Journal Name

ARTICLE TYPE

Cite this: DOI: 10.1039/xxxxxxxxxx

# On the benefits of using multivariate analysis in mass spectrometric studies of combustion-generated aerosols

D. Duca[a], C. Irimiea[b], A. Faccinetto[c], J. A. Noble[a,†], M. Vojkovic[a], Y. Carpentier[a], I. K. Ortega[b], C. Pirim[a] and C. Focsa[a]

Received Date
Accepted Date

DOI: 10.1039/xxxxxxxxxx

www.rsc.org/journalname

Detailed molecular-level analysis of combustion emissions may be challenging even with high-resolution mass spectrometry. The intricate chemistry of the carbonaceous particles surface layer (which drives their reactivity, environmental and health impacts) results in complex mass spectra. Building on a recently proposed comprehensive methodology (encompassing all stages from sampling to data reduction), we propose herein a comparative analysis of soot particles produced by three different sources: a miniCAST standard generator, a laboratory diffusion flame and a single cylinder internal combustion engine. The surface composition is probed by either laser or secondary ion mass spectrometry. Principal component analysis and hierarchical clustering analysis proved their efficiency in both identifying general trends and evidencing subtle differences that otherwise would remain unnoticed in the plethora of data generated during mass spectrometric analyses. Chemical information extracted from these multivariate statistical procedures contributes to a better understanding of fundamental combustion processes and also opens to practical applications such as the tracing of engine emissions.

## 1 Introduction

Multivariate analysis (MVA) methods are powerful tools to unravel trends in complex databases. They have been successfully applied in the past, for instance, to identify drug metabolites in biological fluids[1], to evaluate profiles of volatile compounds present in mainstream tobacco smoke[2], or else, to assess surface water quality[3]. Among the MVA methods commonly used[4] are the principal component analysis (PCA) and the hierarchical clustering analysis (HCA). The former is used to reveal hidden patterns in databases, by emphasising the variance between samples and thus highlighting their differences and similarities[5], whereas the latter searches for patterns in a database by grouping the observables into distinct clusters. Their capability at distinguishing various complex samples, as exemplified for a while now in the field of biology, has recently led to their consideration for unravelling the chemical composition of multifaceted samples of environmental interest.

Atmospheric aerosols are airborne particles consisting of an intricate mixture of chemical constituents whose nature varies greatly depending upon their emission source and evolution within the atmosphere. Carbonaceous particles account for a significant fraction of atmospheric particulate matter in urban areas (typically 30-50% by mass[6–8]). They are mainly formed of soot, i.e. particles generated by the incomplete combustion of hydrocarbon-based fuels or biomass. Accordingly, soot particles hold a multitude of chemical compounds derived from various sources (remnant of fuels, combustion and/or post-oxidation products, etc.) that may have been further transformed (aged) by the time they are analysed due to their continuous interaction with environmental elements (solar rays, water molecules, pollutants, etc.). Soot particles are therefore considered a complex mixture that often needs a concerted analytical scheme to be fully resolved.

Mass spectrometry (MS) based techniques have significantly contributed to better understanding soot chemistry over the years. They are generally robust techniques that do not require extensive sample preparation, and are hence preferred for the analysis of such complex samples. Furthermore, the amount of particulate matter required to perform MS analysis is relatively small. MS based techniques mostly differ by the way the ions

[a] Univ. Lille, CNRS, UMR 8523 – PhLAM – Laboratoire de Physique des Lasers Atomes et Molécules, F-59000 Lille, France. E-mail: dumitru.duca@univ-lille.fr
[b] ONERA – The French Aerospace Laboratory, F-91123 Palaiseau, France.
[c] Univ. Lille, CNRS, UMR 8522 – PC2A – Laboratoire de Physico-Chimie des Processus de Combustion de l'Atmosphère, F-59000 Lille, France.
[†] Current address: CNRS, Aix Marseille Université, PIIM, UMR 7345, 13397 Marseille cedex, France.

transferred to the mass spectrometer are created (e.g. soot particle aerosol mass spectrometry (SP-AMS)[9], two-step laser mass spectrometry (L2MS)[10], time-of-flight secondary ion mass spectrometry (ToF-SIMS)[11,12]), which often condition their specificity to provide information on either bulk or surface chemical composition. Ultra high resolution mass analyzers as Orbitrap, Fourier transform ion cyclotron resonance (FT-ICR) and high resolution quadrupole time of flight MS can reach a resolving power higher than 90 000[13,14]. These techniques were developed mainly for proteomics and pharmaceutical analyses, but lately their application has been extended to many other fields among which they start being used and adapted to atmospheric aerosols[15,16]. However, ultra high resolution mass spectrometry is still very rarely applied to the analysis of combustion products, with only a few examples to date[17]. Ultra high resolution mass analyzers are powerful analytical tools, however they still need validation of the sampling protocols. For instance, the sample transfer into the instrument is based on nanospray desorption electrospray using a polar solvent for Orbitrap, followed more recently by laser desorption for FT-ICR and atmospheric pressure chemical ionization (APCI) for APCI-Orbitrap[13,16,17]. Let us also emphasize that in directed energy (laser, ion beam) desorption methods, beside the analyzer performances, the condensed-gas phase transfer itself plays a critical role in the maximum achievable mass resolution and on the total number of detected signals, trough, e.g., the sample/substrate roughness or conductive properties. We therefore stress the need for a thorough evaluation (and optimization) of the entire analysis chain, from sample collection/deposition on suitable substrates, to sample transfer/ionization into gas phase, ions mass separation and detection, and finally powerful data treatment and interpretation[18,19].

Mass spectra of soot particles can be very complex, featuring hundreds and even thousands mass peaks, which quickly renders the interpretation of mass spectra difficult and therefore limits the potentiality of MS to resolve complex mixtures. Accordingly, resolving sample complexity in MS databases is currently tackled using two main approaches. The first is based on the identification of marker species, i.e. compounds that are directly linked to a source/process and that can thus be considered as their fingerprints, while the second approach relies on statistical methods. In particular, the use of MVA methods in conjunction with MS is a creative combination to exploit all of the information given by a multitude of peaks within a great variety of sample sets. Both approaches are widely used in analysis of mass spectra obtained with aerosol mass spectrometers (AMS)[20–22], proton transfer reaction mass spectrometers (PTR-MS)[23,24], and laser-based MS techniques[19,25,26]. Discrimination using marker species was applied to samples of various sources, proving its effectiveness when comparing soot emitted from wood combustion[20,27], on-road vehicles[25], aircrafts[22–24,28,29], ships[30] or other ambient aerosols[21]. However, since some marker species may not remain stable over the aerosols' life span, especially upon atmospheric ageing[6], this method may misdirect with regards to the origin of samples a priori unknown. To circumvent this limitation, MVA approaches are chosen, as they can discriminate samples regardless of their provenance or evolution. Therefore, MVA can uncover trends and features even in samples of unknown/mixed origins[28,31], which is particularly interesting when analysing natural aerosols.

In constant interaction with their surroundings, aerosols surfaces drive their overall reactivity, and therefore, set their evolution path within the atmosphere (sedimentation, formation of secondary organic aerosols, nucleation, etc.). It is hence imperative to uncover their complex surface composition in order to assess their impact on both human health and the environment[32,33]. For example, some polycyclic aromatic hydrocarbons (PAHs), often found adsorbed on the surface of soot particles, are known to be toxic and to have mutagenic effects[34,35]. In addition, the chemical composition of aerosol surfaces determines their hygroscopicity[36] and therefore their ability to act as condensation nuclei, potentially influencing climate forcing, cloud cover and precipitations.

Our group has been addressing this issue of untangling surface chemical compositions of field-collected or laboratory-generated combustion aerosols for over a decade[10,18,19,26,29,30,37–40]. We recently described an original and comprehensive experimental methodology[18] that we later implemented in combining statistical-based approaches with compound classification techniques[19]. This latter systematic study by Irimiea and coworkers[19] was undertaken to characterise over 100 samples collected from different flames. In this work we developed a comprehensive protocol that allowed significant progress towards the fundamental understanding of soot nucleation and growth. Laboratory flames or standard soot generators are often used to produce soot particles with similar physico-chemical properties to the ones produced by "real world" combustion sources[41]. Laboratory soot particles offer the advantages of a reproducible, easy-access and low-cost production, which is of great importance when testing the robustness of a protocol. Therefore, this necessary step is of paramount importance for further refinements in field-collected combustion-generated particle analyses.

## 2 Experimental

In this section, the choice of the combustion conditions, the sampling approach and the experimental techniques used to characterised the samples are detailed. In particular, L2MS and SIMS are used in parallel to obtain information on the chemical composition of combustion generated aerosols.

### 2.1 Soot samples

Soot samples are generated in different combustion conditions (fuel, burner and sampling method) in order to test the ability of our data treatment protocols to reveal differences and similarities between samples. The sampling procedure, including the substrate choice and its preparation, is optimised according to our previous experience[18]. In particular, the sample-substrate reactivity can lead to the formation of a large number of byproducts that clutter the mass spectrum and make the identification of individual compounds much more difficult. A short description of all analysed samples (summarised in Table 1) is given below. The following soot samples have been used:

- Soot produced by a miniCAST generator (5201c) from Jing


This journal is © The Royal Society of Chemistry [year]

Ltd., which is currently proposed as a means of obtaining "standard" soot easily comparable to other studies [41–43]. The main difference between the miniCAST working points is the oxidation flow (1.50 → 1.15 → 1.00 L min$^{-1}$) resulting in three different combustion conditions ($C_1 \rightarrow C_2 \rightarrow C_3$) [41–43]. The hereby generated particles are subsequently deposited on quartz fibre filters.

- Soot produced by laboratory turbulent diffusion flames supplied with two different liquid fuels: diesel (D1-5) and kerosene (K1-5). Soot particles are sampled from the flame at different height above the burner (HAB) and deposited by impaction on Si wafers. Sampling at various HAB is a means of investigating soot particles of different maturity [38].

- Soot produced by a gasoline single cylinder internal combustion engine (ICE). Operating conditions of this engine (e.g. injection and ignition crank angle, applied load) could be easily changed, thus allowing exhausts sampling at various working regimes. The following operating points were used:

  – normal engine operation, i.e. engine optimised in terms of high efficiency and low particle emissions, with medium (GOM) and high (GOH) applied loads, which simulate different driving regimes;

  – malfunction simulation with a medium load applied: low air/fuel ratio resulting in a high-sooting regime (GEF) and an addition of oil to the combustion chamber (GEO).

  Soot particles are sampled using a cascade impactor (NanoMOUDI) to enable for size selection during sampling, and deposited on Al foils. We analysed the particles collected on the last five stages, having diameter in the range 10-180 nm (Table 1).

Off-line analysis of soot particles requires a careful choice of the deposition substrate, not only to minimise the risk of contaminating the samples, but also to ensure that a high mass resolution can be achieved. In particular, among other factors, the mass resolution is directly linked to the surface roughness of the substrate, and can be maximised by depositing the samples on ultra-flat surfaces such as Si or Ti wafers. Furthermore, the sample-substrate reactivity can lead to the formation of reaction byproducts that may heavily interfere with the assignment of sample-specific signals. Therefore, the careful characterization/choice of the deposition substrate is mandatory and the comprehensive identification of its possible reactivity byproducts is necessary for a valid analytical protocol [18,19]. Regardless of its nature, the substrate should undergo a series of preparation steps before it can be used to collect particulate matter.

### 2.2 Two-step laser mass spectrometry (L2MS)

This laser-based MS technique has been extensively used by our group to characterise the chemical composition of combustion byproducts during the last decade [10,18,26,29,30,37–39]. The main advantages of L2MS are its high sensitivity and selectivity with regards to specific classes of compounds thanks to resonant ionisation processes that can be tuned to reach for instance the sub-fmol limit for the detection of PAHs [10,37]. In addition, the controlled laser desorption process ensures a soft removal of molecules adsorbed on the particle surface (typically sub-monolayer regime), and thus avoids/limits either their fragmentation or the in-depth damaging of the underlying carbon matrix [37]. This qualifies L2MS as a surface-sensitive analysis technique, comparable in limit of detection ($\sim 10^{-6}$ monolayers) with static-mode secondary ions mass spectrometry (SIMS, see below), but with much lower analyte fragmentation. However, our previous L2MS studies were limited by a mass resolution of $m/\Delta m \sim 1000$, significantly lower than the one achievable in SIMS (up to $m/\Delta m \sim 10\,000$, depending on the deposition substrate [18,19]). In the current work, we take benefit of the recent implementation of a new mass spectrometer (Fasmatech S&T) which combines ion cooling, Radio Frequency (RF) guiding and Time of Flight (ToF) analyser to reach a mass resolution of about $m/\Delta m \sim 15000$. In this new experimental setup, the sample, placed under vacuum ($10^{-8}$ mbar residual pressure), is irradiated at 30° angle of incidence by a frequency doubled Nd:YAG laser beam (Quantel Brilliant, $\lambda = 532$ nm, 4 ns pulse duration, $\sim$50 mJ cm$^{-2}$ fluence, 10 Hz repetition rate) focused to a 0.3 $mm^2$ spot on the surface. The desorbed compounds form a gas plume expanding in the vacuum normally to the sample surface, and are ionised by an orthogonal UV laser beam (Quantel Brilliant, $\lambda_i = 266$ nm, 4 ns pulse duration, 10 Hz repetition rate, $\sim$0.3 J cm$^{-2}$ fluence). At this ionisation wavelength, a high sensitivity is achieved for PAHs through a resonance enhanced multiphoton ionisation process 1+1 REMPI [44–46]. Care must be taken on the coupling of the desorption and ionisation steps in this laser-based MS technique [47–49]. Moreover, by changing the ionisation wavelength, one can target different classes of compounds. The generated ions are then RF-guided to a He collision cell for thermalisation and subsequently mass analysed in a time of flight mass spectrometer (ToF-MS).

### 2.3 Secondary Ion Mass Spectrometry (SIMS)

In addition, the samples are characterised by using a commercial IONTOF ToF-SIMS$^5$ secondary ion mass spectrometer with maximum resolving power of $m/\Delta m \sim 10\,000$. In short, samples are placed in the analysis chamber with a residual pressure of $\sim 10^{-7}$ mbar. The surface of the sample is bombarded by a 25 keV $Bi_3^+$ ion beam with a current of 0.3 pA in static mode. A small fraction of the ejected atoms/molecules are ionised (secondary ions) and can thus be analysed using a time-of-flight tube (V mode). Mass spectra are recorded in both positive and negative polarities, to obtain the maximum amount of information on the sample [18,19].

## 3 Data Analysis Methodology and Examples of Applications

The data presented below is analysed following an approach structured in three main points that include: mass defect analysis for identification of unknown compounds (Section 3.1), multivariate analysis for the reduction of the number of dimensions of

**Table 1** Soot samples used to put in evidence the proposed methodology

| Name | Fuel | Source | Substrate | Description | Analysing technique |
|---|---|---|---|---|---|
| C1 | propane | miniCAST | Quartz fibre filters | 1.5 l/min oxidation flow | L2MS + |
| C2 | | | | 1.15 l/min oxidation flow | |
| C3 | | | | 1.0 l/min oxidation flow | |
| D1 | diesel | diffusion flame | Si wafer | HAB = 6mm | SIMS +/- |
| D2 | | | | HAB = 12mm | |
| D3 | | | | HAB = 14mm | |
| D4 | | | | HAB = 18mm | |
| D5 | | | | HAB = 24mm | |
| K1 | kerosene | diffusion flame | Si wafer | HAB = 6mm | SIMS +/- |
| K2 | | | | HAB = 12mm | |
| K3 | | | | HAB = 14mm | |
| K4 | | | | HAB = 18mm | |
| K5 | | | | HAB = 24mm | |
| GOM1 | gasoline | ICE, optimal conditions, medium load | Al foil | ⌀100 - 180nm | SIMS +/- |
| GOM2 | | | | ⌀56 - 100nm | |
| GOM3 | | | | ⌀32 - 56nm | |
| GOM4 | | | | ⌀18 - 32 nm | |
| GOM5 | | | | ⌀10 - 18 nm | |
| GOH1 | gasoline | ICE, optimal conditions, high load | Al foil | ⌀100 - 180nm | SIMS +/- |
| GOH2 | | | | ⌀56 - 100nm | |
| GOH3 | | | | ⌀32 - 56nm | |
| GOH4 | | | | ⌀18 - 32 nm | |
| GEF1 | gasoline | ICE, low Air/Fuel ratio | Al foil | ⌀100 - 180nm | SIMS +/- |
| GEF2 | | | | ⌀56 - 100nm | |
| GEF3 | | | | ⌀32 - 56nm | |
| GEF4 | | | | ⌀18 - 32 nm | |
| GEO1 | gasoline | ICE, addition of oil | Al foil | ⌀100 - 180nm | SIMS +/- |
| GEO2 | | | | ⌀56 - 100nm | |
| GEO3 | | | | ⌀32 - 56nm | |
| GEO4 | | | | ⌀18 - 32 nm | |

the dataset (Section 3.2) and eventually mass peak grouping for uncovering hidden trends and highlight correlations between different classes of compounds (Section 3.3). This section details the proposed data treatment protocol. Mass spectra of the previously described samples have been used to demonstrate its advantages, including its universal character (the ability to be used with mass spectra of various samples, obtained with different experimental techniques). Mass spectra were recorded with either L2MS or SIMS in multiple regions of the sample surface, to ensure the consistency of the method and to build a database allowing a more advanced statistical analysis. Once all the peaks coming from the substrate are removed, the data is ready to be processed.

### 3.1 Mass defect analysis

Mass defect analysis is used to assign a molecular formula to the recorded accurate mass[50,51]. By convention, the mass defect of $^{12}C$ is defined as zero, therefore the mass defect of every other existing isotope is either positive or negative, depending on its relative nuclear binding energy to $^{12}C$. Since each nuclide has unique mass defect, molecules with different isotopic composition have unique exact mass. For example, while a resolving power of around 5000 is sufficient to completely separate $C_{14}H_{10}^+$ and $C_{13}H_6O^+$, for closely spaced ions the required resolving power can easily increase up to $10^5$ or even higher. As the m/z increases, the number of combinations of different elements resulting in the same nominal mass grows very fast. This experimental limitation is already tackled in Irimiea *et al.* [19] when discussing the role of oxygen containing compounds. Nevertheless, a lower mass resolution mass spectrum can provide several helpful information. In particular, in the investigation of soot particles sampled from laboratory flames C, H and O are the major contributors to the total mass of soot, and therefore the mass analysis of peaks with a high signal-to-noise ratio (SNR) can be reasonably limited to $C_mH_nO_p^+$ ions. Identification within 5 ppm, often but not necessarily assumed as "certain"[52], in our work is possible up to m/z $\approx$ 150 – 200. *A priori* knowledge of the samples and experimental conditions can extend this range up to m/z $\approx$ 500 – 550 and lead to self consistent results and coherence with many other works in the literature.

The mass defect analysis can also be used to simplify the visualisation of complex mass spectra (e.g. Figures S1 and S2). This is generally achieved by plotting the mass defects of all peaks versus their nominal mass. The resulting graph (mass defect plot, Figure 1 and S3) enables the visualisation of complex databases in one single plot, and highlights trends that are often invaluable to identify unknown species. For instance aliphatic, aromatic or


This journal is © The Royal Society of Chemistry [year]

polycyclic aromatic hydrocarbons are aligned on different positive slopes corresponding to the addition of H atoms. When analysing samples containing hydrocarbons with different degrees of alkylation, the Kendrick mass defect can be used as an alternative way of presenting the mass defect data[50,51]. Kendrick mass defect is calculated from the re-normalised mass of a repeating molecular fragment to an integer value as shown in Equation 1 for the common case of $CH_2$ ($m$ = 14.01565):

$$m^{Kendrick} = m^{IUPAC} \frac{14.0000}{14.01565} \quad (1)$$

After this conversion, homologous series that contain the repeating fragment have identical Kendrick mass defect and are found aligned on horizontal lines, making their identification even easier[50,53]. This is useful when dealing with repeating alkyl groups for instance, since their mass defect increases regularly with their molecular weight and makes their association to a certain series less intuitive when represented on conventional mass defect plots[50]. The most convenient approach (conventional or Kendrick) heavily depends on the nature of the sample. If the sample is dominated by a variety of different species, the use of the conventional mass defect is more advisable. However, when the mass spectrum contains many species that only differ by a repeating unit such as aliphatic chains for instance (Table S1), Kendrick mass defect is more advantageous (Figure S4).

In this work, mass defect analysis is applied to the data obtained from L2MS and SIMS to demonstrate its effectiveness when dealing with a variety of mass spectrometric data. Figure 1 shows the mass defect plot obtained from sample *C2* analysed by L2MS. The suggested representation merges into one graph important information extracted from the raw mass spectra that include the peaks mass defect (y-axis), nominal mass (x-axis) and relative abundance (dot size). Species that line up in the mass defect plots typically contain a repeating unit. Additionally, the detection of a series of homologous species can help the identification of unknown peaks. This is especially helpful for species with high molecular masses, where the attribution of a chemical formula can be rather delicate.

As PAHs exhibit a high thermodynamic stability[54], they appear in great abundance in all mass spectra and this is amplified by the high sensitivity of the analysis technique to these specific compounds (Figure S1). Since the H/C ratio of PAHs is low compared to other hydrocarbons, they have a relatively small mass defect and are thus easily distinguishable from other hydrocarbons. For instance, aromatic hydrocarbons that contain the same number of hydrogen atoms and progressively increasing number of carbon atoms (e.g. $C_{10}H_8 \rightarrow C_{12}H_8 \rightarrow C_{14}H_8 \rightarrow ... \rightarrow C_{22}H_8$) can be found on the same horizontal line. Besides hydrocarbons, all samples contain oxygen and nitrogen organic derivatives to some extent. As a rule of thumb, in the mass defect plot of combustion generated aerosols, oxygen containing hydrocarbons are often found below the corresponding hydrocarbons due to the large negative mass defect of oxygen. Nitrogen containing hydrocarbons show distinct behaviours. For instance, organic amines are often found mixed to their corresponding hydrocarbons due to the nucleophilicity of nitrogen that results in their tendency to bind one additional hydrogen atom post-ionisation. Organic nitrates, on the other hand, tend to be found at lower mass defect due to the presence of oxygen.

Kendrick mass defect can be used to emphasise some less obvious patterns as shown in Figure S4, in which CH ($m = 13.007825$) is used as the base unit.

### 3.2 Statistical analysis

In this section we detail the chemometric techniques, based on commonly used statistical tools like multivariate analysis, that were adopted by our group to extract chemical information from mass spectrometric data. A mass spectrometry database can contain an extremely variable number of mass spectra (observations), and each of them typically contain up to thousands of peaks (variables). This database structure should be taken into consideration when choosing the most appropriate statistical methods.

#### 3.2.1 Principal component analysis

PCA is a powerful statistical tool that can be used to classify samples and reveal trends and patterns in databases[5], and is often used to increase the readability of very complex data[55]. PCA applied to mass spectrometry is especially useful when many mass spectra are being compared, since it reduces the dimensionality of the database while preserving most of the original information. PCA is a non-parametric analysis, i.e. its output is independent of any hypotheses about data distribution[56]. In this work, PCA is performed on a matrix containing the integrated peaks (variables) against the samples (observations). Before applying PCA, data obtained from mass spectrometry should undergo a special preparation procedure[56,57] that includes calibration, baseline removal, construction of a peak list, peak integration and standardisation. PCA applied to data with no normalisation/standardisation is mostly affected by the largest raw variance, which can skew the overall interpretation of the dataset. Therefore, normalisation techniques are applied to mass spectra prior PCA analysis when there are differences in the samples weight, volume or other properties that may result in additional sources of variance. The most popular and generally recommended normalisation method is the normalisation to the total ion count (TIC), i.e. the integrated ion count over a given mass range[18,58,59].

Care has to be taken when building the peak list as it should only contain species representative of the sample. Minor-abundance isotopes are usually excluded from the peak list, thus allowing to focus on the major-abundance isotopic species[58]. Peaks coming from the substrate and/or originate from the sample-substrate reactivity should also be disregarded. Identifying these peaks, especially the ones corresponding to reaction products, can be a difficult task. One approach to their identification involves comparing mass spectra of the sample deposited in the same experimental conditions but on different substrates (e.g. Si and Ti wafers)[18]. Another possibility relies on the use of PCA: species coming from the sample-substrate reactivity become less prominent as the substrate coverage increases and is less available for the reaction, and are thus likely to be found all clustered in the same principal component.

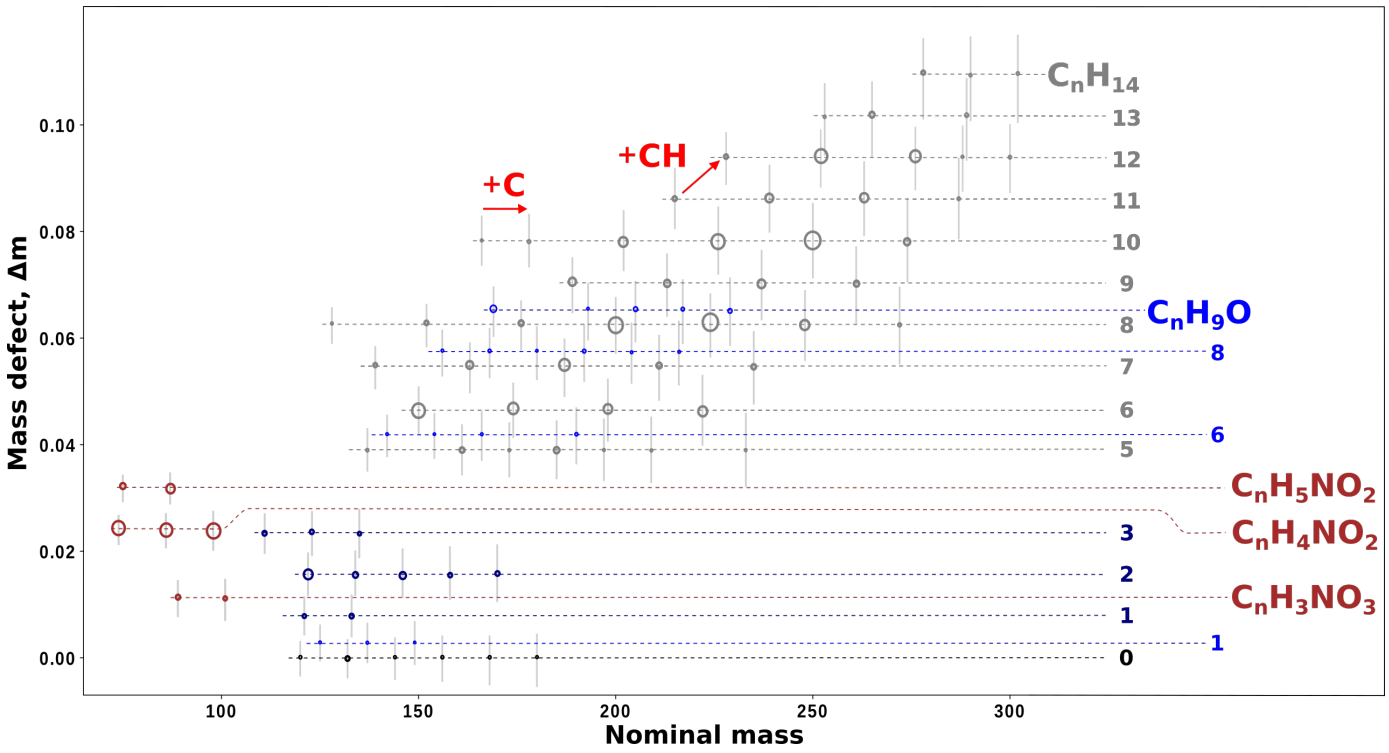


**Fig. 1** Mass defect plot obtained from the L2MS mass spectrum of miniCAST soot, *C2* sample. The data points represent the assigned accurate mass. The size of the data points is proportional to the corresponding peak integrated area, normalised to the total ion count after background subtraction. Molecular formulas of homologous species are displayed. The error bars show the uncertainty on the accurate mass calculated from the obtained mass resolution.

Each principal component (PC) accounts for a defined percentage of the total variance within the data set, are represented in a scree plot and used to select the PCs to take into consideration. The loadings represent the weights of each variable used to calculate the PCs, and are used to understand the contribution of each variable to the selected PC. The distance of an observation from a PC is represented on the scores plot. Scores are obtained for each observation in the database and for each principal component, and are often used as a base to display and classify the samples. In the score plot, similar observations group together and are separated from dissimilar observations. The clustering of the scores is strongly related to the values of the loadings, and they are discussed as a whole. The most challenging part of PCA is the interpretation of individual PCs and their contribution to the investigated processes. To this purpose, there is a vast literature providing general guidelines that should be followed [5,60–62].

To illustrate the potential of this technique, we show below some application to mass spectrometric data of various combustion generated aerosol samples.

#### 3.2.1.1 MiniCAST soot, L2MS

When L2MS mass spectra of miniCAST soot samples are examined, PC1 and PC2 account for ∼ 96% of the total variance, and are therefore only considered for the data interpretation. The three samples are well separated in the PC2 vs. PC1 scores plot (Figure 2). Sample *C1* is highly influenced by $C_{14}H_8$, $C_{14}H_{10}$ and $C_{16}H_{10}$ (high positive PC1 scores) whereas *C2* and *C3* are dominated by higher mass aromatic compounds (negative PC1 scores). It can be noticed that PC2 (∼ 10%) allows for better discrimination between the samples than PC1, especially *C2* and *C3*.

#### 3.2.1.2 Flame and ICE soot, SIMS

PCA is applied to the ensemble of SIMS mass spectra obtained in positive polarity from soot samples generated by the gasoline engine and the laboratory flame (diesel and kerosene fuels). PC1 and PC2 account together for the 73.3% of the total variance. Two main groups are observed in the score plot of both positive and negative ions (Figure 3 and S5). While it was not possible to clearly associate a phenomenon to PC1 (51.7% of total variance), the samples are well separated by the different emission source (engine, GOM, and flame, D and K) in PC2 (21.6% of total variance). At this level of the analysis PCA cannot distinguish soot generated by burning the two different liquid fuels (diesel and kerosene) in laboratory flames, which appear mixed together in negative PC2.

PC1 is mainly associated to high H/C fragment ions (negative contribution, red dots in the loadings plot (Figure 3), and low H/C fragment ions probably resulting from the dissociation of large aromatic hydrocarbons (positive contribution, green dots in the loadings plot). The main contributions to PC2 come from aromatic species (positive contribution, blue dots on the loadings plot), and to a smaller extent to high H/C fragment ions. Therefore, the contribution of high H/C fragment ions, possibly related to the dissociation of aliphatic hydrocarbons, depends less on the fuel and more on the combustion conditions (engine vs. con-


This journal is © The Royal Society of Chemistry [year]

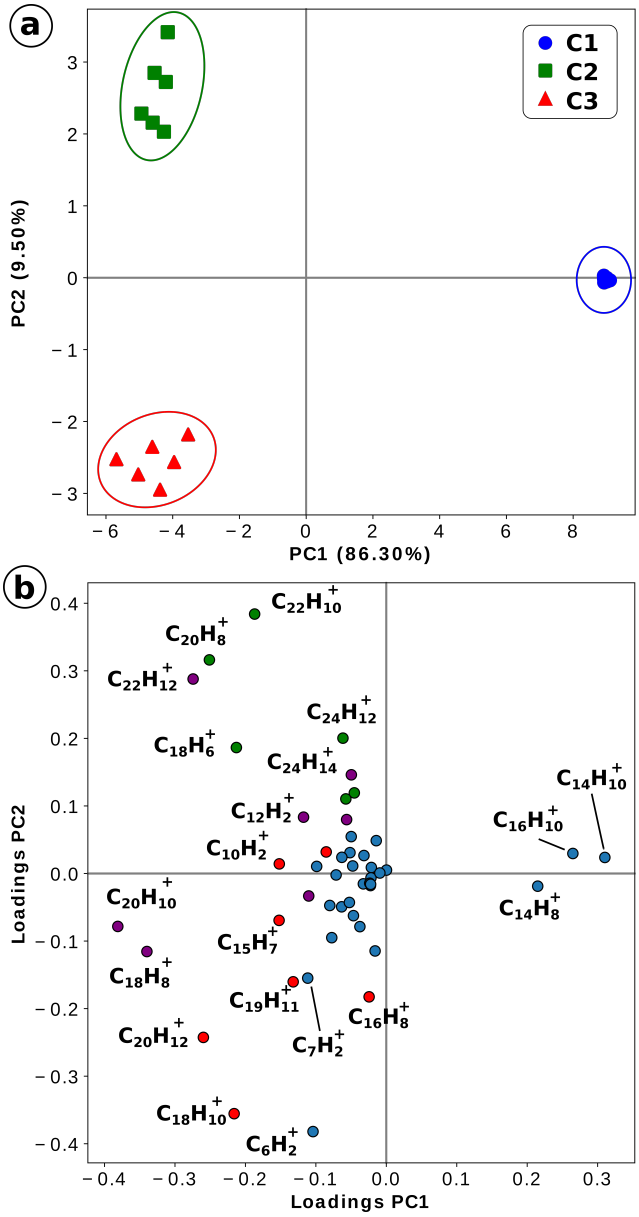


**Fig. 2** Score plots of PC2 vs PC1 for miniCAST soot samples obtained with L2MS – (a). Ellipses highlight data points coming from different samples and are added for visual purposes only. (b) – the corresponding loadings plot of PC2 vs PC1. Several homologous series are highlighted: $C_{n+8}H_n$ – red, $C_{n+10}H_n$ – purple, $C_{n+12}H_n$ – green.

trolled laboratory flames).

Going a step further, PCA is applied to gasoline soot samples obtained in different engine regimes in order to determine their impact on the chemical composition (Figure 4). There is an obvious separation between normal engine operation regimes (GOM, GOH) and the ones which simulate a malfunction (GEF, GEO). A good discrimination is achieved with only the first two components that account for $\sim 98\%$ of the total variance. PC1 alone ($\sim 91\%$) allows the separation of regimes, based on the abundance of aliphatic fragment ions (positive contribution to PC1, marked in red in Figure S6). Consequently, samples that simulate a malfunction (GEF, GEO) are characterised by a higher relative contribution from aliphatic fragment ions compared to optimised engine regimes (GOM, GOH). PC2 is linked to the contribution of aliphatic fragment ions and aromatic species (positive PC2 value), however some aliphatic fragment ions ($C_5H_7$, $C_5H_9$, $C_3H_7$, $C_4H_7$) show a contribution to negative PC2). The data points corresponding to optimal engine regimes form a smaller cluster. This implies that soot produced in conditions simulating engine malfunction shows a much larger variability in chemical composition.

At this point of the analysis, it is clear that the two regimes that simulate a malfunction (GEF, GEO) exhibit similarities, while being well separated from the optimised regimes (upper panel of Figure 4). This implies that the variance of a certain principal component for them is much smaller than the one responsible for the separation between optimised and non-optimised regimes. Consequently, each group should be analysed independently, thus uncovering even smaller contributions to the variance. To demonstrate this concept, the same statistical method was applied a second time to the two non-optimised regimes, and their comparison lead to discriminate between the two main contributors to particulate emissions of the internal combustion engine: fuel and oil, Figure 4. In this case, PC1 ($\sim 71\%$), accountable for the separation of the two regimes, is linked to the contribution of hydrogen-rich hydrocarbons on one side (negative contribution) and of fragment ions and aromatic species on the other (positive contribution). This reveals that oil-related soot particles feature more hydrogen-rich hydrocarbons, while an excess of gasoline leads to the production of more aromatic species, Figure S6. The increase of the contribution of fragment ions in the latter is probably linked to the increase in the aromatic contribution, since the majority of fragment ions can be related to dissociation reactions of PAHs[63]. PC2 ($\sim 20\%$) is associated to the presence of aromatic hydrocarbons (blue dots in Figure S6). One can also notice that samples corresponding to the engine regime with a low air/fuel ratio (GEF1) surprisingly lie in the oil-excess region, while samples GEO3 appear far from the oil-excess region (Figure 4). It is likely that the specific behaviour observed for these samples relates to their particle size (Table 1) but correlating size to chemical composition is out of scope of this paper and will be addressed in a future work.

### 3.2.2 Hierarchical clustering analysis

Hierarchical clustering analysis (HCA) is a MVA method that identifies patterns in a dataset by creating groups of observations called clusters. Unlikely PCA, HCA accounts for the total variance in the database[60,62]. HCA is based on a simple approach for building the clusters that starts with one cluster for each observation and finishes with a single cluster containing the entire database. At each step, the two closest clusters are merged into a single new cluster resulting in a dendrogram representative of the database. In order to decide which clusters to merge, different approaches to measure their distance can be used and give rise to several hierarchical methods[61,62]. In this work, HCA (group average method, Euclidean distances) is applied to the same standardised matrix used for PCA analysis, on both columns (observations) and rows (variables). The HCA output is built in a heatmap organised by the clusters obtained on observations and variables.

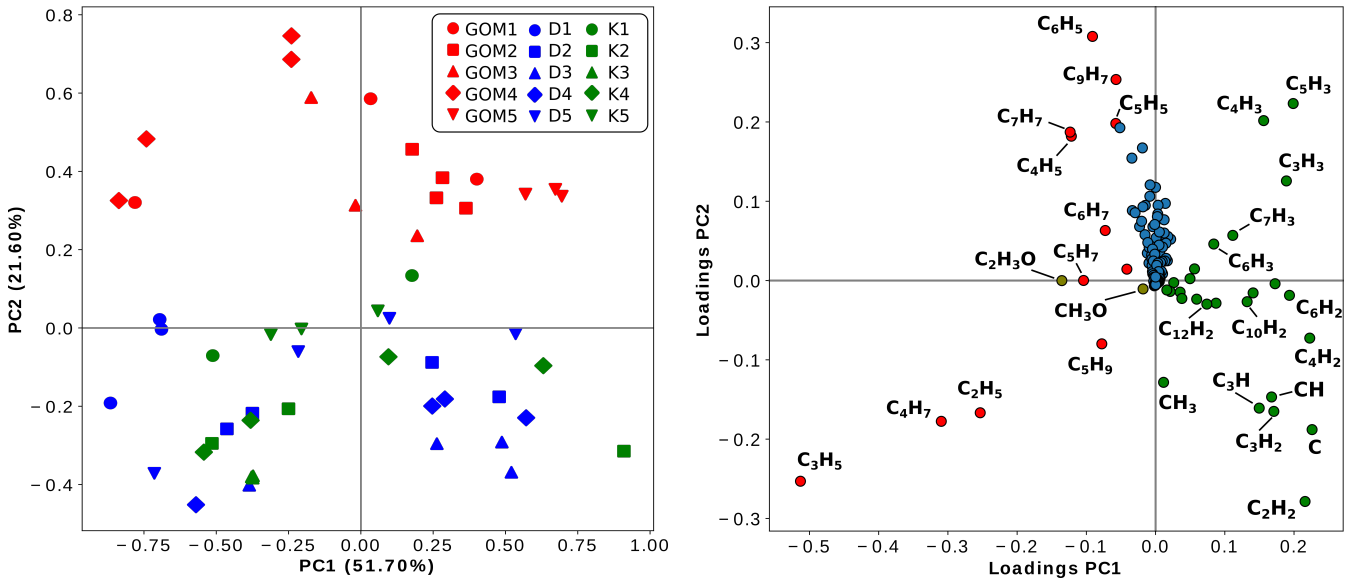


**Fig. 3** Score plot of PC2 vs PC1 for positive ions of soot samples obtained from gasoline engine and laboratory flames (left panel). Corresponding loadings plot of the first two principal components (right panel). For sample description see Table 1.

This representation improves the visualisation of clusters in the multidimensional space, in which each tile represents the value of the correlation between observations and variables.

The heatmap obtained for the samples analysed in SIMS positive polarity is shown in Figure 5. HCA groups the samples in three main clusters (C1, C2 and C3) at distance d1 function of the characteristics of the five clusters of variables (R1, R2, R3, R4 and R5). Cluster C1 is specific to samples GEO1-4, GOM4 and D1 due to the high contribution of compounds with H/C > 1 and identified in the C1-1 cluster. C1-2 is dissimilar from the C1-1 due to the presence of aromatic hydrocarbons and other compounds with low H/C ratio. Soot collected from the gasoline engine in optimal conditions and after the addition of oil are dominated by R5, while there is a shift to R1 and R2 for soot collected from the diesel flame. Contrary to C1, C2 has a high contribution of fragment ions with high (R4) and low (R1) H/C ratio. C2 shows that soot collected from the engine in optimal conditions with high and medium load have similar chemical fingerprint.

This representation offers at once a clustering of the samples function of the three main classes of chemical compounds identified in the mass spectra. For instance, the high content of aromatic hydrocarbons and low H/C fragment ions is specific to soot collected from the kerosene flame. Basically, the addition of oil increases the fraction of high H/C fragment ions in the emissions, the normal operation conditions of the engine have an intermediate content of high H/C fragment ions and a slight contribution of aromatics with four and five aromatic rings, while kerosene soot contains the highest contribution of aromatic compounds and low H/C fragment ions. HCA is also applied to L2MS and SIMS negative polarity data as detailed in the Supplementary Information. In this work, HCA is applied to the raw data corresponding to the selected mass spectra but its usefulness can be extended to more compact data after using another statistical method for sorting the input variables and observations. One of the advantages of this method is that it does not require the raw data set. Moreover, HCA can be used to visualise clusters that form in the principal component space, after applying the PCA, or it can group samples according to other properties (mass defect, contribution from different classes of compounds, etc).

### 3.3 Mass peaks grouping into chemical classes

A detailed description of the soot chemical composition is certainly desirable and can lead to important clues on the soot formation, growth, ageing and reactivity. However, this can rapidly turn into a very cumbersome task, especially if many different samples are analysed. For the sake of simplicity, most of the time, and especially when long time-series of field-collected data are to be treated, individual compounds are grouped in classes (e.g. aliphatics, aromatics, oxygenated, sulphur-containing hydrocarbons and so on). This grouping of mass peaks into appropriate classes allows easier comparison with other experimental measurements (e.g. OC/EC[29]) and facilitates the interaction with modellers that use the data as inputs for various scales simulations. Moreover, this grouping of peaks is also useful when mass spectra of several samples are compared to each other in order to reveal general trends in their chemical composition.

When it comes to the chemical composition of combustion generated aerosols, three non-specific indicators are often considered: amount of ash components (inorganic compounds, IC), amount of carbon associated to the carbonaceous matrix (elemental carbon, EC), and amount of carbon found in organic compounds (organic carbon, OC)[64]. IC alone can sometimes help identify the main source of the emissions. For instance, $K^+$, $Na^+$, $K_2Cl^+$ and $K_3SO_4^+$ in the positive polarity mass spectra and $Cl^-$, $SO_3^-$, $HSO_4^-$ and $KCl_2^-$ in the negative polarity mass spectra are known to be markers of wood combustion[65]. Generally speaking, since IC potentially contains many inorganic compounds, it can


This journal is © The Royal Society of Chemistry [year]

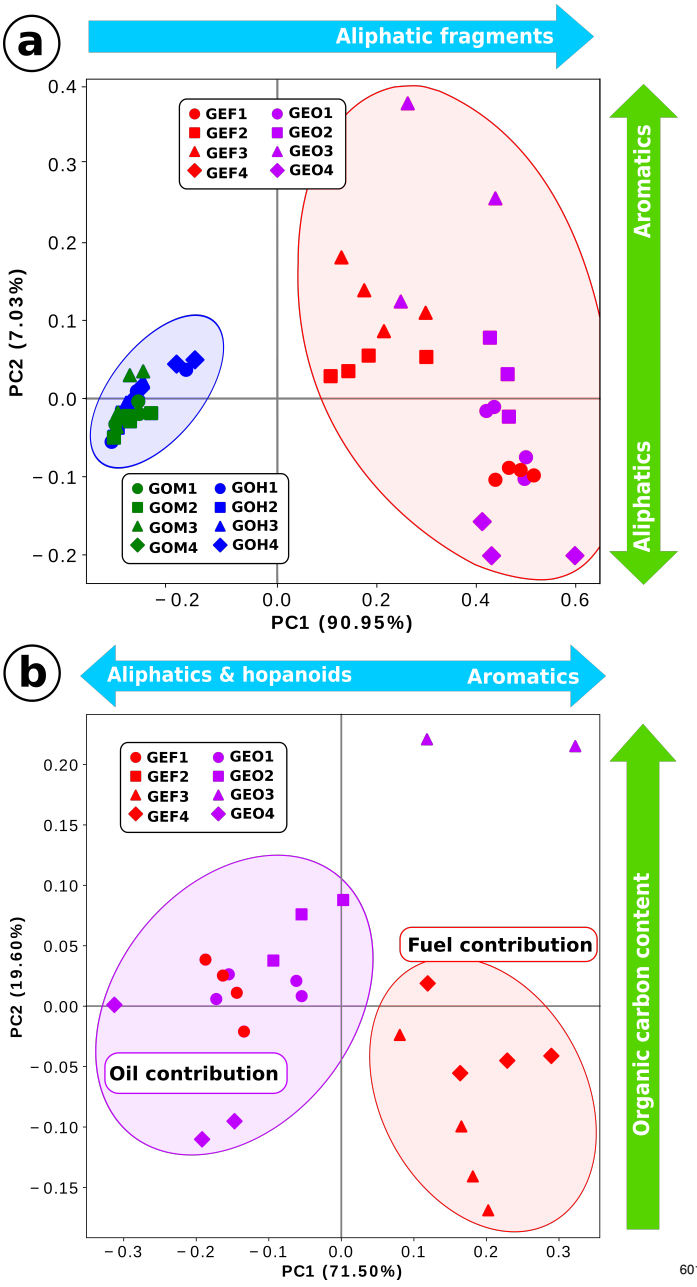


**Fig. 4** Score plots of the first two principal components for soot samples produced by a single cylinder engine. Upper panel – discrimination between different engine regimes, lower panel – particle source discrimination. Ellipses highlight clusters of data points and are for visual purposes only. For sample description see Table 1.

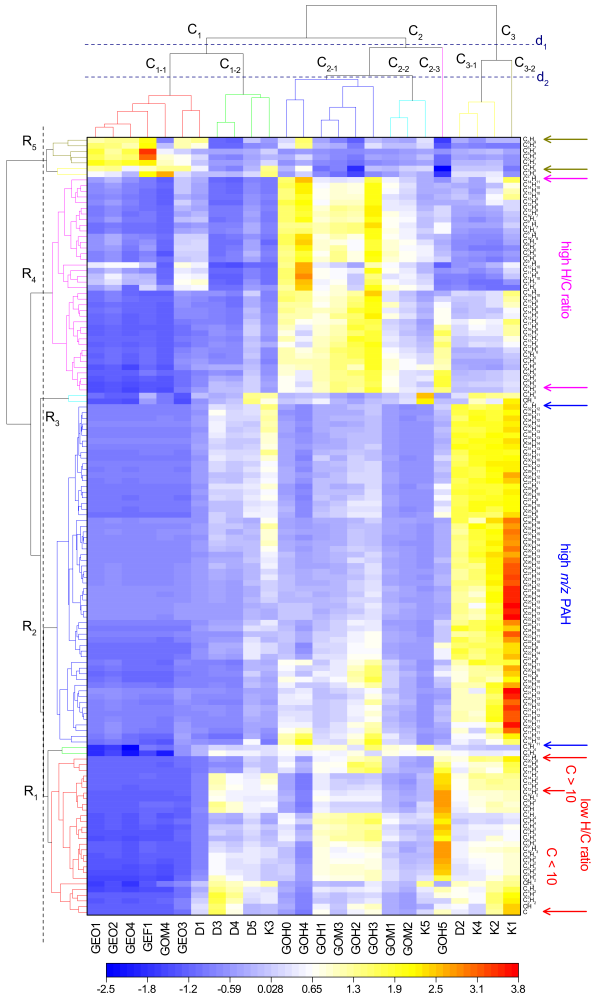


**Fig. 5** Two-way hierarchical clustering heat-map for positive ions of gasoline, diesel and kerosene soot obtained with SIMS. Each column corresponds to the averaged mass spectra obtained for a soot sample. The contribution of each mass in individual samples is expressed as relative value and is represented by the cell colour.

and should be further broken down into source specific groups when characterising complex systems such as internal combustion engines. In this case, accepted grouping of inorganic compounds is: fuel specific (compounds that are coming from fuel additives and trace elements (Na, K) [11,66]), oil specific (detergent and anti-wear additives (P, Ca) [67]) and engine wear tracers (Fe, Al, Cr) [30,67,68]). For addressing the elemental carbon (EC) component, carbon clusters $C_n^-$ (n=2-4) are considered to be appropriate markers in aerosol mass spectrometry [64]. This is also confirmed by the high positive correlation between $C_2^-$, $C_3^-$ and $C_4^-$ signals in the recorded mass spectra [26]. In single particle mass spectrometry, carbon clusters with even higher masses are also considered to be representative of the elemental carbon ($C_5^-$ at 60 u, $C_6^-$ at 72 u and $C_7^-$ at 84 u) [11]. While the handling of IC and EC is relatively straightforward, the OC landscape looks far more complex, with an overwhelming variety of organic compounds, generated in various processes and being themselves main actors of broad-range time-scale reactivity. A subsequent classification of different organic species according to their functional group(s) seems therefore necessary. However, the detailed chemical analysis of a complex mixture of chemicals based on mass spectrometric data only is still an important challenge that requires the identification of the individual ion dissociation patterns. On a practical ground, being able to distinguish these compounds is very important since they all have different sources and roles in the soot formation and ageing mechanisms. For instance, PAHs form during combustion and are well known as building blocks of soot particles and are generally seen as reliable markers of the over-

all OC content[29]. Organic hydroxyl groups are linked to alcohols that are commonly used as additives in gasoline. The presence of many compounds containing carbonyl groups has been proposed as a marker to distinguish fresh emissions from soot particles aged in the atmosphere[69].

A combination of previously described mass peak classification methods is shown in Table 2 along with chemical formula assignments[63]. Detailed classification of molecular ions by functional groups remains difficult by MS alone, however it can be achieved in combination with complementary techniques (e.g. FTIR).[26] Also, for the sake of simplicity, Table 2 displays only the nominal masses, but the peak assignment is based on the exact mass (see mass defect analysis, Section 3.1). The discussion below is based on this grouping of mass peaks.

Depending on the studied samples, the analysis will focus on specific classes from Table 2. For soot samples obtained with the miniCAST standard generator, one may want to address the impact of the oxidation flow. A possible focus is therefore on the evolution of the oxygenated species vs. PAHs (linked to the OC content). Since miniCAST soot is a well-studied standard, it also allows the comparison of mass spectrometric results with the ones reported in the literature based on other experimental techniques. In the present case, Figure 6 clearly shows an increase of the oxygenated species abundance with the oxidation flow, however a low oxidation flow (C2 and C3) leads to the formation of more PAHs, which confirms previous observations on the same set-points of the miniCAST generator[43,70].

Even though examining trends for specific groups can be very informative, when it comes to complex mass spectra containing a multitude of peaks that can be separated in many different ways, not all the groups feature useful trends. It is therefore advisable to first identify the species of interest, groups or individual compounds that can be linked to variations in the chemical composition of the samples. This information can be retrieved from PCA and HCA as discussed in the sections 3.2.1 and 3.2.2, respectively. Based on the statistical analysis of positive polarity SIMS mass spectra of gasoline, diesel and kerosene soot samples, three groups of interest are chosen for further analysis as shown in Figure 6: low-mass and low H/C ions (from the dissociation of aromatic species[63]), low-mass and high H/C ions (from the dissociation of aliphatic species), and finally large aromatic ions (mostly PAHs, stable enough to be detected as molecular ions). Gasoline soot shows higher content of large aromatic compounds, with high and almost constant contribution to all considered particle sizes. Gasoline soot also features the least fragmentation that is well consistent with the higher contribution of large aromatics if compared to diesel and kerosene soot. For the other two fuels, different zones of the flame, corresponding to different stages in the soot formation process, were probed, therefore the variation in aromatic content looks more pronounced. It is clear that the aliphatic content alone cannot be used to discriminate between soot coming from combustion of different fuels, just like it was concluded from PCA. However, it still provides valuable information about different soot maturity. For example, for diesel soot the contribution of aliphatics gradually increases with the sampled HAB ($HAB \geq 12\ cm$). On the other hand the HCA on the negative polarity of SIMS is much easier to interpret because the results clearly discriminate the laboratory flame soot from the one produced with the gasoline engine. The samples belonging to the latest category are clearly evidenced by the presence of sulphur and oxygen containing compounds while the soot from the flames contains mainly OC and EC. Generally speaking, the

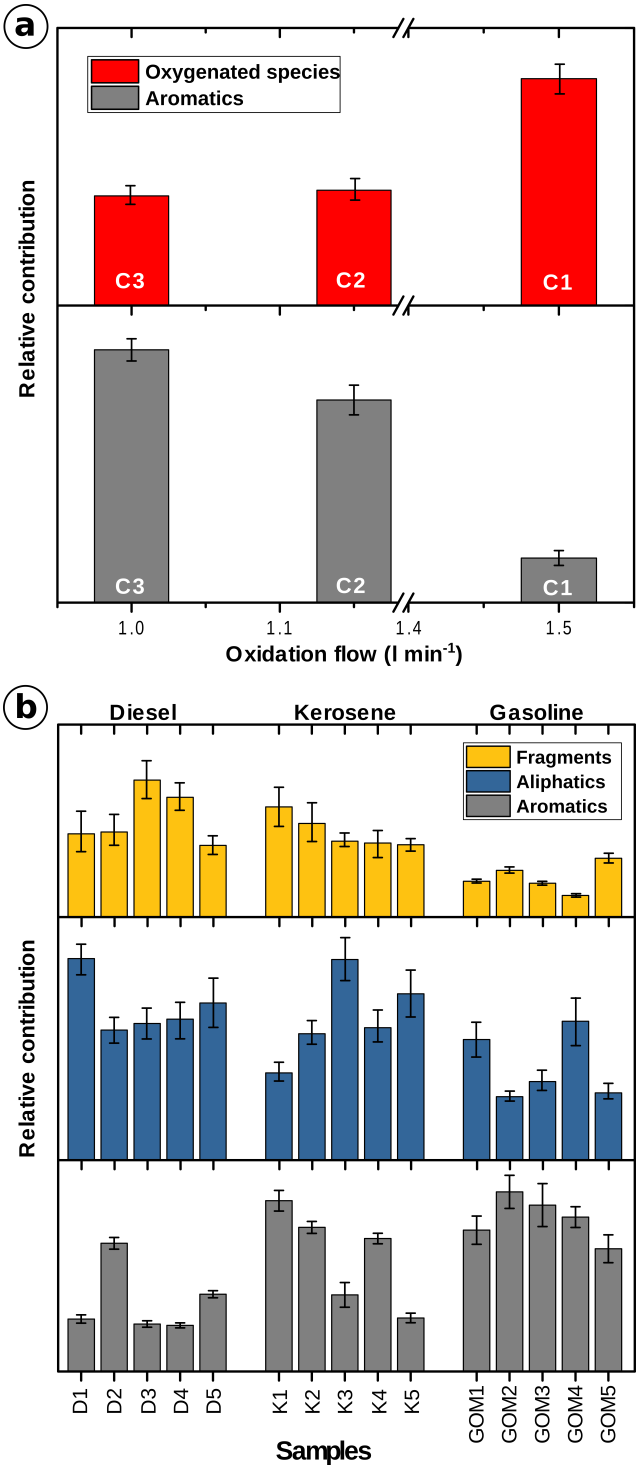


**Fig. 6** Several trends retrieved from mass spectra of: (a) – miniCAST soot (L2MS), (b) – gasoline, diesel and kerosene soot (SIMS).


This journal is © The Royal Society of Chemistry [year]

**Table 2** Grouping of mass peaks into chemical classes

| Category | m/z | Formula | m/z | Formula | m/z | Formula | m/z | Formula |
|---|---|---|---|---|---|---|---|---|
| | 15 | $CH_3$ | 54 | $C_4H_6$ | 71 | $C_5H_{11}$ | 99 | $C_7H_{15}$ |
| Aliphatics | 27 | $C_2H_3$ | 55 | $C_4H_7$ | 81 | $C_6H_9$ | 109 | $C_8H_{13}$ |
| (alkynes, | 29 | $C_2H_5$ | 57 | $C_4H_9$ | 83 | $C_6H_{11}$ | 111 | $C_8H_{15}$ |
| alkene, | 41 | $C_3H_5$ | 67 | $C_5H_7$ | 85 | $C_6H_{13}$ | 113 | $C_8H_{17}$ |
| alkyl, etc.) | 43 | $C_3H_7$ | 68 | $C_4H_8$ | 95 | $C_7H_{11}$ | | |
| | 53 | $C_4H_5$ | 69 | $C_5H_9$ | 97 | $C_7H_{13}$ | | |
| | 26 | $C_2H_2$ | 64 | $C_5H_4$ | 152 | $C_{12}H_8$ | 216 | $C_{17}H_{12}$ |
| | 38 | $C_3H_2$ | 74 | $C_6H_2$ | 154 | $C_{12}H_{10}$ | 228 | $C_{18}H_{12}$ |
| | 39 | $C_3H_3$ | 75 | $C_6H_3$ | 166 | $C_{13}H_{10}$ | 252 | $C_{20}H_{12}$ |
| Aromatics | 40 | $C_3H_4$ | 76 | $C_6H_4$ | 178 | $C_{14}H_{10}$ | 276 | $C_{22}H_{12}$ |
| | 50 | $C_4H_2$ | 78 | $C_6H_6$ | 266 | $C_{21}H_{14}$ | 278 | $C_{22}H_{14}$ |
| | 51 | $C_4H_3$ | 91 | $C_7H_7$ | 190 | $C_{15}H_{10}$ | | |
| | 63 | $C_5H_3$ | 128 | $C_{10}H_8$ | 202 | $C_{16}H_{10}$ | | |
| | 31 | $CH_3O$ | 69 | $C_4H_5O$ | 87 | $C_5H_{11}O$ | 129 | $C_7H_{13}O_2$ |
| | 33 | $CH_5O$ | 71 | $C_4H_7O$ | 89 | $C_5H_{13}O$ | 137 | $C_{10}HO$ |
| | 43 | $C_2H_3O$ | 73 | $C_3H_5O_2$ | 97 | $C_6H_9O$ | 142 | $C_{10}H_6O$ |
| | 45 | $C_2H_5O$ | 73 | $C_4H_9O$ | 97 | $C_5H_5O_2$ | 156 | $C_{11}H_8O$ |
| O-containing | 47 | $CH_3O_2$ | 75 | $C_3H_7O_2$ | 101 | $C_6H_{13}O$ | 166 | $C_{12}H_6O$ |
| (carbonyls, | 47 | $C_2H_7O$ | 75 | $C_4H_{11}O$ | 105 | $C_7H_5O$ | 169 | $C_{11}H_9O$ |
| acids, | 53 | $C_4H_5$ | 81 | $C_5H_5O$ | 109 | $C_7H_9O$ | 180 | $C_{13}H_8O$ |
| ethers, | 55 | $C_3H_3O$ | 83 | $C_5H_7O$ | 111 | $C_6H_7O_2$ | 205 | $C_{14}H_9O$ |
| alcohols, etc.) | 57 | $C_3H_5O$ | 85 | $C_5H_9O$ | 111 | $C_7H_{11}O$ | | |
| | 59 | $C_3H_7O$ | 85 | $C_4H_5O_2$ | 119 | $C_8H_7O$ | | |
| | 61 | $C_2H_5O_2$ | 87 | $C_5H_{11}O$ | 123 | $C_7H_7O_2$ | | |
| | 61 | $C_3H_9O$ | 87 | $C_4H_7O_2$ | 125 | $C_9HO$ | | |
| | 26 | $CN$ | 46 | $CH_4NO$ | 60 | $C_2H_6NO$ | 89 | $C_2H_3NO_3$ |
| N-containing | 29 | $CH_3N$ | 55 | $C_3H_5N$ | 74 | $C_2H_4NO_2$ | 98 | $C_4H_4NO_2$ |
| | 44 | $CH_2NO$ | 55 | $C_2H_3N_2$ | 87 | $C_3H_5NO_2$ | 121 | $C_8H_{11}N$ |
| S-containing | 32 | $S$ | 44 | $CS$ | 46 | $CH_2S$ | | |
| Unclassified | 28 | $C_2H_4$ | 56 | $C_4H_8$ | 84 | $C_6H_{12}$ | 112 | $C_8H_{16}$ |
| hydrocarbons | 42 | $C_3H_6$ | 70 | $C_5H_{10}$ | 98 | $C_7H_{14}$ | | |

trends that are shown herewith are very useful when interpreting the data. However, they are almost impossible to notice in the raw mass spectra. Being able to follow the contribution of a group of related molecules hidden in a much larger ensemble of signals is a powerful feature used to uncover trends that would have remained hidden to a more basic analysis. The fact that PCA and HCA are able to separate the selected samples into categories dependent on their unique pattern of chemical signatures proves that mass spectrometry and MVA provide useful insights into their properties. The usefulness of this approach allows for an easier identification and traceability of combustion generated particles with unknown sources.

## 4 Conclusions

Our recently developed comprehensive methodology (based on mass defect analysis, PCA/HCA multivariate methods)[18] dedicated to the chemical analysis of combustion-generated aerosols is applied here to the study of 30 soot samples generated by three different sources using four different fuels. Laser and secondary ion mass spectrometry techniques are used to probe their surface chemistry. A few examples on the performances of this methodology are provided, showcasing its ability to clearly discriminate samples according to various parameters, such as combustion source, soot maturity, or engine operating conditions. The correlations evidenced by the MVA methods were used for peak clustering to highlight the evolution of grand chemical classes with the combustion conditions. These trends, along with detailed molecular-level information, can further help constrain the processes involved in particulate matter emissions and predict the impact of soot particles on the environment and human health. Moreover, aiming for a standardised (generally accepted) methodology in treating complex mass spectrometry data in aerosol science would certainly allow easier intercomparison and the building of extensive shared databases for further specific developments. An appealing perspective is the possible application of neural networks to this type of big data, which would lead to great advances in automated real-time processing of large dataflows.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

This work was supported by the French National Research Agency (ANR) through the PIA (Programme d'Investissement d'Avenir) under contract ANR-10-LABX-005 (CaPPA – Chemical and Phys-

ical Properties of the Atmosphere), the European Commission Horizon 2020 project PEMs4Nano (H2020 Grant Agreement #724145), and the CLIMIBIO project via the Contrat de Plan Etat Région of the Hauts-de-France region. We thank N. Nuns from the Regional Surface Analysis Platform for assistance with the SIMS measurements.

## References


1 R. S. Plumb *et al.*, *Rapid Communications in Mass Spectrometry*, 2003, **17**, 2632–2638.

2 M. Brokl *et al.*, *Journal of Chromatography A*, 2014, **1370**, 216–229.

3 S. Shrestha and F. Kazama, *Environmental Modelling and Software*, 2007, **22**, 464–475.

4 W. K. Härdle and L. Simar, *Applied Multivariate Statistical Analysis Course*, Springer, 2015.

5 H. Abdi and L. J. Williams, *English*, 2010, **2**, 433–470.

6 J. L. Jimenez *et al.*, *Science*, 2009, **326**, 1525–1529.

7 C. Fountoukis *et al.*, *Atmospheric Chemistry and Physics*, 2014, **14**, 9061–9076.

8 M. Crippa *et al.*, *Atmospheric Chemistry and Physics*, 2014, **14**, 6159–6176.

9 T. B. Onasch *et al.*, *Aerosol Science and Technology*, 2012, **46**, 804–817.

10 A. Faccinetto *et al.*, *Combustion and Flame*, 2011, **158**, 227–239.

11 U. Kirchner *et al.*, *Journal of Aerosol Science*, 2003, **34**, 1323–1346.

12 N. Mayama *et al.*, *Analytical Sciences*, 2013, **29**, 479–482.

13 F. Aubriet and V. Carré, *Analytica Chimica Acta*, 2010, **659**, 34–54.

14 S. Eliuk and A. Makarov, *Annual Review of Analytical Chemistry*, 2015, **8**, 61–80.

15 K. Wang *et al.*, *Atmospheric Environment*, 2018, **189**, 22–29.

16 C. Zuth *et al.*, *Analytical Chemistry*, 2018, **90**, 8816–8823.

17 J. Cain *et al.*, *Physical Chemistry Chemical Physics*, 2014, **16**, 25862–25875.

18 C. Irimiea *et al.*, *Rapid Communications in Mass Spectrometry*, 2018, **32**, 1015–1025.

19 C. Irimiea *et al.*, *Carbon*, 2019, **144**, 815–830.

20 M. F. Heringa *et al.*, *Atmospheric Chemistry and Physics*, 2011, **11**, 5945–5957.

21 J. L. Jimenez, *Journal of Geophysical Research*, 2003, **108**, 8425.

22 M. T. Timko *et al.*, *Journal of Engineering for Gas Turbines and Power*, 2010, **132**, 061505.

23 W. B. Knighton *et al.*, *Journal of Propulsion and Power*, 2007, **23**, 949–958.

24 M. T. Timko *et al.*, *Combustion Science and Technology*, 2011, **183**, 1039–1068.

25 M. Bente *et al.*, *Analytical Chemistry*, 2008, **80**, 8991–9004.

26 S. Gilardoni *et al.*, *Journal of Geophysical Research Atmospheres*, 2017, **34**, 401–409.

27 A. Kortelainen, *PhD thesis*, University of Eastern Finland, 2016.

28 M. Abegglen *et al.*, *Atmospheric Environment*, 2016, **134**, 181–197.

29 D. Delhaye *et al.*, *Journal of Aerosol Science*, 2017, **105**, 48–63.

30 J. Moldanová *et al.*, *Atmospheric Environment*, 2009, **43**, 38–44.

31 C. Giorio *et al.*, *Atmospheric Environment*, 2012, **61**, 316–326.

32 S. S. Lim *et al.*, *The Lancet*, 2012, **380**, 2224–2260.

33 V. Samburova, B. Zielinska and A. Khlystov, *Toxics*, 2017, **5**, 17.

34 R. Niranjan and A. K. Thakur, *Frontiers in Immunology*, 2017, **8**, 1–20.

35 T. Petry, P. Schmid and C. Schlatter, *Chemosphere*, 1996, **32**, 639–648.

36 D. A. Knopf, P. A. Alpert and B. Wang, *ACS Earth and Space Chemistry*, 2018, **2**, 168–202.

37 A. Faccinetto *et al.*, *Environmental Science and Technology*, 2015, **49**, 10510–10520.

38 R. Lemaire *et al.*, *Proceedings of the Combustion Institute*, 2009, **32**, 737–744.

39 Y. Bouvier *et al.*, *Proceedings of the Combustion Institute*, 2007, **31 I**, 841–849.

40 P. Parent *et al.*, *Carbon*, 2016, **101**, 86–100.

41 F. X. Ouf *et al.*, *Scientific Reports*, 2016, **6**, 1–12.

42 J. Yon *et al.*, *Combustion and Flame*, 2018, **190**, 441–453.

43 R. H. Moore *et al.*, *Aerosol Science and Technology*, 2014, **48**, 467–479.

44 R. Zimmermann *et al.*, *Environmental Science and Technology*, 2001, **35**, 1019–1030.

45 O. P. Haefliger and R. Zenobi, *Analytical chemistry*, 1998, **70**, 2660–2665.

46 K. Thomson *et al.*, *Applied Surface Science*, 2007, **253**, 6435–6441.

47 A. Faccinetto *et al.*, *Applied Physics A: Materials Science and Processing*, 2008, **92**, 969–974.

48 C. Mihesan *et al.*, *Chemical Physics Letters*, 2006, **423**, 407–412.

49 C. Mihesan *et al.*, *Journal of Physics: Condensed Matter*, 2008, **20**, 25221.

50 L. Sleno, *Journal of Mass Spectrometry*, 2012, **47**, 226–236.

51 C. A. Hughey *et al.*, *Analytical Chemistry*, 2001, **73**, 4676–4681.

52 A. G. Brenton and A. R. Godfrey, *Journal of the American Society for Mass Spectrometry*, 2010, **21**, 1821–1835.

53 R. Hilbig and R. Wallenstein, *Applied optics*, 1982, **21**, 913–917.

54 S. E. Stein and A. Fahr, *Journal of Physical Chemistry*, 1985, **89**, 3714–3725.

55 T. Adam, R. R. Baker and R. Zimmermann, *Journal of Agricultural and Food Chemistry*, 2007, **55**, 2055–2061.

56 Y. Tanaka, *Communications in Statistics - Theory and Methods*, 1988, 37–41.




This journal is © The Royal Society of Chemistry [year]

57 R. E. Peterson and B. J. Tyler, *Atmospheric Environment*, 2002, **36**, 6041–6049.

58 P. Cejnar *et al.*, *Rapid Communications in Mass Spectrometry*, 2018, **32**, 871–881.

59 T. Alexandrov, *BMC Bioinformatics*, 2012, **13**, S11.

60 L. Pei *et al.*, *Energy and Fuels*, 2008, **22**, 1059–1072.

61 P. Reitz *et al.*, *Journal of Aerosol Science*, 2016, **98**, 1–14.

62 R. Alvin C, *Methods of multivariate analysis - Second Edition*, Wiley - Interscience, 2001, pp. 1–727.

63 F. W. McLafferty and F. Tureek, *Interpretation of Mass Spectra*, University Science Books, Mill Valley, CA, 1993.

64 J. Pagels *et al.*, *Journal of Geophysical Research: Atmospheres*, 2013, **118**, 859–870.

65 J. Pagels *et al.*, *Journal of Aerosol Science*, 2003, **34**, 1043–1059.

66 T. R. Dallmann *et al.*, *Atmospheric Chemistry and Physics*, 2014, **14**, 7585–7599.

67 E. S. Cross *et al.*, *Journal of Engineering for Gas Turbines and Power*, 2012, **134**, 72801.

68 K. Aras, *Atmospheric Environment*, 1994, **28**, 1385–1391.

69 S. Gilardoni *et al.*, *Journal of Geophysical Research: Atmospheres*, 2007, **112**, 1–11.

70 J. Yon, A. Bescond and F.-X. Ouf, *Journal of Aerosol Science*, 2015, **87**, 28–37.