22* (7), 1125-1133. DOI: 10.1007/s13361-011-0125-9. Makarov, A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal Chem* **2000**, *72* (6), 1156-1162. DOI: 10.1021/ac991131p. Denisov, E.; Damoc, E.; Lange, O.; Makarov, A. Orbitrap mass spectrometry with resolving powers above 1,000,000. *International Journal of Mass Spectrometry* **2012**, *325-327*, 80-85. DOI: 10.1016/j.ijms.2012.06.009.
(7) *PetroOrg*; 2012. (accessed.
(8) Orthous-Daunay, F.-R.; Thissen, R.; Vuitton, V. Measured mass to stoichoimetric formula through exhaustive search. *Proceedings of the International Astronomical Union* **2019**, *15* (S350), 193-199. DOI: 10.1017/s1743921319008032.
(9) Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature Methods* **2016**, *13* (9), 741-748. DOI: 10.1038/nmeth.3959.
(10) Niedermeyer, T. H. J.; Strohalm, M. mMass as a Software Tool for the Annotation of Cyclic Peptide Tandem Mass Spectra. *PLoS ONE* **2012**, *7* (9), e44913. DOI: 10.1371/journal.pone.0044913.

(11) Merder, J.; Freund, J. A.; Feudel, U.; Hansen, C. T.; Hawkes, J. A.; Jacob, B.; Klaproth, K.; Niggemann, J.; Noriega-Ortega, B. E.; Osterholz, H.; et al. ICBM-OCEAN: Processing Ultrahigh-Resolution Mass Spectrometry Data of Complex Molecular Mixtures. *Analytical Chemistry* **2020**, *92* (10), 6832-6838. DOI: 10.1021/acs.analchem.9b05659.
(12) Vetter, W.; McLafferty, F. W.; Turecek, F. Interpretation of mass spectra. Fourth edition (1993). University Science Books, Mill Valley, California. *Biological Mass Spectrometry* **1994**, *23* (6), 379-379. DOI: 10.1002/bms.1200230614. Le Maitre, J.; Hubert-Roux, M.; Paupy, B.; Marceau, S.; Ruger, C. P.; Afonso, C.; Giusti, P. Structural analysis of heavy oil fractions after hydrodenitrogenation by high-resolution tandem mass spectrometry and ion mobility spectrometry. *Faraday Discuss* **2019**, *218* (0), 417-430. DOI: 10.1039/c8fd00239h. Chacon-Patino, M. L.; Rowland, S. M.; Rodgers, R. P. Advances in Asphaltene Petroleomics. Part 1: Asphaltenes Are Composed of Abundant Island and Archipelago Structural Motifs. *Energy & Fuels* **2017**, *31* (12), 13509-13518, Article. DOI: 10.1021/acs.energyfuels.7b02873.
(13) Kim, S.; Kramer, R. W.; Hatcher, P. G. Graphical Method for Analysis of Ultrahigh-Resolution Broadband Mass Spectra of Natural Organic Matter, the Van Krevelen Diagram. *Analytical Chemistry* **2003**, *75* (20), 5336-5344. DOI: 10.1021/ac034415p. Van Krevelen, D. W. Graphical-statistical method for the study of structure and reaction processes of coal. *Fuel* **1950**, *29*, 269-284.
(14) Kendrick, E. A Mass Scale Based on CH2 = 14.0000 for High Resolution Mass Spectrometry of Organic Compounds. *Analytical Chemistry* **1963**, *35* (13), 2146-2154. DOI: 10.1021/ac60206a048. Hughey, C. A.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G.; Qian, K. Kendrick Mass Defect Spectrum: A Compact Visual Analysis for Ultrahigh-Resolution Broadband Mass Spectra. *Analytical Chemistry* **2001**, *73* (19), 4676-4681. DOI: 10.1021/ac010560w.
(15) Kroll, J. H.; Donahue, N. M.; Jimenez, J. L.; Kessler, S. H.; Canagaratna, M. R.; Wilson, K. R.; Altieri, K. E.; Mazzoleni, L. R.; Wozniak, A. S.; Bluhm, H.; et al. Carbon oxidation state as a metric for describing the chemistry of atmospheric organic aerosol. *Nature Chemistry* **2011**, *3* (2), 133-139. DOI: 10.1038/nchem.948.
(16) Koch, B. P.; Dittmar, T. From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. *Rapid Communications in Mass Spectrometry* **2006**, *20* (5), 926-932. DOI: 10.1002/rcm.2386.
(17) 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 Communications in Mass Spectrometry* **2021**, *35* (14). DOI: 10.1002/rcm.9113.
(18) Colby, S. M.; Chang, C. H.; Bade, J. L.; Nunez, J. R.; Blumer, M. R.; Orton, D. J.; Bloodsworth, K. J.; Nakayasu, E. S.; Smith, R. D.; Ibrahim, Y. M.; et al. DEIMoS: An Open-Source Tool for Processing High-Dimensional Mass Spectrometry Data. *Analytical Chemistry* **2022**, *94* (16), 6130-6138. DOI: 10.1021/acs.analchem.1c05017.
(19) Letourneau, D. R.; Volmer, D. A. Constellation: An Open-Source Web Application for Unsupervised Systematic Trend Detection in High-Resolution Mass Spectrometry Data. *J Am Soc Mass Spectrom* **2022**. DOI: 10.1021/jasms.1c00371.
(20) 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.
(21) Kitson, E.; Kew, W.; Ding, W.; Bell, N. G. A. PyKrev: A Python Library for the Analysis of Complex Mixture FT-MS Data. *Journal of the American Society for Mass Spectrometry* **2021**, *32* (5), 1263-1267. DOI: 10.1021/jasms.1c00064.
(22) Gavard, R.; Jones, H. E.; Palacio Lozano, D. C.; Thomas, M. J.; Rossell, D.; Spencer, S. E. F.; Barrow, M. P. KairosMS: A New Solution for the Processing of Hyphenated Ultrahigh Resolution Mass Spectrometry Data. *Analytical Chemistry* **2020**, *92* (5), 3775-3786. DOI: 10.1021/acs.analchem.9b05113.

(23) Heberle, H.; Meirelles, G. V.; Da Silva, F. R.; Telles, G. P.; Minghim, R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* **2015**, *16* (1). DOI: 10.1186/s12859-015-0611-3.
(24) McKinney, W. Data Structures for Statistical Computing in Python. 2010, SciPy. DOI: 10.25080/majora-92bf1922-00a.
(25) Harris, C. R.; Millman, K. J.; Van Der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; et al. Array programming with NumPy. *Nature* **2020**, *585* (7825), 357-362. DOI: 10.1038/s41586-020-2649-2.
(26) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. **2011**.
(27) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **2007**, *9* (3), 90-95. DOI: 10.1109/MCSE.2007.55.
(28) Boyer, G. *Chemparse*. 2020.
(29) Raybaut, P. Spyder-documentation. *Available Online at: Pythonhosted. Org* **2009**.
(30) Cortesi, D. PyInstaller Manual — PyInstaller 3.2.1 documentation. **2017**.
(31) 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.
(32) Castilla, C.; Rüger, C. P.; Marcotte, S.; Lavanant, H.; Afonso, C. Direct Inlet Probe Atmospheric Pressure Photo and Chemical Ionization Coupled to Ultrahigh Resolution Mass Spectrometry for the Description of Lignocellulosic Biomass. *Journal of the American Society for Mass Spectrometry* **2020**, *31* (4), 822-831. DOI: 10.1021/jasms.9b00091.
(33) 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.
(34) 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 and Planetary Science Letters* **2018**, *495*, 185-191. DOI: 10.1016/j.epsl.2018.05.014.
(35) Guillemant, J.; Lacoue-Nègre, M.; Berlioz-Barbier, A.; De Oliveira, L. P.; Albrieux, F.; Joly, J.-F.; Duponchel, L. Evaluating the Benefits of Data Fusion and PARAFAC for the Chemometric Analysis of FT-ICR MS Data Sets from Gas Oil Samples. *Energy & Fuels* **2020**, *34* (7), 8195-8205. DOI: 10.1021/acs.energyfuels.0c01104.
(36) Cho, Y.; Kim, Y. H.; Kim, S. Planar Limit-Assisted Structural Interpretation of Saturates/Aromatics/Resins/Asphaltenes Fractionated Crude Oil Compounds Observed by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Analytical Chemistry* **2011**, *83* (15), 6068-6073. DOI: 10.1021/ac2011685.
(37) Lacroix-Andrivet, O.; Moualdi, S.; Hubert-Roux, M.; Loutelier Bourhis, C.; Mendes Siqueira, A. L.; Afonso, C. Molecular Characterization of Formulated Lubricants and Additive Packages Using Kendrick Mass Defect Determined by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Journal of the American Society for Mass Spectrometry* **2022**, *33* (7), 1194-1203. DOI: 10.1021/jasms.2c00050.
(38) Fouquet, T. N. J. The Kendrick analysis for polymer mass spectrometry. *Journal of Mass Spectrometry* **2019**, *54* (12), 933-947. DOI: 10.1002/jms.4480.
(39) Caumette, G.; Lienemann, C.-P.; Merdrignac, I.; Bouyssiere, B.; Lobinski, R. Element speciation analysis of petroleum and related materials. *Journal of Analytical Atomic Spectrometry* **2009**, *24* (3), 263. DOI: 10.1039/b817888g.
(40) Hur, M.; Kim, S.; Hsu, C. S. Petroinformatics. In *Springer Handbook of Petroleum Technology*, Hsu, C. S., Robinson, P. R. Eds.; Springer International Publishing, 2017; pp 173-198.

(41) Jeong, S.; Bendl, J.; Saraji-Bozorgzad, M.; Kafer, U.; Etzien, U.; Schade, J.; Bauer, M.; Jakobi, G.; Orasche, J.; Fisch, K.; et al. Aerosol emissions from a marine diesel engine running on different fuels and effects of exhaust gas cleaning measures. *Environ Pollut* **2022**, *316* (Pt 1), 120526. DOI: 10.1016/j.envpol.2022.120526.
(42) Hur, M.; Ware, R. L.; Park, J.; McKenna, A. M.; Rodgers, R. P.; Nikolau, B. J.; Wurtele, E. S.; Marshall, A. G. Statistically Significant Differences in Composition of Petroleum Crude Oils Revealed by Volcano Plots Generated from Ultrahigh Resolution Fourier Transform Ion Cyclotron Resonance Mass Spectra. *Energy & Fuels* **2018**, *32* (2), 1206-1212. DOI: 10.1021/acs.energyfuels.7b03061.
(43) Atomic Weights and Isotopic Compositions (version 4.1). *Atomic Weights and Isotopic Compositions*.

For Table of Content Only



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[1,2,3], Christopher P. RÜGER[3,4*], 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] 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 author: christopher.rueger@uni-rostock.de**

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

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 whit 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.

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

Table S 1: 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.

2. Datasets details

Hereafter is the tree structure of the data files used in the main body that will also be available on the GitHub repository of the software.

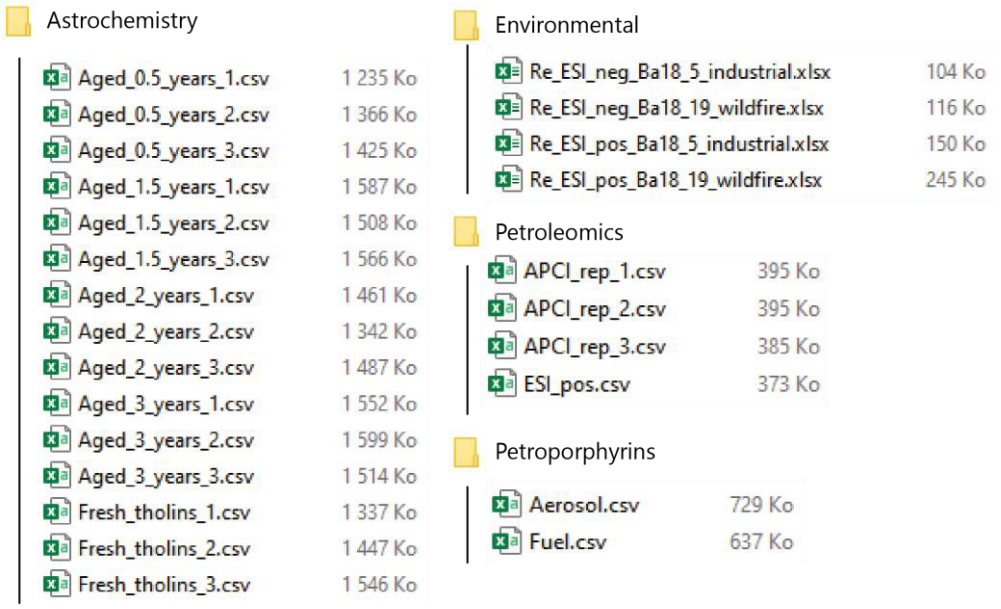


Figure S 1: Tree view of the datasets used

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)	$C_xH_yN_3O_{17}S_2$	$C_xH_yO_2N_{30}$

Table S 2: Details on the datasets used

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 studies.

3. Error plots

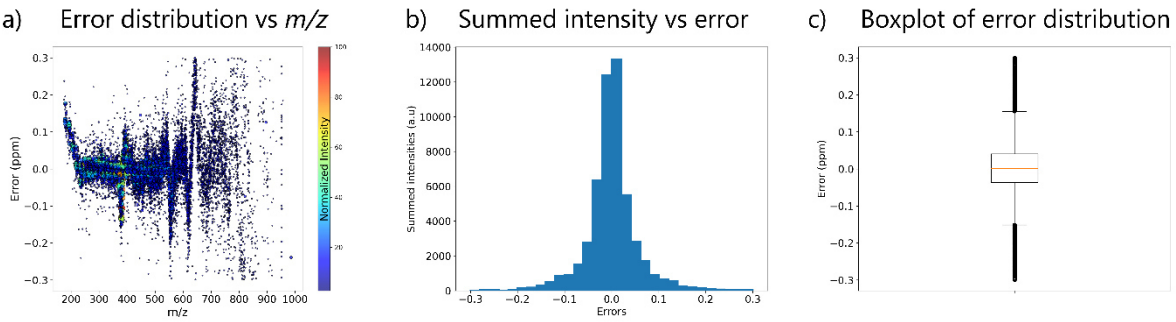


Figure S 2: Three ways to represent the attribution error in PyC2MC: a) error distribution versus m/z ratio with intensity as color code. b) histogram representing the distribution of the intensity versus attribution error. c) Boxplot representation of the error distribution.

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]:

$$MCR = \frac{DBE}{O}, with\ O \neq 0$$

$$If\ O < DBE, then\ MCR = 1$$

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.

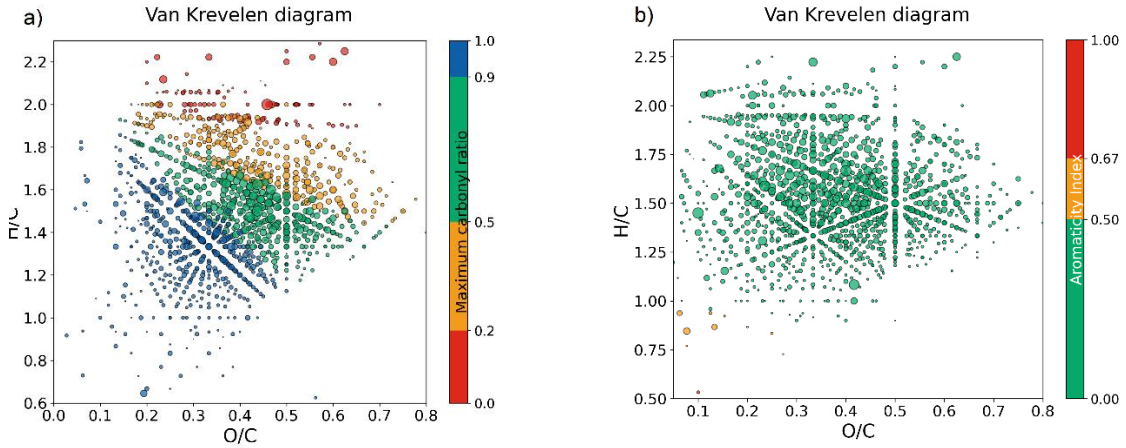


Figure S 3: Van Krevelen diagrams using a) MCR and b)AI as a color-coding variable.

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

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

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

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:

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

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.

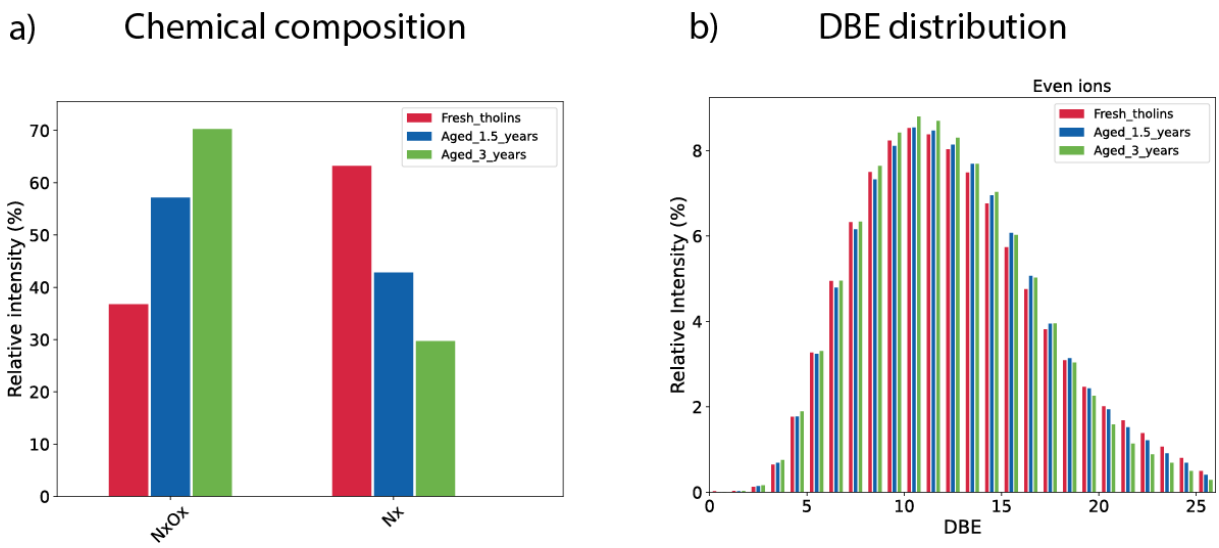


Figure S 4: Demonstration of the inter sample comparison functionalities.

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.

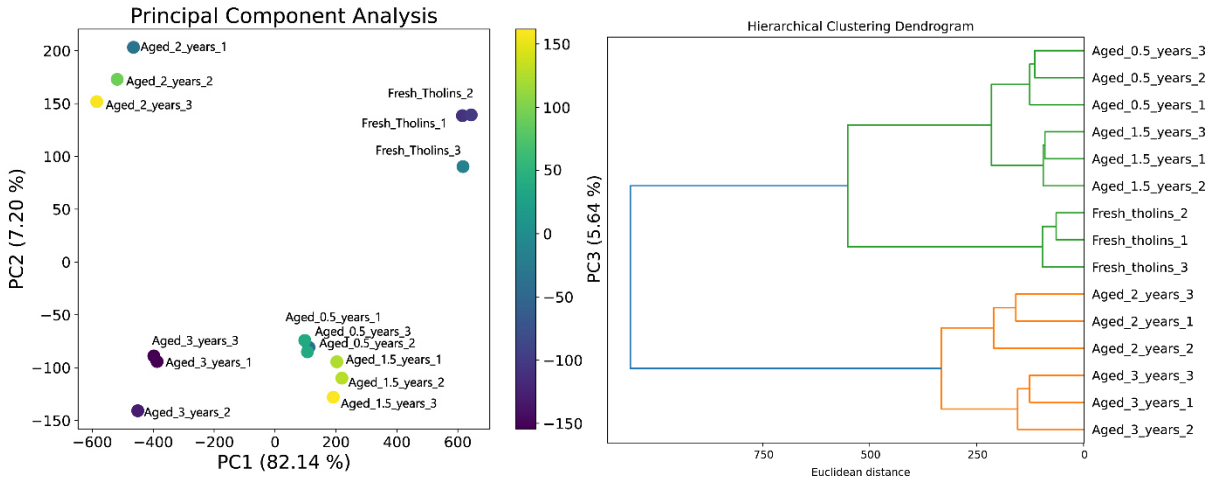


Figure S 5: Statistical analysis features. On the left: Principal Component Analysis (PCA). On the right: Hierarchical Clustering Analysis (HCA).

7. Isotope finder

The additional feature named “Isotope Finder” is provided with its own GUI which is shown in Figure S XX 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 isotopes” 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.

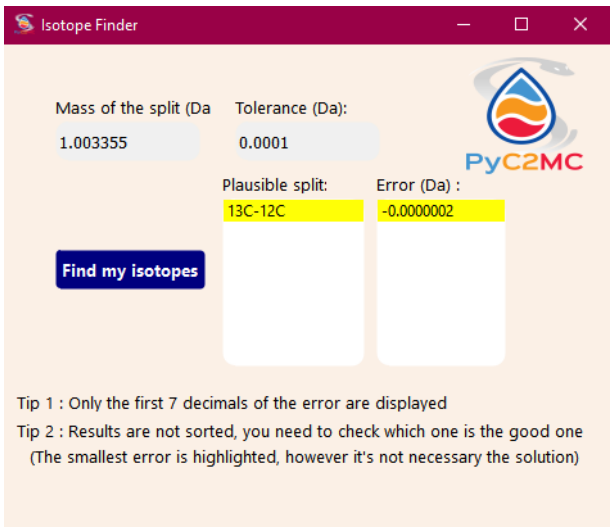


Figure S 6: Isotope finder's GUI developed under PyQT5

References

[1] Mase C, Maillard JF, Paupy B, Farenc M, Adam C, Hubert-Roux M, et al. 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):14828-37.

[2] Sueur M, Rüger CP, Maillard JF, 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.

[3] Schneider E, Czech H, Popovicheva O, Lütdke H, Schnelle-Kreis J, Khodzher T, et al. 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):1095-107.

[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 and Planetary Science Letters 2018;495:185-91.

[5] Zhang Y, Wang K, Tong H, Huang RJ, Hoffmann T. The maximum carbonyl ratio (MCR) as a new index for the structural classification of secondary organic aerosol components. Rapid Communications in Mass Spectrometry 2021;35(14).

[6] Koch BP, Dittmar T. From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. Rapid Communications in Mass Spectrometry 2006;20(5):926-32.