SPIX: a new software package to reveal chemical reactions at trace amounts in very complex mixtures from high-resolution mass spectra data sets

Edith Nicol, Yao Xu, Zsuzsanna Varga, Said Kinani, Stéphane Bouchonnet, Marc Lavielle

SPIX: a new software package to reveal chemical reactions at trace amounts in very complex mixtures from high-resolution mass spectra data sets

Edith Nicol[a]*, Yao. Xu[b,c], Zsuzsanna Varga[a], Said Kinani[d], Stéphane Bouchonnet[a], Marc Lavielle[b,c]

[a] Laboratoire de Chimie Moléculaire, CNRS - IP Paris, Ecole polytechnique, Route de Saclay, 91128 Palaiseau, France

[b] Centre de Mathématiques Appliquées, CNRS - IP Paris, Ecole polytechnique, Route de Saclay, 91128 Palaiseau, France

[c] Inria, École polytechnique, 1 Rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France

[d] Laboratoire National d'Hydraulique et Environnement (LNHE), Division Recherche et Développement, Electricité de France (EDF), 6 Quai de Watier, 78401 Chatou Cedex 01, France

* Corresponding author, e-mail: edith.nicol@polytechnique.edu

Rationale: High-resolution mass spectrometry-based non-targeted screening has a huge potential for applications in environmental sciences, engineering and regulation. However, it produces big data for which full appropriate processing is a real challenge; the development of processing software is the last building-block to enable large-scale use of this approach.

Methods: A new software application, SPIX, has been developed to extract relevant information from high-resolution mass-spectrum datasets. Dealing with intrinsic sample variability and reducing operator subjectivity, it opens up opportunities and promising prospects in many areas of analytical chemistry. SPIX is freely available at: http://spix.webpopix.org.

Results: Two features of the software are presented in the field of environmental analysis. An example illustrates how SPIX reveals photodegradation reactions in wastewater by fitting kinetic models to significant changes in ion abundance over time. A second example shows the ability of SPIX to detect photoproducts at trace amounts in river water, through comparison of datasets from samples taken before and after irradiation.

Conclusions: SPIX has shown its ability to reveal relevant modifications between two series of big data sets, allowing for instance to study the consequences of a given event on a complex substrate. Most of all – and this is to our knowledge the only software currently available allowing that – it can reveal and monitor any kind of reaction in all types of mixture.

Keywords: High-resolution mass spectrometry; non-targeted approach; spectral big data management; software development; chemical reactions; complex mixtures; kinetics

1. Introduction

High-resolution mass spectrometry (HRMS) is now experiencing unprecedented growth. It appeared in the early 1970s with dual-focus devices combining magnetic and electrostatic fields, and continued its development with the introduction of time-of-flight, Orbitraps, and Fourier Transform-Ion Cyclotron Resonance (FT-ICR) analyzers. If FT-ICR mass spectrometers remain the most accurate today, high end QTOFs and QEx Orbitraps provide accuracies below 3 ppm. The resolution of an analyzer reflects its ability to separate ions with close m/z ratios. High-resolution analyzers can thus differentiate isobaric ions: i.e., ions with the same nominal mass but different exact masses, and therefore different chemical formulas, such as $N_2^{+\cdot}$ (m/z 28.0056) and $CO^{+\cdot}$ (m/z 27.9944). High resolution is a very valuable asset: it not only greatly improves the selectivity and specificity of "traditional" detection and quantification methods (in comparison with low-resolution analyzers), but also greatly facilitates structural elucidation by assigning raw formulae to the detected ions.[1]

More recent use of high resolution takes advantage of its ability to separate isobaric ions, in an attempt to break free from separation methods - mainly gas or liquid chromatography, or more rarely capillary electrophoresis or ion mobility – so as to expand the range of molecules detectable in a single analysis. This is particularly interesting in the context of non-targeted analyses – in which the operator does not know which molecules are likely to be present in the sample – because the choice of a chromatographic system focuses the analysis on certain classes of compounds based on their properties (volatile or not, polar or apolar, large or small, etc.) and thereby introduces a bias attributable to the subjectivity of the analyst. Direct introduction (DI) into the ion source without prior pretreatment or chromatographic separation was shown to be a useful alternative for rapid and comprehensive diagnosis of environmental samples, but this approach remains very challenging due to the extreme complexity of environmental matrices and the large number of contaminants likely to be present.[2] On direct introduction of a mixture, different molecules are simultaneously ionized, resulting in mass spectra yielded by the overlapping of spectra of the detectable species. Thus, complex mixture analyses provide mass spectra that can contain tens or hundreds of thousands of ions, even with soft ionization techniques such as electrospray ionization, atmospheric pressure ionization or atmospheric pressure photoionization; these spectra are of no possible use to the operator without the help of adapted software. Finding a molecule showing significant change between two conditions (upstream/downstream or after treatment, for instance) in its trace amounts in an environmental sample is like looking for a needle in a haystack. Being able to quickly evaluate

all of the chemical consequences of an industrial accident on the biotope can be crucial to decision-making. In these situations, non-targeted HRMS-based screening is one of the last resorts for identifying unexpected or unknown contaminants.[3,4,5,6,7,8] This approach has recently been evaluated in a comprehensive collaborative study organized by the NORMAN association, in which a total of 18 institutes from 12 European countries analyzed an extract of the same water sample collected from the Danube River. The results revealed that non-targeted analytical techniques were already widespread and that practices were substantially harmonized between the participants, but that data processing remained complicated and time-consuming.[9] Among the main recommendations formulated to improve the non-targeted approach is the development of robust user-friendly processing software. Likewise, AQUAREF – the French national reference laboratory for aquatic environment monitoring, which works in close concertation with other European reference laboratories – published guidelines for HRMS untargeted analysis, for which SPIX could be a powerful tool.[10]

The first part of this article discusses the notions of uncertainty and subjectivity related to untargeted analysis. The second part presents the general working principle of SPIX software. The third part is dedicated to the presentation of results obtained on real samples. It discusses the strengths and limitations of the software and its specificities compared to the few programs currently commercially available. A brief overview of current computational and statistical approaches to extract relevant information from the *big data* of mass spectrometry analyses is provided in Supplementary information SI-1; it describes Kendrick[11,12] and Van Krevelen[13,14,15] approaches, as well as Multivariate statistical analysis.[16,17,18,19,20,21,22] Multivariate analysis tools enable global understanding of many concomitant variables and of their inter-correlations. Metabolomics processing pipelines often include univariate and multivariate statistical approaches. Univariate analysis is usually used as a pre-processing step, while multivariate analysis is used for classification of samples or features. For example, PCA is used to characterize differences of two groups of metabolomics GC-MS data for the diagnosis of gastric cancer. Wilcoxon rank sum test showed the marker metabolites specific to the tumor group. Multivariate analysis, specifically PCA successfully divided the two groups of samples of normal and malignant gastric tissue.[23] Comprehensive workflow for univariate analysis of LC-HRMS was developed to follow human adult urinary metabolome variations. Univariate analysis was used as a preprocessing step: nonparametric hypothesis testing was used to assess correlations with covariables and Wilcoxon test was used to calculate the median differences between genders. The univariate p-values results together with multivariate importance in projection evidenced that 108 urine metabolites whose concentrations

varied with either age, body mass index, or gender.[24] Concerning direct infusion mass spectrometry a comprehensive workflow for data processing and quality control was developed for metabolomics analysis of cardiac tissue extract. It can be used for different metabolomics analyses as it focuses on the correction of intra- and inter-batch variations and offers best-practice workflows and rigorous quality assessment. The data processing steps include Wilcoxon-test and multivariate analysis.[25] These applications could be extended for environmental samples; however no approach has been reported using univariate or multivariate analysis which focuses on the kinetics of compounds in HRMS datasets. The concept behind multivariate analysis is different from that of the SPIX software: the latter aims at observing all statistically relevant variables individually. Examples of SPIX applications are given in the following article.

2. Notions of uncertainty and subjectivity in modern untargeted analytical approaches, and introduction to SPIX

To illustrate the functionality of the SPIX software, it is necessary to address the notions of uncertainty and subjectivity that are fundamental in analytical chemistry. We propose to take an example in environmental chemistry. Consider a plant located on the bank of a river; it may be a treatment plant or, on the contrary, a source of pollution; the question is whether its presence significantly alters the composition of the water. The question seems simple enough, but providing a relevant answer is much less so. The conventional approach is to take water samples upstream and downstream of the plant, analyze them chemically and compare the results. This approach, while scientifically reasonable, nevertheless raises many questions at each step of the process. How many samples are needed to take account of the spatial and temporal variability of upstream and downstream water composition? Where, when and how to sample? What sample preparation to adopt, given that each choice of solvent, filter, SPE (Solid Phase Extraction) column, chromatographic protocol and mass spectrometry ionization mode conditions the results of the analysis by favoring detection of certain molecules based on their size or polarity. Every single step in the analytical process introduces metrological uncertainties related to the measuring instruments used (balances, pipettes, etc.), but also to the so-called "matrix effect": i.e., the matrix of the reference used to validate the method is generally not rigorously identical to the matrix being analyzed. Stochastic biases and uncertainties are also caused by adsorption, evaporation, etc. The proliferation of sources of error obliges analysts to use internal standards to reduce the overall uncertainty of the results and try to conform to industry-specific standards. Limiting the subjectivity in a method needs to make no assumptions at all, which is in contradiction with the use of an internal

standard; thus, the analytical scientist is left with choosing between limiting subjectivity or limiting uncertainties. To the problem of uncertainties must be added that of operator subjectivity, at two main levels. As mentioned above, this subjectivity comes into play before measurement: when the operator establishes the analytical protocol, choices are made, conditioned by assumptions – the operator's own or those of third parties - as to what might have contaminated the water of the river. Even if the method is not "targeted" (i.e., specifically designed for the selective detection of given analytes), it cannot be considered totally "non-targeted" as there is no effective protocol capable of extracting and detecting everything simultaneously (e.g., both polar and apolar molecules) and any selected protocol effectively excludes some potential analytes. This will lead the analyst to try to minimize sample preparation, with the dual objective of limiting uncertainties and of reducing operator-induced subjectivity; an immediate consequence of this simplification is to increase the complexity of the data. For example, mass spectra recorded from environmental samples will be much more complex if the sample is introduced directly into the mass spectrometer without prior purification and separation. A point that is generally much less considered is operator subjectivity in interpreting results, especially when the data are complex and voluminous, when it comes to manually integrating a peak or comparing two chromatograms or two mass spectra, for example.

In 2019, a visual trial devoted to subjectivity evaluation was carried out during a European winter school on mass spectrometry, on a panel of 37 people with a strong scientific background in analytical chemistry. It consisted of a series of one-minute projections of two images differing by 5 to 22 differences; panelists were asked to note the number of differences they were able to spot. Some images were quite simple (pictures with modified areas) while others were very complex (fractals containing very small differences within complex areas). A set of simulated mass spectra containing 15 differences (variations in peak intensity, addition and removal of peaks) was presented to the panel – in triplicate and not consecutively – without prior notice. The variability between the results of these triplicates gave an average standard deviation of 2.3 observed differences per individual, with mean and median values of 9.6 and 9.7, respectively and a range of 0-19. Considering the variability between panelists, a standard deviation of 20.6 differences was determined over the whole dataset, with mean and median values of 85.6 and 90, respectively, for a total 148 differences to be identified. The number of observed differences ranged from 31 to 122. The number of differences identified varied to the point that one operator would conclude that two spectra were almost identical while another would consider them significantly different![26]

The problem is substantially more complicated when comparing not only spectra but series of spectra corresponding, for example, to samples taken upstream and downstream of a treatment plant. A big variation in an ion count between upstream and downstream spectra may not be significant if the magnitude of variation is equal within and between the downstream and upstream populations; changes in the abundance of this ion reflects only the intrinsic chemical variability of the environment and is not a relevant marker of the impact of the plant. On the other hand, a slight change in the abundance of an ion between "upstream and downstream spectra" may be significant if abundance is almost constant within each population; it then reflects a real effect of the plant on water quality. The SPIX software was created to remedy the observed fact that it is impossible for an operator to determine what makes sense based on simple observation of complex datasets, especially since the data are subject to intrinsic variability. The aim is to extract relevant information from numerous complex data. As explained below, the software can identify significant differences between mass spectra series and track the kinetic evolution of reagents, unknown reaction intermediates, and reaction products at low concentrations in complex mixtures.

3. Materials and methods

3.1. The SPIX software

SPIX was developed in MATLAB 2018a. A stand-alone version is freely available on the website (http://spix.webpopix.org). The source code can be made available on request. Prior to performing any statistical analysis of the data, pre-processing is required to identify and align significant peaks in the data. The method used for detection and alignment actually depends on the type of data available:

- When the device provides data in xml format, this data has already been filtered and contains only the most significant peaks. These peaks are then aligned by using the mspalign function of the Bioinformatics Toolbox (MATLAB) with the "shortest-path" option.
- When the data obtained are raw data (e.g. xy Bruker format in the present study), i.e. intensities measured on a fine and regular grid, the following algorithm is used: considering K series to analyze, the approximate positions of the significant peaks are first roughly determined by building a single series, consisting of the maximum intensities of the K series at all data points, and by thresholding this series. This procedure is used to determine disjoint segments in which the peaks of each of the K series are located.

The position and intensity of each of these peaks are then estimated for each spectrum by fitting a model of the form $A \exp(-\alpha(x-m))$ for which the maximum value A is reached for $x=m$.

SPIX can be used in essentially two situations. The first one allows evidencing some modifications in the composition of a complex mixture over time. The focus here is on how the abundance of certain species varies as a function of a given parameter (time, pH, reagent, etc.). The objective is twofold: to detect ions with significantly varying abundance (in terms of statistical relevance), and to describe how the abundance varies by kinetic modeling. After aligning the peaks as previously described, different kinetic models, including various patterns associated with compound degradation and formation and reaction intermediates, are fitted to the data. A library including seven typical kinetics profiles is currently available; examples of graphical representations are provided in supplementary information SI-2. The selected model minimizes the Bayesian Information Criterion (BIC). The coefficient of determination r^2 is calculated to quantify the part of the variability of the data explained by the model and an ANOVA assesses whether this part of explained variability is statistically significant. The p-value of the F-test and the r^2 value are represented graphically so as to easily visualize ions with abundance accurately fitting a kinetic model.

SPIX also permits to compare 2 series of samples collected under 2 experimental conditions. The objective is to identify the ions with significant differences in intensity and to quantify these differences. The algorithm first consists in identifying the peaks considered significant: i.e., present and above a given threshold in at least 1 of the 2 conditions. For an ion detected in this way, the procedure is as follows: first, the series are locally shifted so that all the peaks are aligned. The maximum intensity at the peak is estimated for each spectrum by fitting a model of the form $A \exp(-\alpha(x-m))$ for which the maximum value A is reached when $x=m$. This provides 2 series of values that can be compared on statistical tests. A t-test detects differences in the mean while a non-parametric Wilcoxon test more generally detects whether the peak intensity tends to be higher in one condition than in the other. A graphical representation of the p-values obtained for all the peaks detected, as well as of the size effects (i.e., differences in mean values between the 2 conditions) provides quick visualization of the chemically significant differences and the statistical relevance of the differences.

Blank correction can be done as follows: The user chooses as a threshold, a ratio and a percentile. By default, the median of intensities is used for the calculations ($p = 0.5$). For the given percentile, the ratio is defined as:

$$r_p(m/z) = \frac{B_p(m/z)}{S_p(m/z)}$$

with $B_p(m/z)$: percentile of order p of the blank intensities' maximum and $S_p(m/z)$: percentile of order p of the experimental data intensities' maximum. If the peak intensity is higher than $S_p(m/z)$ (as a threshold value) in at least one of the experiment spectra, it will be kept as a peak, if not it will be ignored.

With .mat or .xml files, SPIX occupies about 500 MB (it's the MATLAB runtime that takes up all the space). With .xy, formats the .mat conversion stage has to be added (sequentially): if a sub-repertory (time_0 for example) is 250 MB, then SPIX occupies about 750 MB of memory. It does not represent the total volume of all sub-repertories because SPIX loads them and converts them one by one. In all cases, it works very well on a standard PC.

3.2. Chemicals, reagents, irradiation processes and analysis

The ability of SPIX to extract relevant information from sets of complex high-resolution mass spectra is illustrated in two experiments. The first concerned peroxide photocatalyzed degradation of Maprotiline (an antidepressant drug) in a wastewater treatment pilot plant. In this case, the comparison aimed at revealing reagents, intermediates and products using kinetic models, from mass spectra recorded at different irradiation times. The second experiment concerned UV irradiation of Acetamiprid (a neonicotinoid insecticide) in a complex mixture of aqueous fulvic acid to simulate river water; it aimed at revealing Acetamiprid photoproducts at trace levels and evaluating the impact of UV treatment on dissolved organic matter. The comparison covers 2 data sets, for spectra recorded before and after irradiation. Maprotiline and Acetamiprid chemical structures are depicted Figure 1. Supplementary information file SI-3 describes the chemicals, sampling and irradiation processes used for the two experiments.

Figure 1

3.3. High-resolution mass spectrometry analysis

An ultra-high-resolution mass spectrometer, FT-ICR SolarixXR 9.4T (Bruker Daltonics, Bremen, Germany), was used for direct infusion mass spectrometry analysis. The electrospray ion source was set in positive mode and solutions were injected using an automated Acquity HPLC system (Waters, Saint Quentin en Yvelines, France). The injection volume was 10 µL. Elution was carried out using a 0.002 mL/min flow of H_2O/ACN/FA (50/50/0.1). Nitrogen was used as nebulizer and as drying gas, set at 1 bar and 4 L/min, respectively. The drying gas temperature was set at 180°C. The capillary voltage and endplate offset potential were set at -4500 V and -500 V, respectively. Ions were accumulated for 0.2 sec in the collision cell, and 50 scans were summed. Resolution was set at 4 Mpt on a scan range from m/z 57 to m/z 1,000 in order to obtain resolution > 400,000 at m/z 200. A tune mix (Agilent Technologies, Les Ulis, France) was used for mass calibration. Exact formulae were assigned with error < 1 ppm.

4. Results and discussion

4.1. Highlighting chemical reactions in complex mixtures

Degradation of Maprotiline in wastewater under an advanced oxidation process (peroxide/UV) was carried out in a pilot plant, with the aim of testing the ability of SPIX to follow the degradation of contaminants and the evolution of their transformation products. This pilot plan was set up by FACSA, a Spanish company operating water treatment plants, to design, optimize and compare novel water treatment processes; the operational parameters and analytical conditions are given in the Supplementary information SI-3. Considering that the abundances of reagents, intermediates and products of a chemical reaction are not expected to evolve stochastically, an original way to extract relevant information from untargeted analysis consists in filtering the results based on ion abundance trends. Briefly, A-type models are selected to detect molecules with decreasing abundance during the reaction while B-type and C-type models detect the products and intermediates, respectively. The software detects all significant changes over time and provides a kinetic model; statistical data can be exported for further analysis in the Table format described in Supplementary information SI-4. As a first example, from one set of samples using the software default threshold (1.2E+08), SPIX automatically extracted the m/z 278.19056 signal (protonated Maprotiline) for each irradiation time, and associated the fitting model referred to as A1 in SPIX with $r^2 > 0.99$ (Figure 2).

Figure 2

A first attempt to extract transformation products from the background noise, conducted on the basis of 1 set of measurements, gave a few Maprotiline-related peaks, but a thorough study of the raw data showed that using data from duplicate measurements yielded a kinetic model better fitting ion intensity evolution. The aim was to maximize the relevant data obtained while minimizing the number of parallel measurements, to gain valuable analysis time. Using 2 parallel measurements for each irradiation time, 88 ions fitting one of the SPIX kinetic models were extracted with default threshold intensity 1.2E+08; this list was reduced to 12 ions keeping m/z ratios fitting a kinetic model with $r^2 > 0.9$ (Table 1), these fittings being also those corresponding to the lowest p-values. The formulas were assigned using the Bruker software based on accurate mass measurements (sub-ppm accuracy) and isotopic pattern-matching. All the extracted m/z values were related to Maprotiline or its photoproducts (oxidized compounds); they corresponded to singly charged ions with ^{12}C and ^{13}C isotopic contributions. One signal (m/z 92.73055) corresponded to an artifact related to the harmonics of the m/z 278.19054 signal, resulting from signal digitization and Fourier transformation, a phenomenon previously described and explained by Mathur and O'Connor.[27] To study the threshold effect and determine whether additional photoproducts would be found if more peaks were considered, the threshold was halved (6E07) and the same methodology was applied. 197 peaks were thus selected by SPIX and the data were ordered according to statistical relevance. 23 peaks were then selected on the criterion $r^2 > 0.9$ (see Table 1). Here again, m/z values were all 1related to Maprotiline and its photoproducts; a second artifact (m/z 93.06499) was found and attributed to the harmonics of m/z 279.19386.

Table 1. Ions extracted and associated kinetic models related to the photodegradation of Maprotiline in wastewater with $r^2 > 0.9$ (data ordered by decreasing intensity)

m/z	r^2	p-value	Model [a]	Maprotiline-related	Ion formula	Intensity	Relative Intensity (%)
With intensity > 1.2E08							
278.19054	0.93	5.25E-06	A1	Yes	$C_{20}H_{24}N^+$	8350669911	100.0
279.19386	0.94	3.82E-06	A1	Yes	$C_{19}H_{24}N^{\wedge 13}C^+$	1643233281	19.7
294.18552	0.98	3.91E-07	C1	Yes	$C_{20}H_{24}NO^+$	1126960171	13.5
276.17486	0.92	7.58E-05	C1	Yes	$C_{20}H_{22}N^+$	451170002	5.4

292.16992	0.94	2.73E-05	B2	Yes	$C_{20}H_{22}NO^+$	352390361	4.2
92.73055	0.92	1.48E-05	A1	Yes	[b]	339591819	4.1
310.18038	0.94	4.09E-06	B1	Yes	$C_{20}H_{24}NO_2^+$	228987461	2.7
295.18881	0.97	2.30E-06	C1	Yes	$C_{19}H_{24}NO^{\wedge 13}C^+$	218499045	2.6
280.19727	0.94	3.44E-06	A1	Yes	$C_{18}H_{24}N^{\wedge 13}C_2^+$	157067516	1.9
308.16472	0.91	1.69E-04	B2	Yes	$C_{20}H_{22}NO_2^+$	140798543	1.7
344.18594	0.95	1.52E-05	B2	Yes	$C_{20}H_{26}NO_4^+$	135672030	1.6
342.17029	0.95	1.50E-05	B2	Yes	$C_{20}H_{24}NO_4^+$	112723518	1.3
With 1.2E08 > intensity > 6E07							
360.18088	0.96	8.69E-06	B2	Yes	$C_{20}H_{26}NO_5^+$	95678296	1.1
312.19604	0.94	2.71E-05	C1	Yes	$C_{20}H_{26}NO_2^+$	84376971	1.0
328.19096	0.94	2.74E-05	C1	Yes	$C_{20}H_{26}NO_3^+$	79243551	0.9
318.17025	0.94	2.46E-05	B2	Yes	$C_{18}H_{24}NO_4^+$	75240290	0.9
302.1753	0.98	6.44E-07	C1	Yes	$C_{18}H_{24}NO_3^+$	66900122	0.8
93.06499	0.92	9.63E-06	A1	Yes	[b]	66023322	0.8
242.15416	0.95	1.34E-05	B2	Yes	$C_{16}H_{20}NO^+$	63982463	0.8
326.17543	0.92	8.04E-05	C1	Yes	$C_{20}H_{24}NO_3^+$	50496456	0.6
300.17259	0.95	1.27E-06	A1	Yes	$C_{20}H_{23}NNa^+$	20928662	0.2
139.09569	0.91	9.43E-04	C2	Yes	$[C_{20}H_{24}N]^{2+}$	7061477	0.1
336.14927	0.91	1.76E-05	A1	No	$C_{23}H_{18}N_3^+$	104081459	1.2

[a] Currently available SPIX kinetic models are given in Supplementary information SI-2.

[b] m/z 93.06499 and m/z 92.73055 signals correspond to artifact peaks related to the harmonics of m/z 278.19054 and m/z 279.19386 ions, respectively; they resulted from signal digitization.[27]

Using the selection parameters referred to above, one protonated species was detected at m/z 336.14927. Considering Kind & Fiehn's "7 golden rules" and selecting atoms C, H, N, O, P, S, F, Cl, Br, Si, Na and K, the only matching formula was $C_{23}H_{18}N_3^+$.[28] This species is logically assumed not to be related to Maprotiline; it could correspond to a contaminant in high concentration in waste water, degrading under UV radiation. According to the kinetics revealed by SPIX, some photoproducts were present in detectable abundance after 2.5 minutes' irradiation. To estimate the relevance of the results provided by SPIX, 1 of the spectra recorded at this reaction time was selected. After blank subtraction (the blank consisting of wastewater matrix without Maprotiline), the spectrum was exported in .csv format (Bruker's FTICR-MS file format). The spectra were recorded using 8 Mpts, and as the experiments were carried out using secondary treated wastewater, 4,479 peaks were exported by Bruker software which were above the S/N ratio threshold of 4. Out

of these 4,479 peaks, the 11 most abundant ions in the selected spectrum corresponded to the 11 most statistically relevantly changing m/z values extracted by SPIX. The two m/z signals corresponding to harmonics of major ion signals were also extracted with good fit, and ranked 18th and 115th in the original file (overall blank subtracted-spectrum). The photoproducts showing the highest significance (lowest p-values) were those corresponding to m/z 294.18552 and m/z 302.17530 (protonated molecules), fitting is presented in Figure 3 for the former. One of the Maprotiline-related peaks that was not in the list of the highest intensities was not found in the spectrum recorded at 2.5 min of irradiation: it was removed by the blank subtraction process, since the wastewater matrix was very complex. It is thus noteworthy that SPIX does not require blank subtraction to provide valuable information, allowing relevant peaks that coincidentally overlap with some of the matrix peaks not to be removed. The experiment conducted on photodegradation of Maprotiline showed that the SPIX software efficiently revealed the most relevant changes in the composition of the irradiated mixture on the basis of only 12 mass spectra. It was able to automatically detect reagents, intermediates and products at trace levels. Most extracted ions were related either to Maprotiline or to its photoproducts. Only 1 compound was found which was assumed not to be related to Maprotiline on the basis of its molecular formula ($C_{23}H_{17}N_3$); no significant change in the composition of the dissolved organic matter was found, although of course only ESI-ionized species were considered. The photodegradation pathways of Maprotiline have been reported in a study more oriented toward structural elucidation.[29]

Figure 3

4.2. Comparison of two conditions: the example of photodegradation of Acetamiprid in an aqueous solution of fulvic acid

This example demonstrates the ability of the SPIX software to point out relevant changes between two conditions from changes in low-abundance ions within a complex matrix. We studied the effect of UV radiation on Acetamiprid in a complex mixture, simulating river water. A prior photolysis study of Acetamiprid in ultrapure water identified Acetamiprid degradation products in ultrapure water and demonstrated that the presence of other substances in the matrix leads to the formation of different degradation products.[30] These results led us to study the effect of dissolved organic matter on the photodegradation of Acetamiprid as it may happen under real environmental conditions. The peak intensity of Acetamiprid represented 9.2% of the base peak of the mass spectrum before

photolysis and only 1.9% after 30 minutes' irradiation. A sodium adduct, originating from the use of glassware during sample preparation, impurity in solvents or ESI needle for instance, was detected with a relative intensity of 10.4% in the spectrum recorded before and 2.4% after photolysis. These differences, not detectable looking at the whole spectrum, are obvious when zooming on the region from m/z 223.00 to 223.20 (Figure 4).

Figure 4

Sets of spectra recorded before and after photolysis were compared using the SPIX software. Given the intensity of Acetamiprid within the mixture, a peak detection limit was set at only three times the average intensity of spectral noise (average noise at 1E6, detection threshold set at 3E6). This threshold was set as low as possible so as to identify Acetamiprid degradation products in small amounts. Blank spectra were subtracted to eliminate any interference from solvents or instruments. Exported from SPIX, Table 2 lists the ions the abundance of which underwent significant change after irradiation. Here, only ions with a probability of $\geq 95\%$ (p-value ≤ 0.05) were retained: i.e., with significant difference in intensity between the two conditions. In this example, results are organized by increasing p-value, but any parameter can be chosen for presentation of the results. A negative value in the "Difference" column indicates that ion abundance increased after photolysis. Visual comparison (Figure 5) confirms the results displayed in Table 2: the greatest change in intensity - and with the highest significance – was the concomitant decrease of the m/z 223 (MH^+) and m/z 245 (MNa^+) ions. Many other peaks decreased or increased after photolysis, but with lower p-values for the difference in intensity; some were related to Acetamiprid photodegradation, others to changes in dissolved organic matter. The ions at m/z 205.10847 ($C_{10}H_{13}N_4O^+$) and m/z 227.09036 ($C_{10}H_{12}N_4NaO^+$) shown here correspond to the protonated and cationized forms of a photoproduct previously described, the structure of which was elucidated in a study on UV irradiation of Acetamiprid in pure water.[30] This photoproduct was the major one, so it is not possible to say whether others which were previously described were not detected here due to too high detection thresholds or because they are not formed in the presence of fulvic acid. It is interesting to note that SPIX revealed two ions, 284.12725 ($C_{12}H_{19}ClN_5O^+$) and m/z 245.10095 ($C_{10}H_{14}N_4NaO_2^+$) - resulting from ionization of Acetamiprid photoproducts based on their formula - that were not detected by LC-MS in pure water. With a relative intensity below 1% in infusion mode, these two molecules could not have been revealed without SPIX. Fifteen ions with abundance significantly varying before and after UV irradiation were indicated. Based on their chemical formulae, they were assumed not to be related to Acetamiprid or to its photoproducts; they all included a large number

of oxygen atoms (≥ 6) and likely resulted from oxidation of dissolved organic matter. This is of great interest because it opens a way to investigate the global consequences of a depollution treatment, evaluating the treatment - apart from its ability to efficiently degrade pollutants - in terms of biotope preservation.

Figure 5

Table 2. m/z for which intensity significantly varied between series of spectra recorded before and after 30 minutes of irradiation (n = 6). Ions are listed by increasing p-value.

m/z	Difference in intensity	p-value	Ion formula	Related to Acetamiprid [a]
224.07792	10523197	1.47E-06	$C_9H_{12}ClN_4{}^{\wedge 13}C^+$	Yes
223.07459	103518217	3.20E-06	$C_{10}H_{12}ClN_4{}^+$	Yes
225.07161	24820606	1.08E-05	$C_{10}H_{12}N_4{}^{\wedge 37}Cl^+$	Yes
246.05988	10966361	1.30E-05	$C_9H_{11}ClN_4Na^{\wedge 13}C^+$	Yes
247.05358	30754403	2.86E-05	$C_{10}H_{11}N_4Na^{\wedge 37}Cl^+$	Yes
245.05653	107329585	4.50E-05	$C_{10}H_{11}ClN_4Na^+$	Yes
201.03705	-3542730	1.84E-02	$C_6H_{10}NaO_6{}^+$	No
205.10847	-28497851	2.11E-02	$C_{10}H_{13}N_4O^+$	Yes
251.05269	-7062764	2.71E-02	$C_{12}H_{11}O_6{}^+$	No
297.05819	-5696496	2.72E-02	$C_{11}H_{14}NaO_8{}^+$	No
267.04760	-5586613	2.99E-02	$C_{10}H_{12}NaO_7{}^+$	No
283.04253	-4479881	3.20E-02	$C_{10}H_{12}NaO_8{}^+$	No
237.03704	-4871105	3.25E-02	$C_9H_{10}NaO_6{}^+$	No
253.03194	-4495727	3.33E-02	$C_9H_{10}NaO_7{}^+$	No
245.10095	-14590229	3.65E-02	$C_{10}H_{14}N_4NaO_2{}^+$	Yes
211.02138	-5849796	3.65E-02	$C_7H_8NaO_6{}^+$	No
227.09036	-38558966	3.70E-02	$C_{10}H_{12}N_4NaO^+$	Yes
239.05270	-7601354	3.77E-02	$C_9H_{12}NaO_6{}^+$	No
213.03705	-8698910	4.00E-02	$C_7H_{10}NaO_6{}^+$	No
275.05267	-3822758	4.02E-02	$C_{12}H_{12}NaO_6{}^+$	No
284.12725	12781460	4.41E-02	$C_{12}H_{19}ClN_5O^+$	Yes
253.06834	-7132886	4.57E-02	$C_{10}H_{14}NaO_6{}^+$	No
255.08400	-3840684	4.60E-02	$C_{10}H_{16}NaO_6{}^+$	No
325.05308	-3163824	4.66E-02	$C_{12}H_{14}NaO_9{}^+$	No
241.03195	-4410055	4.98E-02	$C_8H_{10}NaO_7{}^+$	No

[a] assumption based on the ion chemical formula

5. Conclusion

The SPIX software aims at extracting relevant data from mass spectra data sets. User-friendly and totally free, it is available for all at: http://spix.webpopix.org. Two features of SPIX are presented in this article, based on examples taken from the field of environmental analysis. The first example showed how SPIX revealed photodegradation reactions by correlating significant changes in ion abundance over time with kinetic models. Thus, the software revealed the reagents, products and intermediate species in a very complex mixture (wastewater). This functionality can be extended to monitor any kind of reaction - even unknown - in all types of mixture. Some features of the SPIX software are still under development. One aims at extending compatibility with many more file formats (wiff, .d, .pkl, .qgd, etc.), in order to be used with most of the marketed mass spectrometers. Another focus of development – already running but requiring some improvements – concerns the extension of SPIX to three-dimensional datasets from hyphenated techniques such as GC-MS, LC-MS or IM-MS. Some commercial software applications allow comparison of chromatograms but, to our knowledge, an approach consisting in extracting relevant data from hyphenated techniques based on fitting with kinetic models has never been reported. The second example showed the ability of SPIX to detect photoproducts at trace amounts in an aqueous solution containing dissolved organic matter, through comparison of datasets for 2 conditions (before/after photolysis in the present case). Regarding this example, the ability of SPIX to deal with intrinsic variability and reduce operator subjectivity opens up promising prospects in all areas of analytical chemistry. It could, in particular, be a very useful tool to assess fragrance and flavor counterfeits, where expertise is very challenging due to the normal variations in abundance of natural substances featuring at low levels in their composition. This feature can also be used to estimate the global consequences of a given treatment on the treated medium, for example to monitor the oxidation of dissolved organic matter and its consequences on biotope preservation.

Acknowledgements

Financial support from the National FT-ICR network (FR 3624 CNRS) and from Inria (Institut national de recherche en informatique et en automatique) for conducting the research is gratefully acknowledged. This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No 765860 (AQUAlity).

Figure captions

Figure 1. Chemical structures of Maprotiline (left hand) and Acetamiprid (right hand)

Figure 2. m/z 278.19056 signal (protonated Maprotiline) extracted for each irradiation time and fitting model associated to degradation kinetics: 8.19E-5+4.37E+10*exp(-0.23*x); r^2 = 0.991

Figure 3. 294.18552 signal extracted for each irradiation time and associated kinetic model

Figure 4. Mass spectra of Acetamiprid in mixture with fulvic acid. a) Mass spectrum before photolysis; b) Mass spectrum after 30 min photolysis; c) Zoom on the protonated Acetamiprid peak (m/z 223.0746) in spectrum a; d) Zoom on the protonated Acetamiprid peak in spectrum b

Figure 5. Visual result provided by the SPIX software after processing of mass spectra series recorded from samples taken before and after 30 min photolysis. The differences in ion intensities are given on the y-axis, positive values corresponding to decreased intensity after irradiation. The associated p-value is given by the color scale: the more the color tends toward red, the more statistically significant the difference.

Associated Content

SI-1. State of the art of modern approaches in managing high-resolution mass spectrometry data

SI-2. Current kinetic models in SPIX

SI-3. Chemicals, reagents and sample preparation

SI-4. Exported data from the SPIX software after assignation of a kinetic model

[1] Hollender J, van Bavel B, Dulio V, Farmen E, Furtmann K, Koschorreck J, Kunkel U, Krauss M, Munthe J, Schlabach M, Slobodnik J, Stroomberg G, Ternes T, Thomaidis NS, Togola A, Tornero V. High resolution mass spectrometry-based non-target screening can support regulatory environmental monitoring and chemicals management. Environ Sci Eur. 2019;31(1):62-67.

[2] Giorio C, Bortolini C, Kourtchev I, Tapparo A, Bogialli S, Kalberer M. Direct target and non-target analysis of urban aerosol sample extracts using atmospheric pressure photoionisation high-resolution mass spectrometry. *Chemosphere*. 2019;224:786-795.

[3] Hoh E, Dodder NG, Lehotay SJ, Pangallo KC, Reddy CM, Maruya KA. Nontargeted comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry method and software for inventorying persistent and bioaccumulative contaminants in marine environments. *Environ Sci Technol*. 2012;46:8001-8008.

[4] Guo J, Chen D, Potter D, Rockne KJ, Sturchio NC, Giesy JP, Li A. Polyhalogenated carbazoles in sediments of Lake Michigan: A new discovery. *Environ Sci Technol*. 2014;48:12807-12815.

[5] Schymanski EL, Singer HP, Longrée P, Loos M, Ruff M, Stravs MA, Ripollés Vidal C, Hollender J. Strategies to characterize polar organic contamination in wastewater: Exploring the capability of high-resolution mass spectrometry. *Environ Sci Technol*. 2014;48:1811-1818.

[6] Gago-Ferrero P, Schymanski EL, Bletsou AA, Aalizadeh R, Hollender J, Thomaidis NS. Extended suspect and non-target strategies to characterize emerging polar organic contaminants in raw wastewater with LC-HRMS/MS. *Environ Sci Technol*. 2015;49:12333-12341.

[7] Kinani A, Kinani S, S. Bouchonnet S. Formation and determination of organohalogen by-products in water. Part III. Characterization and quantitative approaches. *TRAC-Trend Anal Chem*. 2016;85:295-305.

[8] Hollender J, van Bavel B, Dulio V, Farmen E, Furtmann K, Koschorreck J, Kunkel U, Krauss M, Munthe J, Schlabach M, Slobodnik J, Stroomberg G, Ternes T, Thomaidis NS, Togola A, Tornero V. High resolution mass spectrometry-based non-target screening can support regulatory environmental monitoring and chemicals management. *Environ Sci Eur*. 2019;31:2-52.

[9] Schymanski EL, Singer HP, Slobodnik J, Ipolyi IM, Oswald P, Krauss M, Schulze T, Haglund P, Letzel T, Grosse S, Thomaidis NS, Bletsou A, Zwiener C, Ibáñez M, Portolés T, De Boer R, Reid MJ, Onghena M,

Kunkel U, Schulz W, Guillon A, Noyon N, Leroy G, Bados P, Bogialli S, Stipanicev D, Rostkowski P, Hollender J. Non-Target Screening with High-Resolution Mass Spectrometry: Critical Review Using a Collaborative Trial on Water Analysis. *Anal Bioanal Chem.* 2015, 407, 6237-6255.

[10] https://www.ineris.fr/sites/ineris.fr/files/contribution/Documents/AQUAREF_Togola_NTS.pdf

[11] Kendrick E. A mass scale based on CH_2 = 14.0000 for high resolution mass spectrometry of organic compounds. *Anal Chem.* 1963;35(13):2146-2154.

[12] Sleno L. The use of mass defect in modern mass spectrometry. *J Mass. Spectrom*. 2012;47:226-236.

[13] Van Krevelen DW. Graphical-statistical method for the study of structure and reaction processes of coal. *Fuel* 1950:29;269-84.

[14] Kew W, Blackburn JWT, Clarke DJ, Uhrín D. Inter-active van Krevelen diagrams-advanced visualisation of mass spectrometry data of complex mixtures. *Rapid Commun Mass Spectrom.* 2017;31:658-662.

[15] Wu Z, Rodgers RP, Marshall AG. Two- and three-dimensional van Krevelen Diagrams: A graphical analysis complementary to the Kendrick mass plot for sorting elemental compositions of complex organic mixtures based on ultrahigh-resolution broadband Fourier Transform ion cyclotron resonance mass measurements. *Anal Chem.* 2004;76:2511-2516.

[16] Massart DL, Vandeginste BGM, Deming SM, Michotte Y, Kaufman L. Regression Methods in Chemometrics: a textbook. 1988, pp. 165-182.

[17] Jurs PC. Pattern recognition used to investigate multivariate data in analytical chemistry, *Science* 1986;232:1219-1224.

[18] Abdi H, Williams LJ. Principal component analysis. *WIREs Comput Stat*. 2010;2:433-459.

[19] Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification y. *J Chemom.* 2006;20:341-351.

[20] Araujo F, Peres P, Fogliatto FS. Variable selection methods in multivariate statistical process control: A systematic literature review. *Comput Ind Eng*. 2017;115:603-619.

[21] Zou H, Hastie T, Tibshirani R, Ou HZ, Astie TH, Ibshirani RT. Sparse principal component analysis. *J Comput Graph Stat.* 2012;8600:265-286.

[22] Filzmoser P, Gschwandtner M, Todorov V. Review of sparse methods in regression and classification with application to chemometrics. *J Chemom*. 2012;26:42-51.

[23] Wu H, Xue R, Tang Z, Deng C, Liu T, Zeng H, Sun Y, Shen X. Metabolomic investigation of gastric cancer tissue using gas chromatography/mass spectrometry. *Anal Bioanal Chem*. 2010;396:1385-1395.

[24] Thevenot EA, Roux A, Xu Y, Ezan E, Junot C. Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *J Proteome Res*. 2015;14(8):3322-3335.

[25] Kirwan JA, Weber RJM, Broadhurst DI, Viant MR. Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control. *Sci data*. 2014;1:140012.

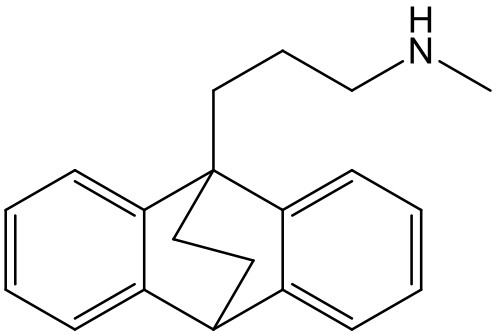
[26] Nicol E, Xu Y, Varga Z, Grosshans R, Lavielle M, Bouchonnet S. Variability and subjectivity in analytical chemistry. Results presented at the International Winter School on Mass Spectrometry. March 2019, Palaiseau, France.

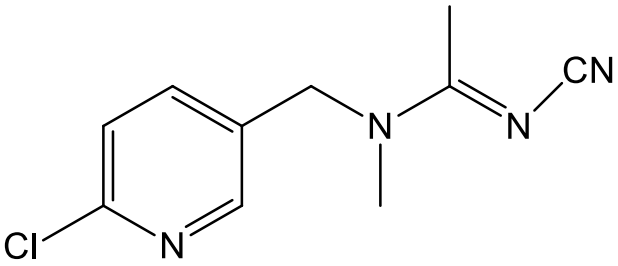
[27] Mathur R, O'Connor PB. Artifacts in Fourier transform mass spectrometry. *Rapid Commun Mass Spectrom*. 2009;23: 523–529.

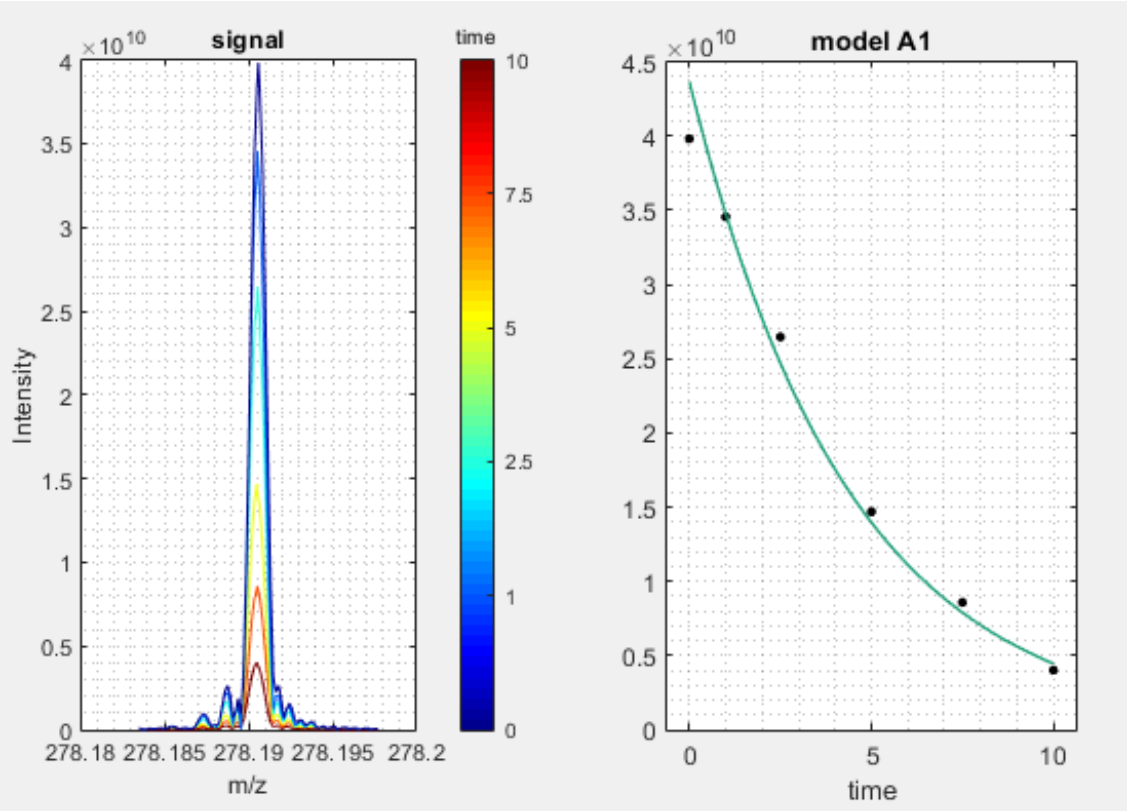
[28] Kind T, Fiehn O. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinform.* 2007;8:105.

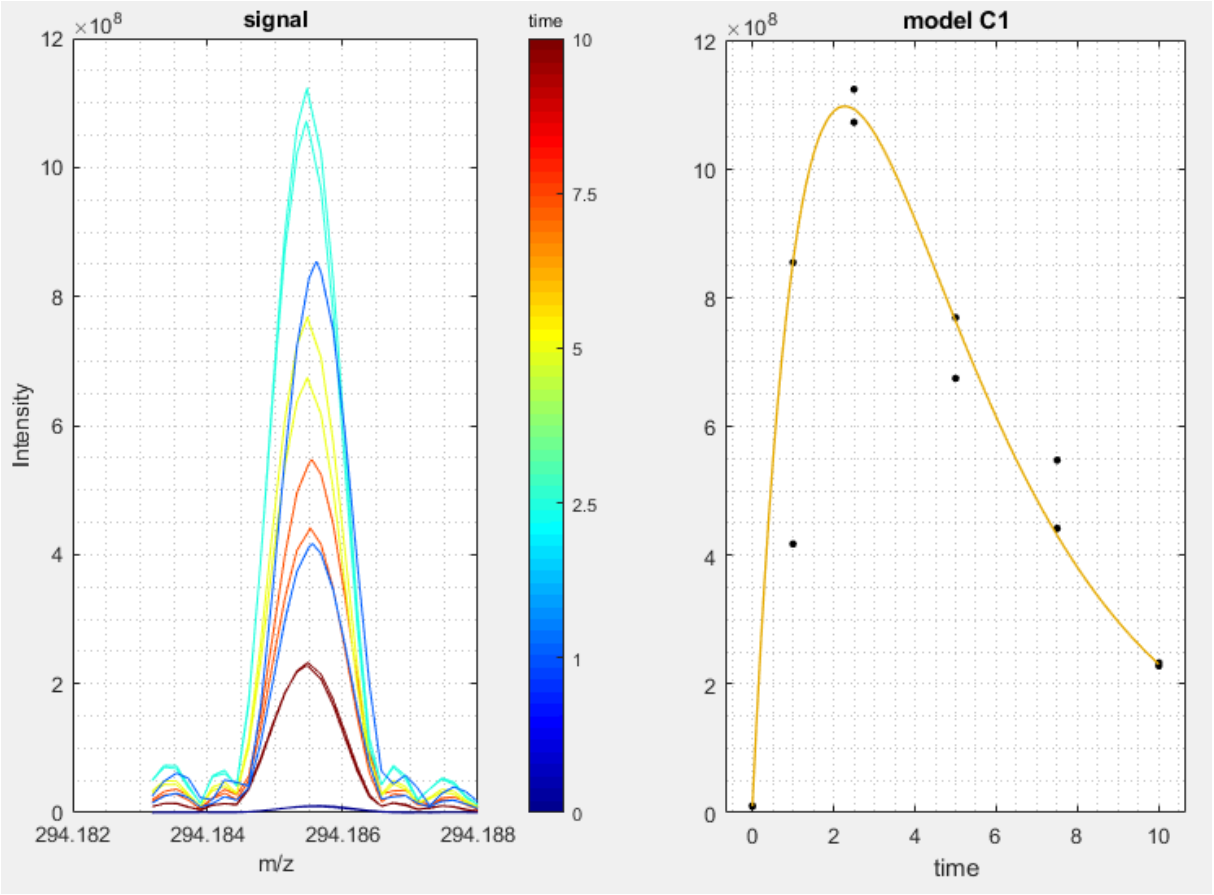
[29] Gonçalves NPF, Varga Z, Bouchonnet S, Dulio V, Alygizakis N, Dal Bello F, Medana C, Calza P. Study of the photoinduced transformations of maprotiline in river water using liquid chromatography high-resolution mass spectrometry. *Sci. Total Environ.* 2020;6:143556.

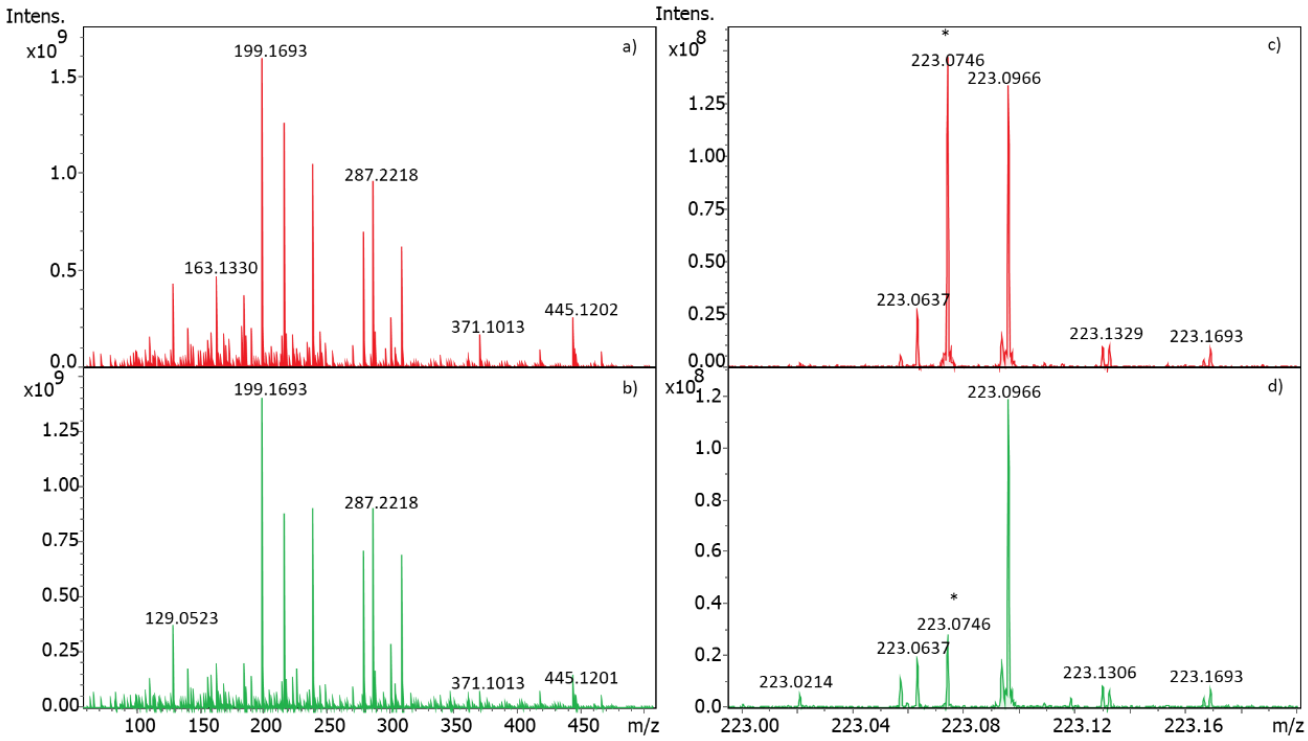
[30] Nicol E, Varga Z, Vujovic S, Bouchonnet S. Laboratory scale UV-visible degradation of Acetamiprid in aqueous marketed mixtures - Structural elucidation of photoproducts and toxicological consequences. *Chemosphere* 2020;248:126040.

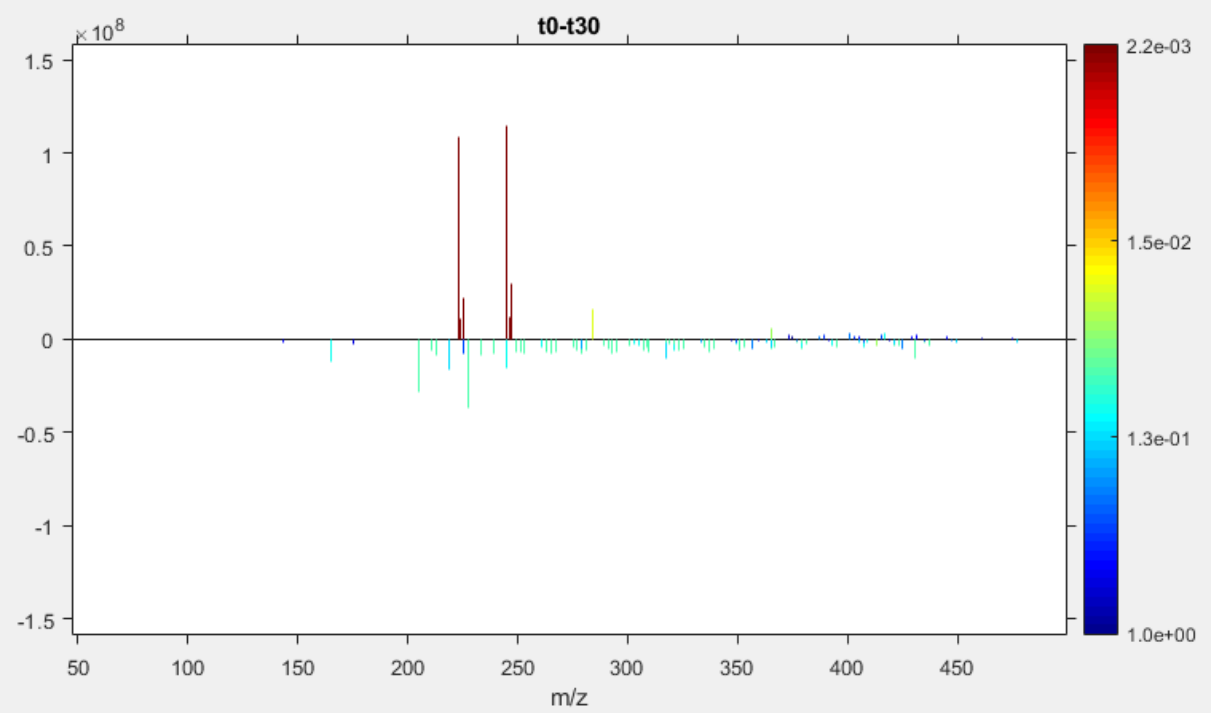

H
N


CN
N
N
Cl
N


signal
×10^10
Intensity
m/z
278.18
278.185
278.19
278.195
278.2
time
model A1
time


signal
×10^8
Intensity
m/z
294.182
294.184
294.186
294.188
time
model C1
time


Intens.
x10^9
199.1693
a)
287.2218
163.1330
371.1013
445.1202
b)
129.0523
445.1201
m/z
Intens.
x10^8
*
223.0746
223.0966
c)
223.0637
223.1329
223.1693
d)
223.0214
223.1306
223.00
223.04
223.08
223.12
223.16


t0-t30
×10^8
1.5
1
0.5
0
-0.5
-1
-1.5
50
100
150
200
250
300
350
400
450
m/z
2.2e-03
1.5e-02
1.3e-01
1.0e+00

Supplementary information

SPIX: a new software package to reveal chemical reactions at trace amounts in very complex mixtures from high-resolution mass spectra data sets

Edith Nicol, Yao Xu, Zsuzsanna Varga, Said Kinani, Stéphane Bouchonnet, Marc Lavielle

SI-1. State of the art of modern approaches in managing high-resolution mass spectrometry data

In mass spectrometry, the emergence of high-resolution analyzers has enabled analysis of samples of ever-increasing complexity. Whether in direct infusion or with hyphenated techniques (LC-MS and GC-MS couplings), the amount of information issuing from high-resolution analysis of complex mixtures requires computer processing and simplified data representation. Direct infusion of a sample can thus provide a mass spectrum including several thousands of distinct ions. Various representations are commonly used to simplify the visualization and comparison of samples analyzed by mass spectrometry. As the approaches are so diverse, this section does not seek to be exhaustive, but restricts itself to presenting the Kendrick and Van Krevelen diagrams and the multivariate statistical analyses most commonly used by high-resolution mass spectrometry specialists.

Kendrick diagrams

The Kendrick diagram allows easy identification, from a mass spectrum, of series of compounds that include the same number of heteroatoms and unsaturations but differ from each other by the number of -CH_2- groups.[1] The diagram is built by plotting the Kendrick mass defect (KMD) for each ion (eq. 1) as a function of the Kendrick mass (KM) (eq. 2).

$$KMD = (Nominal\ Kendrick\ Mass - Exact\ Kendrick\ Mass) \quad \text{(eq. 1)}$$

$$KM = IUPAC\ mass * \left(\frac{14}{14.01565}\right) \quad \text{(eq. 2)}$$

Compounds in the same series (i.e., with the same number of heteroatoms and degrees of unsaturation) will have the same KMD. In the diagram, each series is aligned horizontally with a deviation of 14 that reflects a difference of one -CH_2- pattern. A shift of 0.01340 on the vertical axis corresponds to implementation of 1 unsaturation. Originally reported for the investigation of

[1] Kendrick, E. A mass scale based on CH_2 = 14.0000 for high resolution mass spectrometry of organic compounds. *Anal. Chem.* **1963**, *35(13)*, 2146-2154

petroleomics-type samples in the early 2000s,[2,3] the use of the Kendrick diagram has been extended and adapted over the years for complex environmental samples,[4,5,6] metabolomic studies,[7,8] and proteomics on phosphopeptides.[9] As needs differ between environmental chemistry and petroleomics, many studies have focused on modification of the mass defect[10] in order to characterize reaction products such as oxidation or chlorination.[11,12] Even with a mass measurement accuracy of 1 ppm, a

[2] Marshal, A.G.; Rodgers, R.P. Petroleomics: the next grand challenge for chemical analysis. *Acc. Chem. Res.* **2004**, *37*, 53-59

[3] Hughey, C.A.; Hendrickson, C.L.; Rodgers, R.P.; Marshall, A.G. Kendrick mass defect spectrum: a compact visual analysis for ultrahigh-resolution broadband mass spectra. *Anal. Chem.* **2001**, *73*, 4676-4681

[4] Chu, F.L.; Pirastru, L.; Popovic, R.; Sleno, L. Carotenogenesis up-regulation in Scenedesmus sp. using a targeted metabolomics approach by liquid chromatography – high-resolution mass spectrometry. *J. Agric. Food Chem*. **2011**, *59*, 3004-3013

[5] Sleighter, R.L.; Hatcher, P.G. The application of electrospray ionization coupled to ultrahigh resolution mass spectrometry for the molecular characterization of natural organic matter. *J. Mass Spectrom*. **2007**, *42*, 559-574

[6] Kramer, R.W.; Kujawinski, E.B.; Hatcher, P.G. Identification of black carbon derived structures in a volcanic ash soil humic acid by Fourier transform ion cyclotron resonance mass spectrometry. *Environ. Sci. Technol*. **2004**, *38*, 3387-3395

[7] Ni, S.; Qian, D.; Duan, J.; Guo, J.; Shang, E.; Shu, Y.; Xue, C. UPLC-QTOF/MS-based screening and identification of the constituents and their metabolites in rat plasma and urine after oral administration of Glechoma longituba extract. *J. Chromatogr. B* **2010**, *878*, 2741-2750

[8] Zhang, H.; Zhang, D.; Ray, K.; Zhu, M. Mass defect filter technique and its applications to drug metabolite identification by high-resolution mass spectrometry. *J. Mass Spectrom*. **2009**, *44*, 999-1016

[9] Bruce, C.; Shifman, M.A.; Miller, P.; Gulcicek, E.E. Probabilistic enrichment of phosphopeptides by their mass defect. *Anal. Chem*. **2006**, *78*, 4374-4382

[10] Sleno, L. The use of mass defect in modern mass spectrometry. *J. Mass. Spectrom*. **2012**, *47*, 226-236

[11] Jobst, K.J.; Shen, L.; Reiner, E.J.; Taguchi, V.Y.; Helm, P.A.; McCrindle, R.; Backus, S. The use of mass defect plots for the identification of (novel)halogenated contaminants in the environment, *Anal. Bioanal. Chem*. **2013**, *405*, 3289-3297

[12] Taguchi, V.Y.; Nieckarz, R.J.; Clement, R.E.; Krolik, S.; Williams, R. Dioxin analysis by gas chromatography-Fourier transform ion cyclotron resonance mass spectrometry (GC-FTICRMS). *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 1918-1921

mass of 200 Da is assigned only 1 raw formula, while a mass of 500 Da is assigned 21.[13] Regarding this issue, the Kendrick diagram can significantly increase the number of single raw formulae that can be assigned from m/z values in a mass spectrum. From the raw formula of the first compound, the identification of homologous series aids in assigning the raw formula of other compounds in the series regardless of their mass. This allows a complex spectrum to be recalibrated, to obtain the best possible accuracy and thus assign as many raw formulae as possible (e.g., prior to a principal component analysis).[14,15]

Van Krevelen diagrams

The van Krevelen diagram, originally used in petroleomics to control oil and kerosene quality, represents the H/C ratio as a function of the O/C or N/C ratio for each ion of a complex mixture.[16] This allows the composition of a sample to be quickly estimated based on constituent molecular families (lipids, proteins, sugars, carbohydrates, lignin, tannins, etc.).[17] Today, this representation is commonly used for environmental samples to track their evolution following an event such as

[13] Kind, T.; Fiehn, O. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinform.* **2006**, 7, 234-243

[14] Ajaero, C.; McMartin, D.W.; Peru, K.M.; Bailey, J.; Haakensen, M.; Friesen, V.; Martz, R.; Hughes, S.A.; Brown, C.; Chen, H.; McKenna, A.M.; Corilo, Y.E.; Headley, J.V. Fourier transform ion cyclotron resonance mass spectrometry characterization of Athabasca oil sand process-affected waters incubated in the presence of wetland plants. *Energy Fuels* **2017**, *31*, 1731-1740

[15] Ajaero, C.; Peru, K.M.; Hughes, S.A.; Chen, H.; McKenna, A.M.; Corilo, Y.E.; McMartin, D.W.; Headley, J.V. Atmospheric pressure photoionization fourier transform ion cyclotron resonance mass spectrometry characterization of oil sand process-affected water in constructed wetland treatment. *Energy Fuels* **2019**, *33*, 4420-4431

[16] van Krevelen, D.W. Graphical-statistical method for the study of structure and reaction processes of coal. *Fuel* **1950**, *29*, 269-284

[17] Kew, W.; Blackburn, J.W.T.; Clarke, D.J.; Uhrín, D. Interactive van Krevelen diagrams – advanced visualisation of mass spectrometry data of complex mixtures. *Rapid Commun. Mass Spectrom.* **2017**, *31*, 658-662

treatment or pollution.[18,19,20,21,22] Some studies have extended the van Krevelen diagram over 3 dimensions to achieve better classification of compounds and better differentiation between complex mixtures. In some cases, this approach is associated with other t-test type statistical tests.[23,24]

Multivariate statistical analysis

Multivariate statistical analysis is a versatile tool for dealing with high-dimensional datasets, and many methods can be used to extract valuable information, perform data compression, assess subclasses and compare groups of samples assessing relationships between variables. For quantitative datasets, two categories of model can be distinguished in terms of the relationship between variables and response, according to the parameters: linear and non-linear. An example is UV-Vis absorbance analysis of a complex mixture, where absorbance depends on the concentrations of all the compounds present in the mixture, based on a linear relationship; in this case, multivariate linear regression can

[18] Minor, E.C.; Swenson, M.M.; Mattson, B.M.; Oyler, A.R. Structural characterization of dissolved organic matter: a review of current techniques for isolation and analysis. *Environ. Sci.-Proc. Imp.* **2014**, *16*, 2064-2079

[19] D'Andrilli, J.; Foreman, C.M.; Marshall, A.G.; McKnight, D.M. Characterization of IHSS Pony Lake fulvic acid dissolved organic matter by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry and fluorescence spectroscopy. *Org. Geochem*. **2013**, *65*, 19-28

[20] Wozniak, A.S.; Bauer, J.E.; Sleighter, R.L.; Dickhut, R.M.; Hatcher, P.G. Technical Note: Molecular characterization of aerosol-derived water-soluble organic carbon using ultrahigh resolution electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Atmos. Chem. Phys.* **2008**, *8*, 5099-5111

[21] Hertkorn, N.; Benner, R.; Frommberger, M.; Schmitt-Kopplin, P.; Witt, M.; Kaiser, K.; Kettrup, A.; Hedges, J.I. Characterization of a major refractory component of marine dissolved organic matter. *Geochim. Cosmochim. Acta* **2006**, *70*, 2990-3010

[22] 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. *Anal. Chem*. **2003**, *75*, 5336-5344

[23] Martins, N.; Jiménez-Morillo, N.T.; Freitas, F.; Garcia, R.; Gomez da Silva, M.; Cabrita, M.J. Revisiting 3D van Krevelen diagrams as a tool for the visualization of volatile profile of varietal olive oils from Alentejo region, Portugal. *Talanta* **2020**, *207*, 120276-120285

[24] Wu, Z.; Rodgers, R.P.; Marshall, A.G. Two- and three-dimensional van Krevelen Diagrams: A graphical analysis complementary to the Kendrick mass plot for sorting elemental compositions of complex organic mixtures based on ultrahigh-resolution broadband Fourier Transform ion cyclotron resonance mass measurements. *Anal. Chem.* **2004**, *76*, 2511-2516

describe the correlations.[25] On the other hand, supervised and unsupervised methods are applied for qualitative datasets.[26,27] Principal component analysis, an unsupervised method, is usually implemented as a first approach for visualization, dimensionality reduction, classification, and finding patterns of similarities in the dataset.[28,29] Supervised methods require *a priori* information, meaning that classes, determined by specific qualitative properties, are known in advance, and this information is used to sharpen the distinction between the given classes. A subclass of these methods, discriminant analysis, studies why the classes are different and which variables drive their separation, bearing the largest discriminatory power (e.g., Partial Least Squares Discriminant Analysis).[30] The main areas of application in mass spectrometry data interpretation include food analysis and

[25] Massart, D.L.; Vandeginste, B.G.M.; Deming, S.M.; Michotte, Y.; and Kaufman, L. Chemometrics: a textbook. **1988**, 165-182

[26] Jurs, P.C. Pattern recognition used to investigate multivariate data in analytical chemistry. *Science*, **1986**, *232*, 1219-1224

[27] Kemsley, E.K. Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemom. Intell. Lab. Syst.* **1996**, *33*, 47-61

[28] Abdi, H.; Williams, L.J. Principal component analysis. *WIREs Comput. Stat.* **2010**, *2*, 433-459

[29] Ringnér, M. What is principal component analysis?, *Nat. Biotechnol*. **2008**, *26(3)*, 303-304

[30] Bylesjö, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J.K.; Holmes, E.; Trygg, J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J. Chemom.* **2006**, *20*, 341-351

authentication,[31,32,33,34] environmental sample analysis,[35,36] proteomics,[37,38] metabolomics in diagnostics and biology,[39,40,41] and imaging.[42,43] However, when dealing with high-dimensionality data (e.g., direct infusion HRMS) it is difficult to assess and visualize which variables account for the

[31] Callao, M.P.; Ruisanchez, I. An overview of multivariate qualitative methods for food fraud detection. *Food Control* **2018**, *86*, 283-293

[32] Marti, M.P.; Busto, O.; Guasch, J. Application of a headspace mass spectrometry system to the differentiation and classification of wines according to their origin, variety and ageing. *J. Chromatogr. A* **2004**, *1057*, 211-217

[33] Kenar, A.; Çiçek, B.; Arslan, F.N.; Akin, G.; Karuk Elmas, S.N.; Yilmaz, I. Electron impact-mass spectrometry fingerprinting and chemometrics for rapid assessment of authenticity of edible oils based on fatty acid profiling. *Food Anal. Methods* **2019**, *12*, 1369-1381

[34] Rubert, J.; Lacina, O.; Zachariasova, M.; Hajslova, J. Saffron authentication based on liquid chromatography high resolution tandem mass spectrometry and multivariate data analysis. *Food Chem.* **2016**, *204*, 201-209

[35] Karpuzcu, M.E.; Fairbairn, D.; Arnold, W.A.; Barber, B.L.; Kaufenberg, E.; Koskinen, W.C.; Novak, P.P.; Rice, P.J.; Swackhamera, D.L. Identifying sources of emerging organic principal components analysis. *Environ. Sci. Process. Impacts* **2014**, *16*, 2390-2399

[36] Corilo, Y.E.; Podgorski, D.C.; McKenna, A.M.; Lemkau, K.L.; Reddy, C.M.; Marshall, A.G.; Rodgers, R.P. Oil spill source identification by principal component analysis of electrospray ionization fourier transform ion cyclotron resonance mass spectra. *Anal. Chem*. **2013**, *85*, 9064-9069

[37] Gaspari, M.; Verhoeckx, K.C.M.; Verheij, E.R.; van der Greef, J. Integration of two-dimensional lc-ms with multivariate statistics for comparative analysis of proteomic samples. *Anal. Chem.* **2006**, *78*, 2286-2296

[38] Wang, X.; Chambers, M.C.; Vega-montoto, J.; Bunk, D.M.; Stein, S.E.; Tabb, D.L. QC metrics from CPTAC Raw LC-MS/MS data interpreted through multivariate statistics. *Anal. Chem*. **2014**, *86*, 2497-2509

[39] Wang, C.; Kong, H.; Guan, Y.; Yang, J.; Gu, J.; Yang, S.; Xu, G. Plasma phospholipid metabolic profiling and biomarkers of type 2 diabetes mellitus based on high-performance liquid chromatography/electrospray mass spectrometry and multivariate statistical analysis. *Anal. Chem*. **2005**, *77*, 4108-4116

[40] A. Kiss, A.; Lucio, M.; Fildier, A.; Buisson, C.; Schmitt-Kopplin, P.; Cren-Olivé, C. Doping control using high and ultra-high resolution mass spectrometry based non-targeted metabolomics - a case study of salbutamol and budesonide abuse. *PLoS One* **2013**, *8*, 1-13

[41] Tsugawa, H.; Tsujimoto, Y.; Arita, M.; Bamba, T.; Fukusaki, E. GC/MS based metabolomics: development of a data mining system for metabolite identification by using soft independent modeling of class analogy (SIMCA). *BMC Bioinform.* **2011**, *12*, 131-144

differences. Usually, variable selection[44] or sparse methods[45] must be applied, which are able to remove or suppress variables that are irrelevant to response prediction or classification.[46] These methods proved their efficiency but have to be used with caution to avoid losing valuable information, to prevent overfitting, and to handle chance correlations correctly. In summary, multivariate analysis tools enable global understanding of many concomitant variables and of their inter-correlations. The concept behind multivariate analysis is different from that of the SPIX software: the latter aims at observing all statistically relevant variables individually.

[42] Dill, A.L.; Eberlin, L.S.; Zheng, C.; Costa, A.B.; Ifa, D.R.; Cheng, L.; Masterson, T.A.; Koch, M.O.; Vitek, O.; Cooks, R.G. Multivariate statistical differentiation of renal cell carcinomas based on lipidomic analysis by ambient ionization imaging mass spectrometry. *Anal. Bioanal. Chem.* **2010**, *398*, 2969-2978

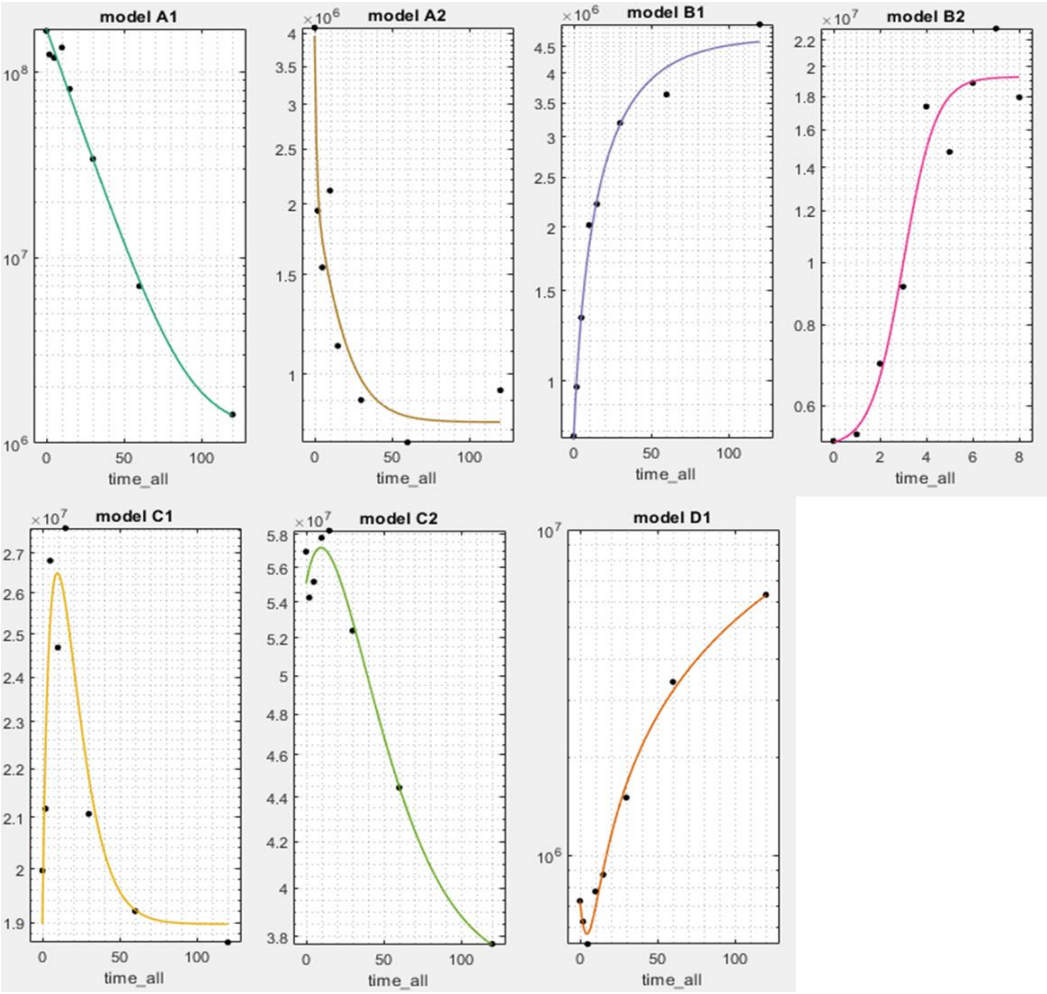
[43] Alexandrov, T. MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinform.* **2012**, *13*, S16-S11

[44] Peres, F.A.P.; Fogliatto, F.S. Variable selection methods in multivariate statistical process control: A systematic literature review. *Comput. Ind. Eng.* **2018**, *115*, 603-619

[45] Zou, H.; Hastie, T.; Tibshirani, R. Sparse principal component analysis. *J. Comput. Graph. Stat.* **2006**, *15*, 265-286

[46] Filzmoser, P.; Gschwandtner, M.; Todorov, V. Review of sparse methods in regression and classification with application to chemometrics. *J. Chemom.* **2012**, *26*, 42-51

SI-2. Current kinetic models in SPIX



Model	Theoretical equation	Experimental equations for the examples shown
A1	f=f0+A*exp(-k*t) k>0	1.18E+06 + 1.70E+08 * exp(-5.50E-02 * t)
A2	f=f0+A1*exp(-k1*t)+A2*exp(-k2*t) k1>0 ; k2>0	8.22E+05 + 1.89E+06 * exp(-9.61E-01 * t) + 1.25E+06 * exp(-6.93E-02 * t)
B1	f=f0+B*(1-exp(-k*t)) k>0	4.69E+06 - 3.91E+06 * exp(-3.17E-02 * t)
B2	f=A+B/(1+C*exp(-k*t)) k>0	1.21E-12 + 1.47E+07 / (1 + exp(-4.83E-02 * (t - 65.23)))
C1	f=f0+A*(exp(-k1*t)-exp(-k2*t)) k2 > k1 > 0	1.90E+07 + 3.25E+08 * (exp(-1.00E-01 * t) - exp(-1.06E-01 * t))
C2	f=f0+A1*exp(-k1*t)-A2*exp(-k2*t) k2 > k1 > 0 ; A1 > A2	3.64E+07 + 3.70E+08 * exp(-3.96E-02 * t) - 3.52E+08 * exp(-4.29E-02 * t)
D1	f=f0+A1*exp(-k1*t)-A2*exp(-k2*t) k1 > k2 > 0 ; A2 > A1	1.01E+09 + 6.22E+05 * exp(-2.14E-01 * t) - 1.01E+09 * exp(-5.12E-05 * t)

SI-3. Chemicals, reagents, irradiation processes and sample preparation

Chemicals and reagents

Acetamiprid ((E)-N-(6-chloro-3-pyridylmethyl)-N'-cyano-N-methylacetamidine), maprotiline hydrochloride (*N*-methyl-3-(1-tetracyclo[6.6.2.$0^{2,7}$.$0^{9,14}$]hexadeca-2,4,6,9,11,13-hexaenyl) propan-1-amine;hydrochloride), acetonitrile (ACN) and formic acid (FA) (chromatographic grade purity > 99.99% for both) were purchased from Sigma-Aldrich (Saint Quentin Fallavier, France). Suwannee River Fulvic Acid Standard was purchased from International Humic Substances Society (Denver, Colorado, USA). Ultrapure water (specific resistance, 18 MΩ cm^{-1} at 25 °C) was produced by a Purelab Chorus 1 water purification system purchased from Veolia Water Technologies (Wissous, France).

Sample preparation

1. Peroxide/UV photodegradation of maprotiline in wastewater

The UV photocatalyzed degradation of maprotiline was carried out in a 45-liter pilot plant with continuous flow at a wastewater treatment plant operated by the FACSA company in Alhama de Murcia (Spain). The molecule was submitted to peroxide/UV advanced oxidation process. The pilot plant included a tank for mixing with a stirrer, a reactor with UV lamp (model UVLA-325-4, controlled by Synergy 3 control panel - ATG UV Technology), a pump for water circulation, a rotameter to assess the water flow, and a compressor (Metabo Basic 250-50 W) to provide airflow in the rotameter equipped-reactor. The wastewater, secondary treated water originating from municipal and industrial sources, was transferred into the pilot plant after undergoing preliminary treatment (screening, sand and grease removal): decantation, biological treatment and sand filtration, before the chlorination step. At this point there were still bacteria in the mixture as well as micropollutants, which the traditional wastewater treatment methods are not able to remove. The wastewater was spiked with Maprotiline hydrochloride at 5 ppm, the total organic carbon content of the mixture was measured as 37.2 mg/L. 4 mL of hydrogen peroxide of technical grade (33 v/w%, VWR chemicals, Llinars del Vallès, Spain) were added to the pilot plant; it corresponds to the stoichiometric amount needed for Maprotiline mineralization. Since the reaction with peroxide radical is usually fast, the reaction time was 10 minutes and the sampling for HRMS (2 x 1 mL) was done at the following

times: t0 (3 samples), 1 min, 2.5 min, 5 min, 7.5 min and 10 min. The only sample preparation before direct infusion mass spectrometry analysis was the addition of 0.1 mL of acetonitrile and 0.1% of formic acid to the 1 mL samples, to achieve better ionization and solubility. Using direct infusion HRMS aimed at suppressing many sample preparation steps, to gain a considerable amount of time and avoid too much variability and operator subjectivity.

2. UV irradiation of acetamiprid in an aqueous solution of humic acid

Acetamiprid has been detected in agriculture water samples at concentrations of up to 44 µg/L.[47] Therefore, a 40 µg/L acetamiprid solution was prepared using an aqueous solution of fulvic acid at 20 mg/L, a mean value regarding the amount usually found in natural waters.[48] Six glass tubes of the solution were simultaneously irradiated for 30 minutes in a laboratory-made reactor equipped with a UV-Vis high-pressure mercury lamp HPL-N125W/542 E27 SC (Philips, Ivry-sur-Seine, France) emitting light on wavelengths ranging from 200 nm to 650 nm, with a maximum irradiation wavelength at 254 nm and a radiation flux of 6200 lm. Each tube contained 50 mL of solution to ensure good surface irradiation. 1 mL of solution was taken twice from each tube before and after irradiation. Samples were analyzed in ESI-MS using automated direct infusion with a solvent made of 50% H_2O/AF (0.1%) and 50% ACN/AF (0.1%) at a flow of 0.002 mL/min. The six replicates and blanks (H_2O/ACN 50/50 v/v) were randomly analyzed and data were extracted in a «.xy » text format so that they can be treated with the SPIX software.

[47] Anderson, T.A.; Salice, C.J.; Erickson, R.A.; McMurry, S.T.; Cox, S.B.; Smith, L.M. Effects of landuse and precipitation on pesticides and water quality in playa lakes of the southern high plains. *Chemosphere*, **2013**, *92, issue 1*, 84-90

[48] Thurman, E.M. Amount of organic carbon in natural waters. In: Organic geochemistry of natural waters, *Springer, Dordrecht*, **1985**, *2*, 7–65

SI-4. Exported data from the SPIX software after assignation of a kinetic model to a m/z ratio

Segment	m/z	Intensity	Time	File
65	278.19055	3.53E+10	0.0	Maprotiline WW H2O2 0min B.xy
65	278.19057	3.39E+10	1.0	Maprotiline WW H2O2 1min B.xy
65	278.19050	8,34E+09	2.5	Maprotiline WW H2O2 2min B.xy
65	278.19052	3,74E+09	5.0	Maprotiline WW H2O2 5min B.xy
65	278.19057	1,95E+09	7.5	Maprotiline WW H2O2 7min B.xy
65	278.19053	8,84E+08	10.0	Maprotiline WW H2O2 10min B.xy

m/z	r^2	p-value	model
278.19053	0.99	0.000776	A1