Unveiling trends in soot nucleation and growth: when secondary ion mass spectrometry meets statistical analysis

Author Affiliation. Cornelia Irimiea[1,2,†], Alessandro Faccinetto[1,*], Xavier Mercier[1], Ismael-Kenneth Ortega[3], Nicolas Nuns[4], Eric Therssen[1], Pascale Desgroux[1], Cristian Focsa[2]

[1]Univ. Lille, CNRS, UMR 8522 - PC2A - Laboratoire de Physico-Chimie des Processus de Combustion et de l'Atmosphère, 59655 Villeneuve d'Ascq, France.

[2]Univ. Lille, CNRS, UMR 8523 - PhLAM - Laboratoire de Physique des Lasers, Atomes et Molécules, 59655 Villeneuve d'Ascq, France.

[3]DMPE, ONERA, Univ. Paris Saclay, F-91123 Palaiseau, France.

[4]Univ. Lille, CNRS, UMR 2638, Institut M. E. Chevreul, F-59000 Lille, France.

[†]Currently affiliated to [3].

*Correspondence to: alessandro.faccinetto@univ-lille1.fr

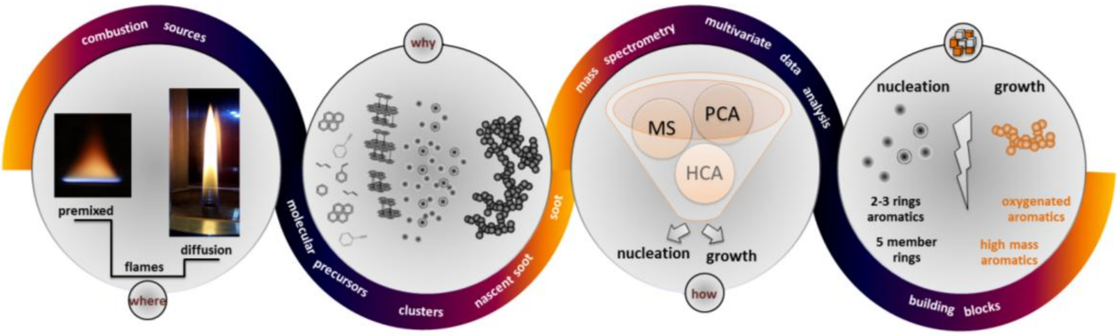
***Corresponding Author.** Alessandro Faccinetto

Contact information: Tel: +33 (0)3 20 43 49 85, mail: alessandro.faccinetto@univ-lille1.fr

Unveiling trends in soot nucleation and growth: when secondary ion mass spectrometry meets statistical analysis

Abstract (200 words max). At the present day, ultrafine soot particles have become the object of increasing attention due to their well-documented adverse effects on human health and climate. In particular, understanding soot nucleation is one of the most challenging problems toward a more controlled and cleaner combustion. Detailed information on the chemistry of nascent soot particles (NSPs) is expected to provide clues on the soot formation and growth reaction pathways. Herein, the early steps of soot formation in flames are addressed by investigating the chemical composition of NSPs and their molecular precursors by secondary ion mass spectrometry. The originality of this work lies on the combination of several factors that include: two different approaches to gain access to soot nucleation in premixed flames (analysis of a nucleation flame in which NSPs nucleate all along the reaction coordinate without growing in size) and in diffusion flames (separation of the precursors and soot regions); a sampling procedure that allows a rough separation of the condensable gas phase from soot particles; a large database of collected samples and high mass resolution that enable the use of an improved data reduction protocol based on hierarchical cluster analysis and principal component analysis for the data mining and interpretation.

Graphical abstract:



1. Introduction

Soot nucleation is a term commonly used in combustion research to describe the transition from molecular precursors in the gas phase to particulate matter in the condensed phase [1–5]. Soot nucleation is a wonderfully complex phenomenon that remains poorly understood to the present day. The detection of the molecular precursors that participate to soot formation is a major experimental challenge. In addition, the criteria used to identify soot particles in the condensed phase are often based on the experimental measurements used to identify them, thus over the years a variety of definitions partially overlapping or contradicting each other arose [6–11]. The emission of thermal radiation is one of the possible criteria to identify nascent soot particles (NSPs), which can be operationally defined as the smallest entities in the condensed phase capable of emitting radiation described by Planck's law (often cited as blackbody-like or graybody radiation) [8].Soot particles are generally at thermal equilibrium with the combustion gas, and therefore emit thermal radiation in a continuous spectrum that can be fitted with Planck's law. This emission can be used to identify soot particles even as their size approaches the limit of large organic molecules or small molecular clusters [12,13]. Hence, the yellow-orange luminosity of flames becomes an intuitive and effective criterion for assessing the presence of soot particles, including NSPs that can be as small as 1-3 nm diameter [6,14]. Although the detailed structure of NPSs remains essentially unknown, their small size and the generally amorphous nature of young soot tend to suggest a disordered rather than a core-shell internal nanostructure similar to mature soot primary particles [15–17].

Despite many recent advances, the formation of NSPs from the gas phase remains an elusive process. Mainstream hypotheses range from small aromatic molecules stacking into progressively larger compact [18–20] or 3D cross-linked [21,22] clusters. However, because of the intrinsic experimental difficulty in discerning the structure of NSPs, so far no conclusive evidence could be provided to completely support or disprove one model. On the molecular level, polycyclic aromatic hydrocarbons (PAHs) are long since well established [23], from experiments [24–27] and modeling [28,29], to be important molecular building blocks of soot. PAHs are found both in the gas phase as molecular precursors, and adsorbed on the surface or integrated as structural elements in the nascent particles [30,31]. Recent advances revealed for the first time the chemical structure of individual molecules found on NSPs [32]. Experimental evidences that oxygen-containing hydrocarbons (OCHs) are associated to soot have also been found [33], and thus they have been proposed as potential candidates for soot nucleation in addition to PAHs [34,35]. The molecular precursors, and more in general the compounds detected on freshly nucleated particles, typically feature high hydrogen to carbon ratio $H/C > 0.8$ [15,16,36]. Once soot particles are formed, the soot mass rapidly increases by surface growth, while the maturing primary particles coagulate and coalesce to form progressively larger aggregates characterized by complex fractal-like morphology [10,37]. In parallel, the H/C ratio of the compounds detected on the particles and aggregates surface gradually decreases to values $H/C < 0.4$ typical of mature soot [25,9].

From the experimental point of view, NSPs are difficult to detect and their properties are challenging to measure [26]. Commonly used diagnostics in combustion give access to a variety of information that range from the soot volume fraction [38,39] to the particle size distribution [6,40], structure and morphology [37,41]. The measurement of the mass of the compounds in the gas phase or adsorbed on the particle surface provides important chemical information on the species involved in soot formation and growth. As a result, countless implementations of mass spectrometry exist that give access to the chemistry of both the gas and the particulate phase [24,27,31,42–44]. In this work, we propose an innovative approach based on ex-situ secondary ion mass spectrometry (SIMS) to probe the chemistry of soot particles and condensable gas in parallel to obtain information on NSPs and their molecular precursors. To achieve such goal, a method to separate the condensable gas from the soot particles has been developed first [45]. Then, a large database of samples obtained in a variety of experimental conditions has been analyzed. It is well known that mass spectrometry returns large sets of data that have to be reduced into chemical information. Hence, data mining becomes fundamental to extract hidden information [46,47]. Furthermore, increasing the size of the database is arguably the most effective means of reducing and identifying sampling errors and obtaining a stable and reliable factor analysis representative of the total population [46]. The complementary information on the soot volume fraction profiles is obtained by in-situ laser induced incandescence (LII) mapping [12] that relies on the enhancement of the natural emission of soot particles in the flame by laser heating. Information on the spatial distribution of the aromatic precursors in the gas phase is obtained by in-situ laser induced fluorescence (LIF).

Isolating soot nucleation from other flame processes represents an additional challenge. In particular, in sooting flames the chemical reactions occurring at the surface of particles tend to build up as soon as NSPs are formed, and to quickly convert them into mature soot primary particles, de facto concealing the nucleation process. In this work, to avoid such interference and to identify independently the compounds involved in soot nucleation and growth, different approaches based on the analysis of two low-pressure premixed and two atmospheric diffusion laboratory methane flames are implemented. First, in the premixed nucleation flame, a dynamic equilibrium between the processes that increase the soot particle size (nucleation, surface growth, coagulation and coalescence) and those that decrease it (reversibility of nucleation, surface oxidation) is reached by carefully choosing the flame equivalence ratio, and results in the absence of any measurable increase in diameter of the NSPs against the reaction time [8,48,49]. Thus, the comparison of the nucleation flame with a second sooting flame stabilized in the same conditions represents a powerful tool to obtain information on the chemistry of NSPs. Second, in the diffusion flames the formation regions of molecular precursors and soot are separated and sufficiently far to be probed independently and with little interference from one another.

A total of 152 samples are collected from the four laboratory flames at different reaction times and analyzed by SIMS. Advanced statistical methods are employed to reduce this large database and reveal similarity and specificity. The comparison of soot sampled from the premixed and diffusion flames shows that PAHs and OCHs play a prominent role in soot nucleation. In particular, the smallest compounds involved in soot nucleation are consistent with PAHs having as few as 2-3 aromatic rings. Moreover, many strong signals come from 5-member ring PAHs, which further stresses the importance of these particular compounds. The important role of the post-nucleation surface growth is isolated in the premixed flames and leads to the definition of two distinct lists of compounds involved in the formation of NSPs and their growth. The obtained data are compared to kinetic modeling in the literature [50] showing that in the growth of both molecular precursors and compounds adsorbed on soot, compact PAHs are preferred over open structures.

2. Methodology

2.1. Burners and flames

2.1.1. Premixed flames

Two low-pressure laminar, premixed, methane/oxygen/nitrogen flames are investigated having equivalence ratio $\phi = 1.95$ (nucleation flame, PRE-N for short) and $\phi = 2.32$ (sooting flame, PRE-S for short). Flame PRE-N and PRE-S were already object of investigations in the past [51,14,8]. In particular, the soot volume fraction profiles of both flames are already known [8]. Pictures of the flames are shown in Fig. 1a and b, while the flow conditions are detailed in Table 1. The premixed flames are stabilized on a standard commercial Holthuis (previously McKenna) burner equipped with a 60 mm diameter sintered bronze porous plug at pressure $p = 267 \pm 1$ mbar and total gas flow $Q_{\text{tot}} = 3.96$ L min^{-1}. The burner is installed on a vertical linear translation stage located inside a low pressure vessel.

2.1.2. Sampling the premixed flames

A non-diluting quartz extractive microprobe enters the low-pressure vessel through the top flange to reach the flames axially. The thin end of the probe has a 0.3 mm diameter aperture that can be positioned in the flame with ±0.1 mm accuracy. The probed flame depth is estimated as twice the diameter of the probe aperture [52]. Changing the vertical position of the burner changes the relative distance between the burner surface and the probe, and enables positioning the tip of the probe at different height above the burner (HAB). Soot particles and combustion gas are pumped, undiluted, through the probe and a heated (80°C) stainless steel line to reduce condensation during the transfer to the sample holder. During the sampling, the pressure inside the line is approximately 0.5 mbar. Flame PRE-S is sampled at six HABs on the flame axis: before (7 mm) and at the beginning (9, 12 mm) of the soot incandescence zone, and up to the burned gas region (15, 20, 40 mm). Flame PRE-N is sampled at the same HABs.

2.1.3. Diffusion flames

Two atmospheric laminar diffusion methane flames are stabilized on a burner similar to the one used by Siegmann et al. [53,54]. The configuration of the burner and more details on the first flame (non-smoking flame, DIF-S for short) are given in Irimiea et al. [45]. A picture of flame DIF-S is shown in Fig. 1c. The second diffusion flame (quenched flame, DIF-Q for short), which features an additional quenching with argon in the reactive gas mixture, is used as a reference since soot volume fraction and mass spectrometry data at different locations in the flame axis are available in the literature [53,54]. Flame DIF-S is supplied with N5.0 grade methane only (Q_{fuel}), while flame DIF-Q is supplied with a mixture of N5.0 grade methane (Q_{fuel}) and argon (Q_{dil}) [53]. The flow conditions for the two diffusion flames are shown in Table 2.

2.1.4. LII/LIF measurements on flame DIF-S

In-situ diagnostics commonly used for the detection of soot and precursors include optical techniques as laser induced incandescence (LII) and laser induced fluorescence (LIF) that are used to measure the soot volume fraction and the size of the primary particles [12,55,56] and to map the molecular precursors in the gas phase [57], respectively. In this work, flame DIF-S has been subject to detailed in-situ investigations since no information on the soot volume fraction or on the chemical composition of the gas phase was available beforehand. Soot is detected by LII at 1064 nm, while its precursors are detected by LIF at 488 nm and 266 nm excitation wavelength. The fundamental of a pulsed Nd:YAG laser (Quantel Brilliant B, 6 ns pulse width, 10 Hz) equipped with a frequency quadrupling module and an optical parametric oscillator (Panther-EX OPO Continuum) pumped by the 3rd harmonic of a Nd:YAG laser (Powerlight 8000 Continuum, 12 ns pulse width, 10 Hz) are used. Soot particles are heated below their sublimation threshold (F = 150 mJ cm^{-2}) and precursors in the gas phase are excited without activating the incandescence of soot (F = 7.7 mJ cm^{-2} at 488 nm and F = 2.2 mJ cm^{-2} at 266 nm). The LII/LIF emission is recorded at 90° incidence with respect to the laser beam propagation, using a system of two achromatic lenses for the imaging of the selected volume on the entrance of a spectrometer equipped with a 150 groves mm^{-1} grating blazed at 300 nm. A fast ICCD camera (Princeton Instruments 500PI MAX) placed at the exit of the spectrometer records an image corresponding to 500 laser pulses. The LII/LIF detection is performed during a 50 ns gate after the peak of the laser pulse. The spectral response of the detection system is calibrated in intensity with an integration sphere (SphereOptics CSTM-LR-6-M), while the wavelength calibration is performed with an Hg-Ar pen-ray line source. The LII signal is integrated over a 5 nm wavelength window centered at 750 nm, while the Vis-LIF and UV-LIF signals are integrated over the same spectral window centered at 550 nm. The LII and LIF signals are recorded at different HAB by translating the flame on the vertical axis. The emission of the flame is subtracted from each LII/LIF image, and the LII/LIF signal is normalized to the maximum detected value. It is therefore possible to localize and distinguish between the gas phase precursors and soot particles. A schematic representation of the setup is shown in Fig. 2a, while Fig. 2b shows the evolution of the

LII/LIF against HAB. In flame DIF-S, the UV-LIF signal is detected very low in the flame from around 5 mm HAB, it increases up to a maximum at 50 mm HAB, and finally disappears above 70 mm HAB. The Vis-LIF signal starts at 20 mm, reaches a maximum at 65 mm HAB and completely disappears at 85 mm HAB. The LII signal starts at 62 mm HAB, very close to the maximum of the Vis-LIF, reaches a maximum at 80 mm HAB, then quickly decreases and disappears around 105 mm HAB. Both diffusion flames are non-smoking.

2.1.5. Sampling the diffusion flames

Flame DIF-S is characterized by LII/LIF herein to select the optimal HABs for the sampling, while flame DIF-Q is used as a reference since the soot profile and mass spectrometry data at different locations in the flame axis are available in the literature [53,54]. The extraction system is detailed in [45]. Briefly, it consists of a sampling probe having 0.25 mm diameter aperture coupled to an automatic pressure regulation system inserted radially into the flame, engineered to combine high dilution ratio up to 3×10^4 [40] while minimizing flame perturbations [58]. The diffusion flames are sampled at three HABs each in the flame axis, at the maximum of the UV fluorescence detected by LIF (48 mm for flame DIF-S and 25 mm for flame DIF-Q), at the beginning of the soot incandescence zone (65 mm for flame DIF-S and 45 mm for flame DIF-Q), and near the maximum of the soot incandescence zone in a region with weak signal of fluorescent species excited by visible wavelength (80 mm for flame DIF-S and 55 mm for flame DIF-Q). The large soot volume fraction of the diffusion flames requires dilution during sampling to quench chemical reactions and limit particle coagulation and coalescence.

2.2. Preparation of the samples and SIMS measurements

Online and ex-situ diagnostics often depend on mass spectrometry to gather chemical information from the samples. Online techniques relying on the generation and analysis of a molecular beam have been available for almost two decades now [42,59]. Their modern implementations, like aerosol mass spectrometry for instance, are now compact and commercially available devices that can provide semi-quantitative measurements in real time and therefore have become widespread for the analysis of environmental samples [60–64]. The applications of surface techniques to the ex-situ analysis of soot are too many to describe herein, and range from thermal [65,66] to laser desorption mass spectrometry [67], possibly after solvent extraction [43] and on a variety of deposition substrates [27,44,68]. SIMS has been independently used for the analysis of the chemical composition of aerosols [69,70] and PAH containing samples [71], although it never was the technique of choice for systematic investigations of the chemical composition of soot particles.

In this work, samples obtained by impaction of soot and condensable gas on silicon and titanium wafers are characterized using the ToF.SIMS[5] instrument available at the Regional Platform of Surface Analysis at the University of Lille. A detailed description of the adopted analytical protocol is provided in Irimiea et al. [45]. SIMS is a high sensitivity analytical technique used to determine the chemical composition of unknown samples that can probe a surface layer thinner than 1 nm. The sample is sputtered with a primary ion beam and the ejected secondary ions are mass-analyzed with a time of flight mass spectrometer having maximum resolving power ~ 10^4. The distinction between soot and the condensable gas phase is obtained by following the protocol proposed in Irimiea et al. [45]. During impaction deposition, particles deposit preferentially at the outlet of the impactor, while condensable gas diffuse all over the wafer surface, including the already deposited soot. The mapping of the sample surface and the comparison of the mass spectra of the soot impacting region and surroundings by image reconstruction can be exploited to get information on the chemical composition of the two phases.

2.3. HCA and PCA in a nutshell

Mass spectra of soot contain a significant amount of chemical information that is difficult to extract because of the large number of signals they typically feature [72]. As high resolving power mass spectrometers get easier and easier to access, advanced statistical tools have become invaluable to refine large databases of raw mass spectra and to reduce them into chemical information. Principal component analysis (PCA) is used to classify the data and explain the variance of the database. Since PCA is based only on a fraction of the total variance, its output can be affected by the presence of outliers. Therefore, to increase the confidence level of the data reduction, hierarchical clustering analysis (HCA), which is based on the total variance of the dataset, is applied in parallel following the approach proposed by Pei et al. [73].

HCA attempts to build a hierarchy of clusters of observations based on their similarity. A measure of the similarity between clusters and a set of linkage criteria are required to decide which clusters should be combined, and can be provided in a variety of different ways conceptually based on the calculation of the distance between pairs of observations. For instance, a common approach is based on Euclidian distances between observations, then HCA is computed with the group average method. The average linkage criterion is commonly used because is space-conserving and delivers the most robust data hierarchy after using a multitude of algorithms [73,74]. The results of HCA are organized in distinctive-looking tree diagrams (dendrograms) that show the distances at which the data clusters are merged.

PCA is a mathematical tool used in exploratory data analysis to identify relatedness among populations [75,76]. In PCA, an original set of possibly correlated variables is transformed into a second set of linearly uncorrelated variables (principal components) that retains the maximal amount of information of the original data. PCA is performed on a data correlation or covariance matrix usually after a normalization of the original data to zero average and unit variance. The normalization is a particularly delicate step since PCA only uses linear transformations, and therefore PCA on a non-normalized dataset would be dominated by the largest numeric values. The percentage of the total variance explained by each principal component is stored in the eigenvalues and visualized in the *scree plot* (eigenvalues against principal components). Essentially, the scree plot orders the eigenvalues from the largest to the smallest, and it is often used to select the number of principal components to be further considered. The relative weights of each original variable used to calculate the principal components are stored in the eigenvectors and visualized in the *loading plots* (weights against original variables). The loading plots are fundamental to quickly identify which variables affect each principal component. The loadings can take negative or positive value: the more a variable influences one of the principal components, the larger the corresponding loading in absolute value. Conversely, the less a variable influences one of the principal components, the closer the corresponding loading is to zero. The scores are linear combinations of the data calculated from the coefficients for each principal component, and represent the distance of an observation from a principal component. There is one score for each observation in the data set and for each principal component. A *score plot* shows the scores of two principal components one against the other, and can be used as a base to classify the data. For instance, if the first two components account for the largest part of the variance of the database, the PC2 against PC1 score plot is generally sufficient to identify data clusters and trends.

In this work, HCA and PCA are implemented on the normalized integrated peaks outcome of the mass spectra. The peak ion counts are normalized by the total ion count of each spectrum after background removal. Then, they are organized in a matrix containing the observation (rows) by the peak ion counts (columns). Finally, HCA (group average clustering method, Euclidean distances) and PCA (covariance matrix) are implemented on the data matrix. Routines in OriginPRO 2016 (OriginLab, Northampton, U.S.A.), Matlab (MathWorks, Natick, U.S.A.) and the freeware ToFTools have been used for the peak deconvolution and computation of the statistical methods.

3. Results and discussion

3.1. Understanding the mass spectra

A selection of the recorded mass spectra sorted by flame (rows) and sampling HAB (columns) is shown in Fig. 3. The mass spectra of the condensable gas phase and soot are shown in the top and bottom panels, respectively. The mass spectra feature a quasi-bimodal distribution of peaks separated by a gap around *m*/*z* 150-200. The first mode spans *m*/*z* 0-150 and its strongest signals are attributed to fragment ions as discussed below. This mode tends to be more prominent in the condensable gas sampled at high HAB, and in flame PRE-N at any HAB. The second mode roughly spans *m*/*z* 150-800 and features signals mainly attributed to $C_mH_n^+$ and $C_mH_nO_p^+$ molecular ions. The second mode is more prominent in the diffusion than the premixed flames. The relative intensity of the peaks and the position of the modes in the *m*/*z* axis are both flame and HAB dependent. Within the modes, the signal peaks always recur in groups corresponding to the increasing number of carbon atoms m. Each group contains several peaks corresponding to ions having a progressively increasing number of hydrogen atoms n. Neighboring groups are separated by gaps in which weaker but still well resolved peaks are detected. The gaps become progressively smaller as *m*/*z* increases, and disappear from *m*/*z* 700 upwards.

The analytical protocol adopted in our previous work and based on soot and condensable gas deposition by impaction [45] results in a mass resolution ($m/\Delta m \sim 7000$) large enough to enable using mass defect analysis for the identification of unknown molecular ions. Despite being richer in low mass signals, the SIMS mass spectra well match our previous two-step laser mass spectrometry (L2MS) measurements [27] and early tests with SIMS where deposition on filters was used [8]. The strongest peak in a group sharing the same m depends on the parity of m. If m is even, then the strongest peaks have even nominal mass, while if m is odd, then the strongest peaks have odd nominal mass. This behavior is consistent with the attribution of the signal peaks to PAH ions. In fact, odd m PAHs contain (at least) one non-aromatic carbon atom, and therefore post-ionization hydrogen elimination reaction result in increased stability by extended electron resonance.

Flame DIF-Q is similar to the work of Siegmann et al. [53,54] and provides a comparison with the literature. In this work, a significant extension of the upper mass limit could be obtained in a range important for the nucleation and growth of NSPs. Although the peak relative intensities are significantly different in the two works, the main mass sequences attributed to benzenoid PAHs are well consistent. The peak relative intensities are characteristic of the ionization scheme. In Siegmann et al. [54], resonant enhanced multiphoton ionization (REMPI) is used that enables the efficient ionization and hence a highly selective detection of resonant compounds. In particular, ionization at 248 nm excitation wavelength is very effective for the detection of PAHs in the range *m*/*z* 200-300 [77], and the obtained distribution of peaks is centered around *m*/*z* 200 consistently with REMPI of low mass PAHs. Furthermore, the smooth energy transfer typical of low irradiance laser ionization greatly limits post-ionization dissociation, resulting in almost fragment free mass spectra. This is a major difference with SIMS, where the harsher ionization produces fragment-rich mass spectra.

In this work, peaks are identified and assigned a molecular formula based on their mass defect [78]. A working peak list containing 771 peaks is generated following the protocol established in Irimiea et al. [45]. Only peaks having signal–to-noise ratio (SNR) larger than three times the standard deviation of the background noise are included in the data reduction. One mass spectrum and the corresponding mass defect plot are shown as examples in Fig. 4a and b, respectively. High mass defect ions are identified as $C_mH_n^+$ from their typical pattern and assigned a molecular formula. The difference between the measured accurate mass and the exact mass of the assigned molecular formula is typically in the 0.1-5 ppm range. A large number of weaker signals characterized by lower mass defect are also detected and attributed to OCHs .

SIMS analyses are often affected by post-ionization dissociation reactions that generate a potentially large number of fragment ions and increase the complexity of the mass spectra. Fragment ions should be identified and treated independently of the molecular ions. To this end, hierarchical clustering analysis (HCA, group average clustering method, Euclidean distances) is used to find general trends and similarities in the behavior of the peak signals, and to search correlations between individual peaks and/or groups of peaks. HCA found three main clusters of peaks that share close behavior that are shown in different colors in Fig. 4a and b. Cluster 1 (black data) only contains low mass defect ions that are easily identified thanks to their distinctive pattern in the mass defect plot. Their mass defect is too low to be consistent with molecular ions (strongest signals are C_mH^+, $C_mH_2^+$ and $C_mH_3^+$, $m \in [1, 16]$). Cluster 2 (red data) contains higher mass defect ions that have mass consistent with unsaturated and aromatic compounds. It is worth to notice that many small PAHs from $C_9H_8^+$ up to $C_{24}H_{16}^+$ are in cluster 2. Finally, cluster 3 (green data) contains signals only attributed by mass defect analysis to large PAHs and OCHs. The three clusters coexist in *m*/*z* 200-300. At lower *m*/*z*, clusters 1 and 2 dominate, while the vast majority of signals at higher *m*/*z* are part of cluster 3. The peak intensity scatter matrix gives further information on the correlation coefficients (Pearson's R^2) calculated on the normalized peak intensities. The peaks classified by HCA in the same cluster are highly correlated to each other ($R > 0.8$). The peaks classified in clusters 1 and 2 are generally anti-correlated to those classified in cluster 3, but feature much weaker correlation coefficients ($-0.7 < R < 0.1$) than the previous case. In particular, it is interesting to notice that the anti-correlation of clusters 1 and 3, together with the very low mass defect of the peaks in cluster 1, can be explained in terms of post-ionization dissociation reactions. Dissociation reactions remove molecular ions from the ion plume, therefore the higher the counts of fragment ions (cluster 1) and the lower the counts of precursor ions (cluster 3), and vice versa. The negative correlation between clusters 2 and 3 is more difficult to explain as it does not seem to be related to the analysis process, and likely have a more fundamental explanation (for instance, small hydrocarbons in cluster 2 might be depleted to generate the larger species in cluster 3). In any case, this result greatly strengthens the hypothesis that different compounds are involved in different processes, and that advanced statistical analysis is necessary to classify the different behaviors in view of finding a detailed explanation.

3.2. From peaks to molecules

Once attributed a molecular formula based on their mass defect, $C_mH_n^+$ ions can be represented as points of coordinates (n/m, *m*/*z*) in a two-dimensional space. The resulting graphs are an adaptation of the well known C-H diagrams [79], and are especially useful in this context as they provide the space of all possible combinations of (n, m) and an immediate, visual feedback on the molecular formula of all detected ions. Furthermore, C-H diagrams are a convenient way to compare experimental results with the predictions of kinetic modeling [80]. Fig. 5 shows the data reduced from the mass spectra in Fig. 3 sorted by flame and sampling HAB. The n/m ratio roughly spans the range [0.2, 0.8], clearly decreasing as *m*/*z* increases. The molecular formulae $C_mH_n^+$ obtained from the mass defect analysis are consistent with PAHs that contain 5- and 6- member rings and side-substituted PAHs. In Fig. 5, the PAH distributions are compared with the model proposed by D'Anna et al. [50] to follow the molecular growth in terms of C and H atom numbers. The black dashed lines represent the limits of benzene oligomers (top line, high n/m) and pericondensed PAHs (bottom line, low n/m). In all flames, the identified $C_mH_n^+$ ions are found scattered between the benzene oligomers and the pericondensed PAHs limits, generally much closer to the latter. The few data points found below the pericondensed PAHs limit are possibly fragment ions not identified by HCA as discussed above, or alternatively substituted PAHs or pericondensed PAHs with additional 5-member rings on a zigzag edge. Soot sampled from flame PRE-N only shows signals up to *m/*z 400-450, while soot sampled from the other three flames shows signals up to *m/*z 850-900.

The position in a mass spectrum at which the ion SNR falls below 3 is denoted as $(m/z)_{max}$ and is a convenient variable to monitor the mass sequence. As shown in Fig. 6, $(m/z)_{max}$ depends on the flame type, sampling HAB and analyzed zone (condensed gas phase or soot). In Fig. 6, the filled dots represent soot while empty dots represent the condensable gas phase obtained in the absence of soot (for instance, flames PRE-S at HAB ≤ 9 mm or PRE-N at any HAB) or by image reconstruction whenever soot coexists with the gas phase (for instance, flames DIF-S and DIF-Q at 65 and 45 mm HAB, respectively).

In flame PRE-N (nucleation flame) no differences are observed between condensable gas and soot samples, and additionally little to no evolution against HAB is observed as shown in Fig. 6a. This absence of net molecular growth is consistent with a soot formation mechanism driven by the generation of NSPs all along the flame reaction coordinate and by the virtual absence of soot particle size increase [8]. The existence in flame PRE-N of PAHs up to *m*/*z* 400-450 in parallel to NSPs detected by LII further supports the hypothesis that soot nucleation in premixed flames does not require large hydrocarbons to occur. On the other hand, $(m/z)_{max}$ obtained from the analysis of soot and condensable gas in flame PRE-S greatly differ. As shown in Fig. 6b, $(m/z)_{max}$ of the condensable gas features a local maximum at 12-15 mm HAB that corresponds to the beginning of the soot nucleation zone determined by LII (gray highlighted region). Then, $(m/z)_{max}$ progressively decreases as the HAB further increases down to the same value of flame PRE-N at 40 mm HAB. $(m/z)_{max}$ of soot starts at *m*/*z* 850-900 at the beginning of the nucleation zone, then progressively decreases against HAB. The diffusion flames reveal a different evolution of $(m/z)_{max}$ than the premixed flames as shown in Fig. 6c and d. Signals up to *m*/*z* 850-900 are identified upstream the soot inception zone in both diffusion flames from very low HAB. At higher HAB, $(m/z)_{\max}$ monotonically decreases, faster in the condensable gas than in soot.

3.3. Disclosing the chemistry of soot nucleation and growth

HCA and PCA are used to find correlations between subsets of the total sample population. HCA and PCA are initially used to refine the complete dataset of 152 samples, and in particular to identify sampling and analysis artefacts. A detailed discussion can be found in the Supporting Information (*SI. Local factors affecting the structure of the database*), while herein a reduced dataset of 101 samples obtained after removal of the identified artefacts is discussed.

Fig. 7a and b compare the HCA dendrogram to the PC2 vs. PC1 score plot obtained after removing the identified fragment ions. The loading plots of PC1 and PC2 are shown in Fig. 7c and d, respectively. The loadings correspond to signals attributed by mass defect analysis to PAHs and OCHs, and are represented by black and red data, respectively. To provide an immediate visual feedback, the data clusters identified by HCA are color coded to match the groups of samples in the PCA score plot. It is significant that the cluster structure identified by HCA and PCA are consistent and all data clusters are well separated and appear adjacent to each others in the score plot. The correspondence of the data clusters with the samples is further elaborated in Table 3. In the score plot and dendrogram, samples from different flames and HABs are most of the times classified in independent data clusters. For instance, data clusters 1, 2 and 8 only contain samples from the diffusion flames, both condensable gas at low HAB (data cluster 1), and young and mature soot at higher HAB (data clusters 2 and 8). Data clusters 3, 4 and 7 contain soot samples from flame PRE-S well separated by HAB. Data cluster 5 contains all the samples characterized by high negative loadings of PC1 in the low *m*/*z* region, and low negative loadings in the high *m*/*z* region of PC2 as shown in Fig. 7c and d. Data cluster 5 notably includes the majority of samples from flame PRE-N, the condensable gas from the tip of the diffusion flames and all samples from flame PRE-S at 40 mm HAB despite originating from very different combustion conditions. Finally, data cluster 6 contains all the samples characterized by high negative loadings of PC1 in the low *m*/*z* region, and positive loadings in the low *m*/*z* region of PC2 as shown in Fig. 7c and d. These samples include all the condensable gas from flame PRE-S except 40 mm HAB, and flame PRE-N at 40 mm HAB. Two samples from flame PRE-N at 15 mm HAB are also in this data cluster.

PC1 (52.6%) and PC2 (23.5%) are by far the most interesting principal components of PCA that together explain the 76.1% of the variance of the database. The lower variance principal components (PC3, PC4, etc.) are not considered in the discussion. A sharp change of sign of PC1 occurs at *m*/*z* 165–245. PC1 > 0 contains signals attributed to PAHs and OCHs, while PC1 < 0 contains signals attributed to low mass PAHs. PC2 shows a much slower change of sign in *m*/*z* 294–450. The loadings on PC2 > 0 are much larger than PC2 < 0 and mostly contain signals attributed to PAHs, while the loadings on PC2 < 0 are closed to zero and mainly attributed to OCHs. The highest loading of PC1 > 0 is *m*/*z* 239.086 ($C_{19}H_{11}^+$), and the highest loading of PC1 < 0 is *m*/*z* 115.055 ($C_9H_7^+$). In the ionization conditions the experiments are performed, post-ionization H-elimination reactions of more than two hydrogen atoms are never observed during the analysis of analytical standards (pure PAHs), and therefore these signals are attributed to the product of H-elimination from $C_{19}H_{12}^+$ and $C_9H_8^+$, respectively. Many similar cases exist in the mass spectra. It is worth to notice that odd numbers of carbon atoms are compatible with polyaromatic structures containing 5-member ring(s), at least one non-aromatic carbon atom, or both. The existence of such signal as base peak is remarkable given the general higher thermodynamic stability of benzenoid PAHs [23], and suggests that structures containing 5-member rings or non-aromatic carbon atoms play an important role on soot formation and growth and are stable enough to be found in the gas phase or adsorbed on young soot particles at potentially large concentration. This has been very recently confirmed by a combined experimental and theoretical study [81]. Whether chemical species that are removed from the flame still carry information on the flame processes remains a central problem of the data interpretation. The extraction procedure in ex-situ diagnostics causes irreversible modifications to the samples. Recombination reactions result in the loss of all information on reactive species (radicals, ions) as soon as they are extracted from the flame, and furthermore volatile compounds partially desorb in the mass spectrometer during the creation of the vacuum before the mass analysis can take place. On the other hand, by applying high dilution factors during sampling, extensive post-sampling recombination can be avoided. Regardless, the good agreement of the sequence of hydrocarbons detected on any sampling conditions with the prediction of kinetic modeling is a strong evidence that at least the information on the large hydrocarbons, in the gas phase or adsorbed at the particle surface, is preserved, and this information can be linked to the composition of the particulate matter (NSPs and primary soot particles).

To help explaining the results of PCA it is generally helpful to add as much information as it is available to the score plots. Fig. 8 shows the PC2 vs. PC1 score plot with data classified by flame and sampling HAB. The data point color and shape represent the sampled flame, PRE-N (red rounds), PRE-S (violet stars), DIF-S (green squares) and DIF-Q (blue triangles) while the data point interior represents soot (filled dots) and condensable gas (empty dots). PC1 and PC2 are the most important principal components that explain the 52.6% and the 23.5% of the total variance of the database, respectively:

- PC1 < 0 (*m*/*z* 93-245) identifies samples that contain very small carbon-based clusters and/or nanoparticles in the gas phase (Fig. 7). Notably, such entities include NSPs from flame PRE-N resulting from the nucleation process and having constant size all along the flame axis. Furthermore, PC1 < 0 contains soot nanoparticles resulting from the re-oxidation of larger soot aggregates occurring at high HAB in the diffusion flames (80 mm in DIF-S and 55 mm in DIF-Q, both diffusion flames are non-smoking).
- PC1 > 0 (*m*/*z* 165-731) is characterized by high loadings of PAHs with more than three aromatic rings (Fig. 7), and its most remarkable feature is being clearly linked to the analysis of visible soot deposits and to the low HAB samples from the diffusion flames (48 mm in DIF-S and 25 mm in DIF-Q in Fig. 8). The high loadings of PC1 > 0 point to the existence of large PAHs upstream the soot nucleation region in the diffusion flames, but not in the premixed flames. The high scores obtained for the gas phase in flames DIF-S and DIF-Q confirms the predominance of aromatic compounds with mass in the range *m*/*z* 165-731 well below the soot inception zone. The lower scores on PC1 obtained for samples collected in a sooting region imply that these compounds are consumed or converted as the HAB increases. This phenomenon is reversed in flame PRE-S, in which high mass compounds are found on soot rather than in the gas phase.
- PC2 < 0 (*m*/*z* 294-731) contains all samples of flame PRE-N except 40 mm HAB and all the condensable gas from the tip of the other three flames at high HAB (PC1 < 0), and samples not clearly clustered of flames PRE-S, DIF-S and DIF-Q (PC1 > 0). The very small loadings of PC2 < 0 (Fig. 7) indicate that no statistically

significant differences exist between samples with high score in PC2 < 0, and that the corresponding variables are not involved in the phenomenon described by PC2. This is particularly interesting for data clusters 2 and 5 that contain samples obtained from several flames. In Table 4, the structure of data cluster 2 suggests that soot obtained from the tip of the two diffusion flames has virtually the same composition despite the different flame height. However, data cluster 5 contains a larger variety of samples having in common very low concentration of high mass PAHs. As shown in Table 4, these samples include flame PRE-N at all HAB except 40 mm, flame PRE-S at 40 mm, and all condensable gas sampled from the tip of the diffusion flames.

- PC2 > 0 (*m*/*z* 90-450) contains well separated condensable gas samples of flame PRE-N at 40 mm HAB and all the condensable gas of flame PRE-S (PC1 < 0), and again samples not clearly clustered from flames PRE-S, DIF-S and DIF-Q (PC1 > 0). It is interesting to notice how a slight increase of the peak intensity in the mass spectra collected in flame PRE-N at 40 mm HAB, but not of $(m/z)_{max}$, ultimately results in this HAB to be the only one to appear in PC2 > 0. Among other things, this shows that PCA is sensitive to small changes of the peak intensity.

The data collected at all HABs except 40 mm in flame PRE-N coexist in the same data cluster, i.e. no statistically significant differences in the mass spectra can be observed at different HABs. We believe this to be a potentially important finding. Flame PRE-N is characterized by the existence of NSPs having nearly constant diameter in the range of 1-3 nm all along the reaction coordinate as already established by in-situ LII investigations [14,48,51] and supported by kinetic modeling [49]. The lack of particle size growth and the very slow evolution of the soot chemical composition as shown by the SIMS analysis strongly point toward the formation of a steady state for both the particle size and their chemical composition. After the particle formation and initial growth, the surface oxidation processes become progressively more important as the particle grows in size. At some point, the overall oxidation matches the same rate as the overall growth, and the particle diameter ceases to increase. After the steady state is reached, a hypothetical particle heavily affected by surface oxidation would still be indistinguishable from a freshly nucleated particle with the currently available diagnostics, or, in other words, it is not possible to find a direct link between the reaction time and the particle size. Therefore, we speculate that in the premixed flames the change of sign of PC1 is related to the transition from a regime dominated by the nucleation of NSPs to a regime dominated by soot surface growth. The high loadings of PC1 < 0 (see Fig. 7) suggest that the composition of NSPs mainly consists of small hydrocarbons having molecular mass as low as *m*/*z* 90.047 ($C_7H_6^+$). Therefore, if this hypothesis is correct, by following the change of sign of PC1 it is possible to mark the beginning of the sequence of PAHs mainly involved in soot surface growth that eventually result in the soot deposits visible even to the naked eye. The complete peak lists with corresponding attributions can be found in the Supporting Information.

Globally, Fig. 8 shows that the same classes of molecular precursors, although spanning different mass ranges, provide the building blocks of soot in different combustion environments and the final product is remarkably homogeneous from a statistical point of view (PCA scores). The molecular precursors involved in the formation of NSPs show a remarkably similar chemical composition in both premixed flames characterized by a relatively narrow distribution of small PAHs. For the diffusion flames, the pool of PAHs is broadened toward higher masses. However, in the flames characterized by prominent particle size growth (PRE-S, DIF-S and DIF-Q) the detected mass range is again expanded to a significantly higher value in the young soot and mature soot regions. The explanation of this enlargement is still not clear, and our current observations do not allow ascertaining if these high mass compounds belong exclusively to the gas phase. These masses could either correspond to large PAHs formed during the PAHs growth processes or to small molecular clusters of medium size PAHs potentially participating to the soot nucleation. Experimentally, the latter are expected to be difficult to detect by in-situ techniques, and from a fluid mechanics point of view they are extremely hard to separate from the gas phase for ex-situ diagnostics. However, this step of molecular clustering could be a crucial one for understanding the transition from the gas to the condensed phase, and therefore each chemical insight we can bring has paramount importance. To this purpose, the role of PAH dimerization in the formation of NSPs is worth of particular attention being a currently much debated topic [82,83,33]. The formation of Van der Waals dimers in flames is indeed questionable because of the very weak bond energies, which do not theoretically enable them to survive at the flame temperature. However, different authors regard as physically realistic the formation of dimers in flames [5,83,84]. For instance, by using a semi-empirical molecular dynamic approach, Schuetz et al. [84] showed that dimers of pyrene can be stabilized by the development of free internal rotors in the colliding PAHs. More recently, Kholghy et al. [83] proposed a model in which soot nucleation is not completely reversible and involves a stabilization pathway of the dimers by the formation of covalent bonds through a sequence of hydrogen abstraction and carbon-carbon bond formation. In this last work, which is also based on data obtained from flame PRE-N, the authors conclude that only the formation of dimers of small PAHs can provide soot quantities comparable to the available measurements. This conclusion is therefore in excellent agreement with the results of HCA and PCA in our present work that clearly highlights the prominence of small PAHs (PC1 < 0 in Fig. 7, *m*/*z* 93-245) in the nucleation zone before the appearance of NSPs. Hence, if dimers were produced in this flame, they would necessarily originate from small PAHs as suggested by our treatment. However, more efforts are still to be made to lead to the precise identification of these species.

3.4. Oxygen-containing hydrocarbons

The mass defect analysis allows attributing a $C_mH_n^+$ molecular formula with incertitude below 5 ppm to around half the total number of detected peaks. The remaining peaks are assigned to OCHs based on their low mass defect that places them below the pericondensed PAH limit in the mass defect plot in Fig. 4b, and based on their isotopic patterns that systematically rule out low H/C fragment ions. A molecular formula $C_mH_nO_p^+$ is not attributed to the individual peaks for the following reasons. Contrary to $C_mH_n^+$ ions, the peaks attributed to OCHs are often asymmetrical, suggesting that more than one ion contributes to the peak structure. In turn, the incertitude on the assignment of a molecular formula can be as high as 15 ppm. In many cases the mass resolution is not deemed sufficient for attempting a full peak deconvolution, and therefore only the accurate mass of the mode is selected for the mass defect analysis. Even in the most favorable cases of well resolved and symmetrical peaks, univocally attributing a $C_mH_nO_p^+$ molecular formula is not always possible with the available resolving power. For instance, $C_{m-4}H_nO_3^+$ and $C_{m-7}H_{n+4}O_5^+$ require mass resolution as high as 25000 and 70000, respectively, to be resolved from $C_mH_n^+$ at *m*/*z* 400.

As shown in the loading plots in Fig. 7c and d, OCHs (red data) are mostly found associated to PC1 > 0 and PC2 < 0, and correspond to soot deposits and low HAB samples from the diffusion flames as already discussed above. However, their loadings in PC2 < 0 are small enough that they do not contribute significantly to the (unknown) phenomenon described by PC2. In general, OCHs closely follow the behavior of PAHs but are characterized by much lower ion counts in all samples and in particular in condensable gas. Premixed and diffusion flames once again behave differently. In the premixed flames, OCHs are mostly associated to the soot growth but the nucleation of NSPs occurs at lower HAB and only involves the smaller PAHs identified in PC1 < 0. On the other hand, in the diffusion flames a large amount of OCHs is found already in the gas phase before soot nucleation as well as associated to NSPs. These findings are consistent with the works of Cain et al. [9] and Commodo et al. [85] where it is shown that NSPs in diffusion flames contain large amounts of oxygen.

The origin of OCHs is still being debated [34,35,9,85]. Although some works suggest that data obtained with ex-situ techniques are subject to post-sampling oxidation [9], a recent study points to the marginal role of the oxidation during sampling compared to the combustion process [86]. In addition, both simulations [28,33] and experiments [34,85] show that high mass OCHs potentially play an important role on soot nucleation. However, even if the presence of OCHs represented evidence of post-sampling surface oxidation, a recent study proves that OCHs are less likely to condense on soot particles [33] than PAHs, and therefore those found on soot are likely to have taken part to soot formation at some point.

4. Summary and conclusions

Detailed information on the chemistry of nascent soot particles (NSPs) is expected to provide clues on the soot formation and growth reaction pathways. The structure of many laboratory flames is nowadays well understood, and this knowledge makes them very desirable for both experimentalists and modelers aiming to investigate fundamental processes. Herein, we discuss the data obtained from the secondary ion mass spectrometry (SIMS) analysis of samples extracted from four different laboratory flames, two low-pressure premixed and two atmospheric diffusion methane flames. The premixed flames are stabilized in conditions that result in the soot nucleation beginning at the same reaction time, or equivalently at the same height above the burner (HAB). However, in the premixed sooting flame (PRE-S) soot primary particles are formed at short reaction time, while in the premixed low-pressure nucleation flame (PRE-N) the equivalence ratio is chosen so that NSPs nucleate all along the flame without growing in size [8,49]. To complete the set of investigated flames, a sooting diffusion flame (DIF-S) and a diffusion flame quenched with argon (DIF-Q) are stabilized in similar aerodynamic conditions. In contrast to the previous case, in which soot and molecular precursors coexist all along the reaction coordinate, the laser induced incandescence/fluorescence (LII/LIF) mapping of the diffusion flames shows a net stratification of the formation regions of soot and molecular precursors that can be probed separately. Flames PRE-N, PRE-S [8] and DIF-Q [53,54] were already subject to ex-situ mass spectrometry investigations in the past. The much larger database of collected samples and the higher mass resolution enable here the use of an improved data reduction protocol based on multivariate analysis for the data mining and interpretation.

A database containing 152 samples is first analyzed to assess how the flame chemistry influences the structure of the mass spectra. Hierarchical cluster analysis (HCA) and principal component analysis (PCA) are used to classify the samples and find correspondences between sample populations. PCA shows the changes in the total variance of a group of observables (the samples) calculated from a set of variables (the peak ion counts). The loadings plots identify the variables that have the largest effect on each principal component. Therefore, if a principal component can be associated to a physical phenomenon, then PCA suggests a set of variables this phenomenon depends on. This additional information is provided by LII/LIF that maps the formation regions of soot and molecular precursors, and by the presence of visible soot deposits on the sample surface.

The presence in the mass spectra of a sequence of high mass species is well represented by the first principal component (PC1, 52.6% of the total variance explained). PC1 is characterized by a rather sharp change of sign at *m*/*z* 165–245. Globally, samples rich in high mass species feature high scores in PC1 > 0 and correspond to strong LII signals and the formation of visible soot deposits. Conversely, samples poor in high mass species feature high scores in PC1 < 0 and lack of LII signal or visible soot deposits. Flame PRE-N is a notable exception as it features high scores in PC1 < 0 that are representative of the gas phase, but at the same time the detection of LII signal shows that NSPs nucleate. In the diffusion flames, large masses are detected at low HAB before the soot formation region identified by LII. This behavior strongly implies a correlation of PC1 with the presence of NSPs in a steady state (PC1 < 0) and with the coagulation and coalescence typical of growing soot particles and the nucleation in the diffusion flames (PC1 > 0). In particular, the change of sign of PC1 can be used to distinguish two different regimes of soot nucleation, and leads to the definition of two mass lists compatible with the formation of NSPs in the investigated premixed and diffusion flames. In conclusion, the presence of large molecules in the gas phase is neither a necessary condition (NSPs nucleate from low mass PAHs in flame PRE-N) nor a sufficient condition (large PAHs are detected in the gas phase in the absence of soot in flames DIF-S and DIF-Q) for soot nucleation.

It was not possible to clearly associate a physical phenomenon to the second principal component (PC2, 23.5% of the total variance explained).

On a molecular level, polycyclic aromatic hydrocarbons (PAHs) and oxygen-containing hydrocarbons (OCHs) are identified by mass defect analysis and correlated to specific loadings, and thus indirectly to specific soot nucleation conditions. PAHs dominate the mass spectra, although a large array of OCHs is also detected in agreement with theoretical calculations and predictions of modeling from the literature. The existence of a large number of compounds containing an odd number of carbon atoms gives indirect evidence of the importance of non-benzenoid or substituted PAHs in the nucleation and growth of NSPs. In particular, many strong signals attributed to PAHs are consistent with 5-member rings that further stress their importance in soot formation and growth [7]. Comparison of the data to kinetic modeling shows that in the growth of both molecular precursors and compounds adsorbed on soot, compact PAHs are preferred over open structures. The comparison of soot sampled from different flames shows that the molecular compounds involved in soot nucleation and growth, when classified according to their H/C ratio, all lie clustered around or slightly above the curve representing the maximally condensed aromatics. The excellent agreement with the predictions of kinetic modeling of D’Anna et al. [50] suggests a similarity in behavior that can be explained by a common origin for instance. However, the hydrocarbons involved in the nucleation of NSPs in premixed and diffusion flames span significantly different mass ranges, below m/z 245 in the premixed flames (left part of the curve) and up to m/z 731 in the diffusion flames (right part of the curve). In the diffusion flames, these large PAHs are found upstream the soot nucleation zone in the gas phase. OCHs are detected in all soot samples. However, PCA shows that they are only associated to the nucleation of NSPs in the diffusion flames and to the particle growth in the premixed flames (PC1 > 0), while they are virtually absent from the pool of molecules associated to the nucleation of NSPs in the premixed flames (PC1 < 0). This behavior is in good agreement with the works of Cain et al. [9] and Commodo et al. [85], and further stresses the differences in the nucleation of NSPs in premixed and diffusion flames. In the premixed flames small

PAHs are sufficient to nucleate NSPs, while in the diffusion flame large amount of OCHs are found from the very early steps of NSP nucleation.

In conclusion, large PAHs and OCHs are not required to nucleate soot. This effect is very evident in at least one premixed flame (PRE-N), in which the formation of condensed phase NSPs involves only PAHs having mass smaller than 400-450 u. Whether the dimer hypothesis is correct, the PAHs involved in the process have mass below 200-225 u, much smaller than expected from thermodynamics-based calculations to further support the idea that soot formation cannot be a thermodynamically controlled process.

5. Acknowledgments

This work was supported by the Agence Nationale de la Recherche through the LABEX CAPPA (ANR-11-LABX-0005), the Ministry of Higher Education and Research, Hauts-de-France Regional Council and European Regional Development Fund (ERDF) through the Contrat de Projets Etat-Region (CPER CLIMIBIO), the H2020 project Portable Nano Particle Emission Measurement System (PEMs4Nano, Grant Agreement no. 724145).

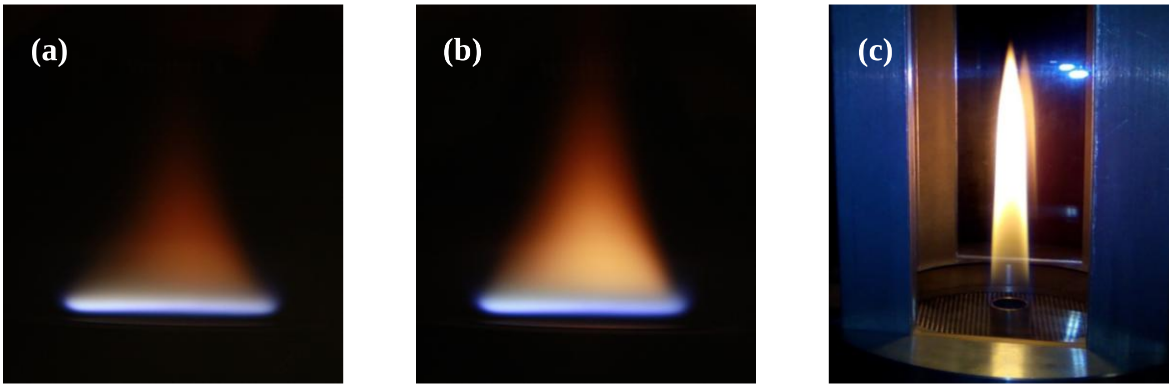


Fig. 1. Investigated flames. (a) Flame PRE-N and (b) flame PRE-S inside the low pressure vessel. (c) Flame DIF-S surrounded by a quartz windowed chimney [45]. During the soot sampling procedure one of the windows is removed for the sampling probe insertion. Flame DIF-Q is very similar to flame DIF-S, and therefore is not shown in the picture.

Flame	Q_{tot} / L min^{-1}	$x_{methane}$	x_{oxygen}	$x_{nitrogen}$
PRE-N	3.96	0.425	0.435	0.140
PRE-S	3.96	0.462	0.398	0.140

Table 1. Laminar premixed flames experimental conditions. Q_{tot} is the total flow, $x_{methane}$, x_{oxygen} and $x_{nitrogen}$ the gas volume fractions. Both flames are stabilized at p = 267 ± 1 mbar.

Flame	Q_{fuel} / L min^{-1}	Q_{dil} / L min^{-1}	Q_{shield} / L min^{-1}	h / mm
DIF-S	0.470	/	87	115
DIF-Q	0.425	0.335	76	70

Table 2. Laminar diffusion flames experimental conditions. Q_{fuel} is the methane flow, Q_{dil} is the dilution argon flow premixed to the methane flow, Q_{shield} is the air shield flow and h the flame visible height measured from the injector exit plane to the tip of the flame.

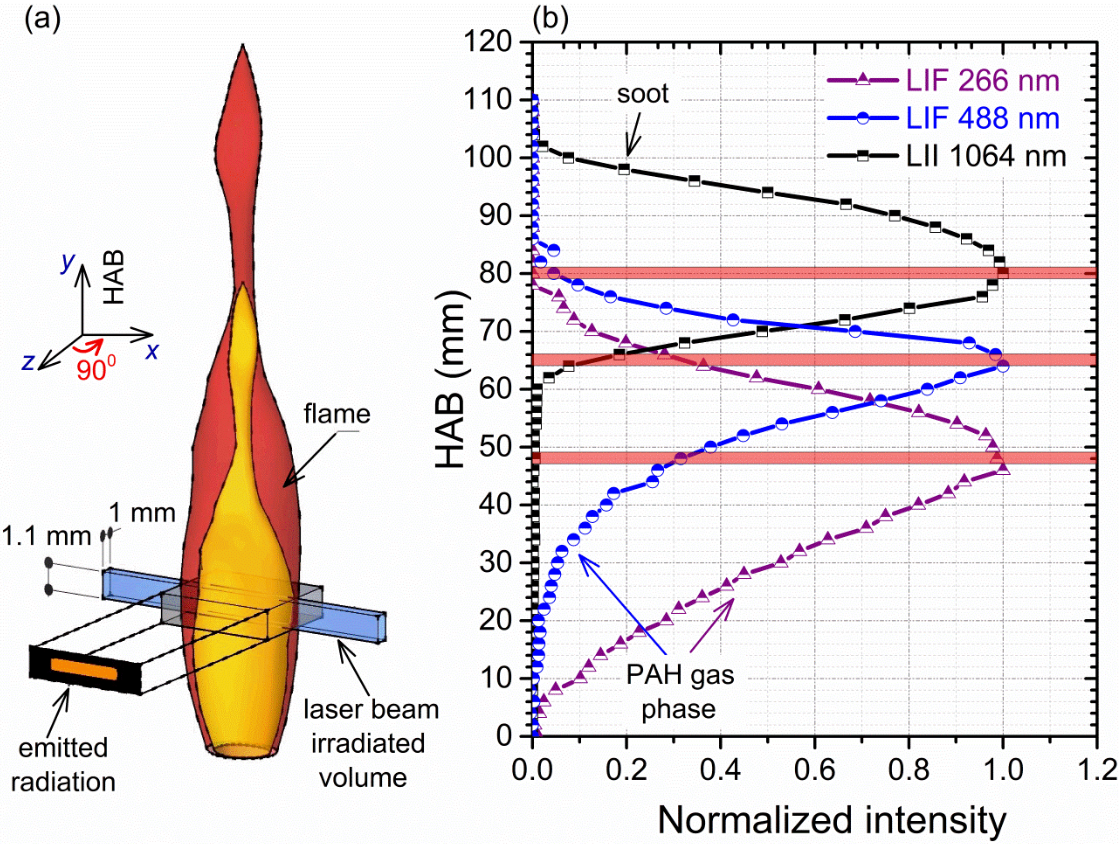


Fig. 2. DIF-S flame. (a) sketch of the LII/LIF working principle: the flame axis is irradiated with a top hat laser beam profile and the emitted radiation is collected at 90° incidence. The emission is collected from a 2 mm deep region on the centerline of the flame. (b) LII/LIF signal against HAB with highlighted sampling locations.

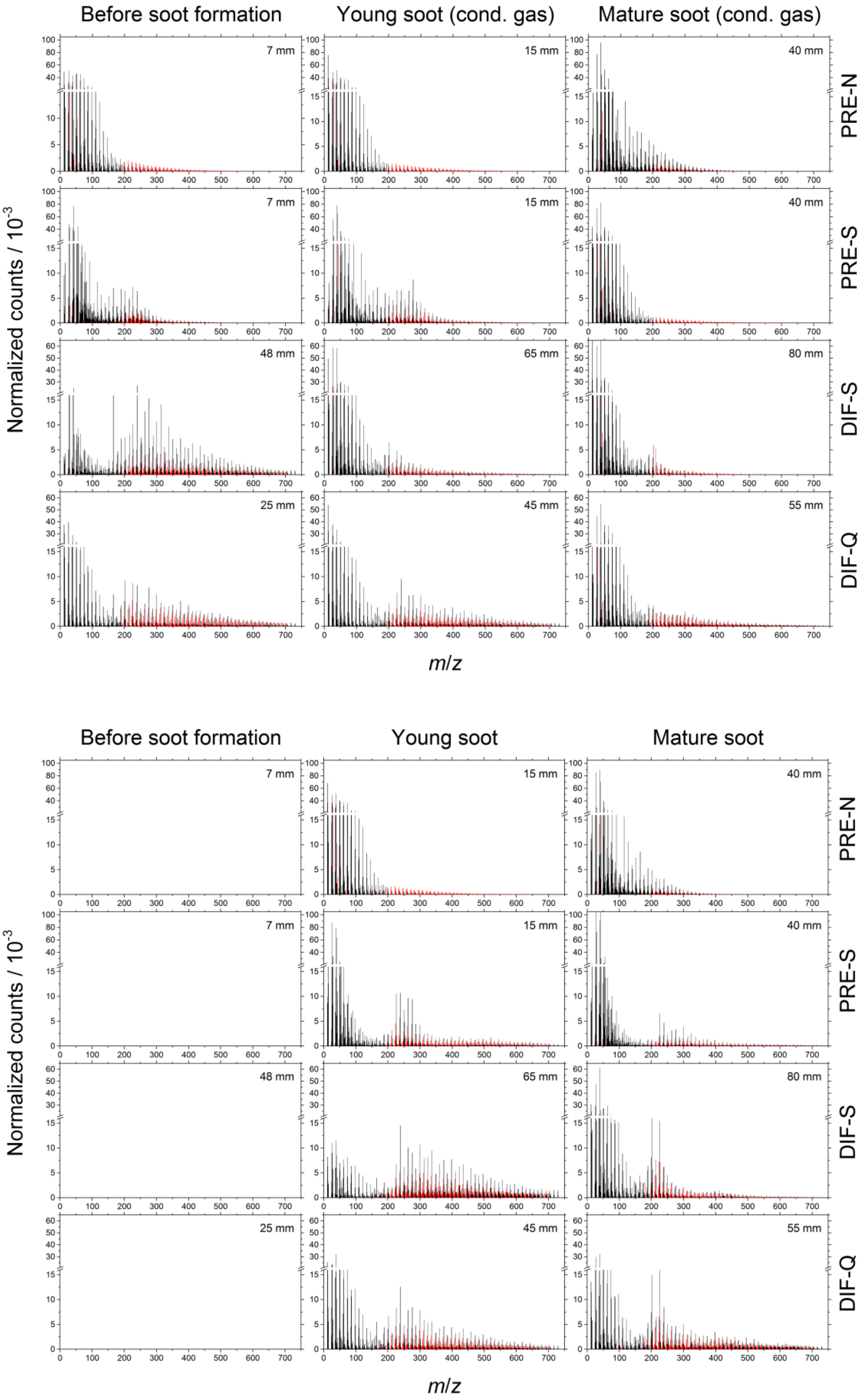


Fig. 3. Mass spectra obtained from the four investigated flames. Each column represents a flame region, from before soot nucleation (left column), to the beginning of the soot nucleation zone (middle column) and up to the mature soot region (right column). Mass spectra of condensable gas and soot are shown in the top and bottom panel, respectively. The intensity count is normalized to the partial ion count calculated from the peak list after background removal. Black and red signals represent hydrocarbons and oxygen-containing hydrocarbons, respectively.

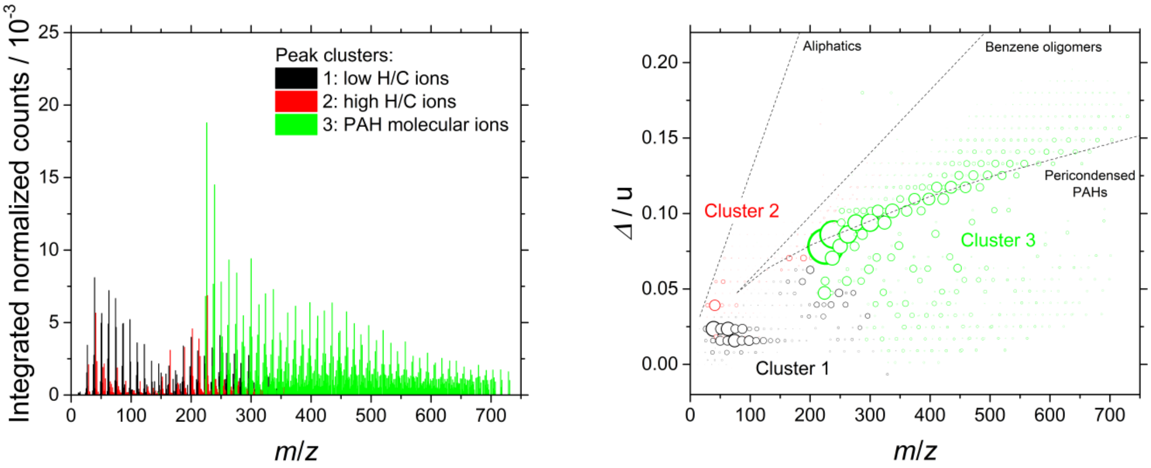


Fig. 4. (a) Example of HCA classification applied to one mass spectrum (flame DIF-S, 65 mm HAB), and (b) to the corresponding mass defect plot. Three patterns are identified by HCA: low H/C ions (black data) are mainly attributed to fragment ions, high H/C ions (red data) to unsaturated hydrocarbons and low-mass PAHs, and high-mass ions (green data) to PAHs and OCHs. The dot size in (b) is proportional to the integrated area of the selected peaks. The dashed lines represent the limits of aliphatic, benzene oligomers and pericondensed PAHs.

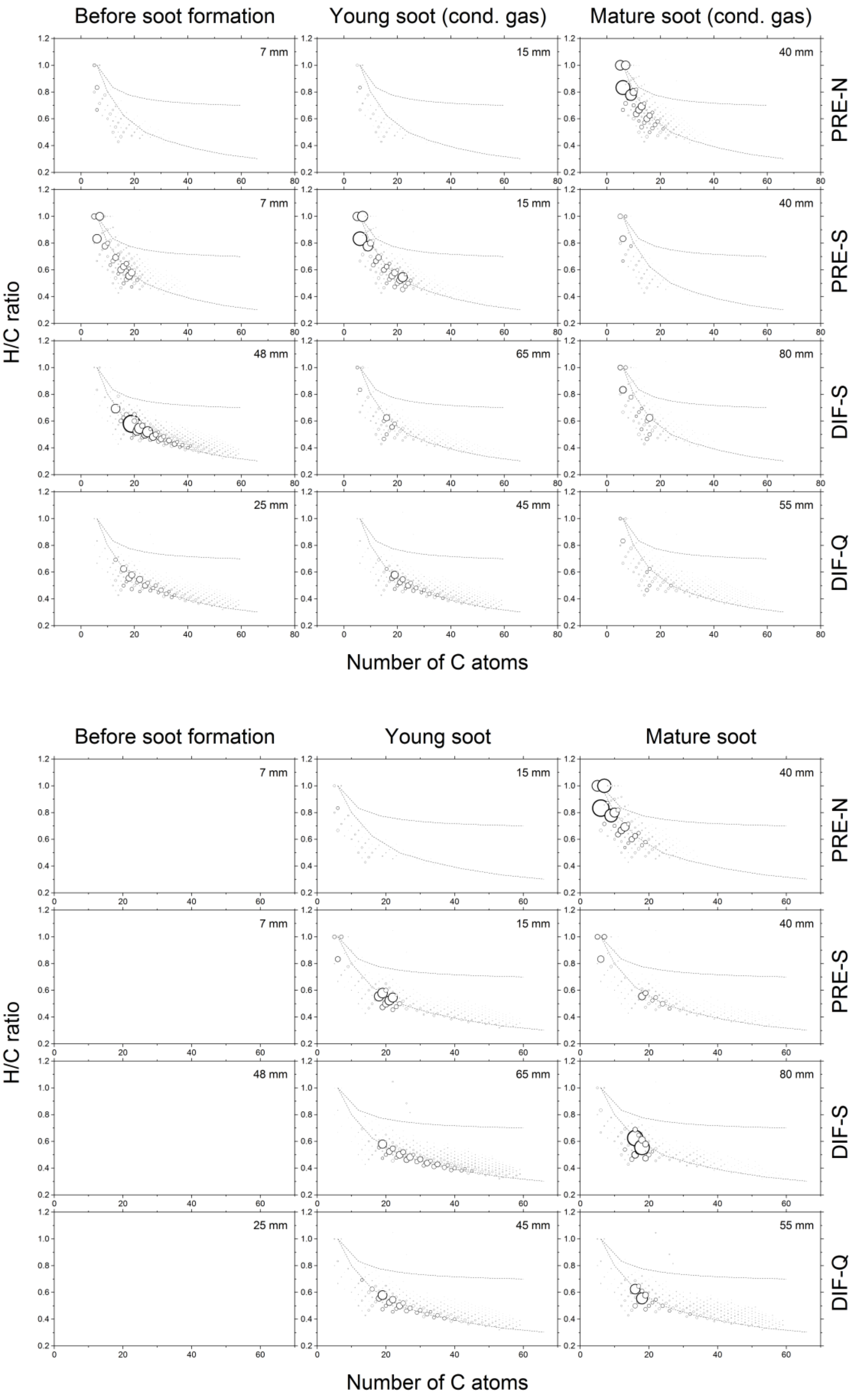


Fig. 5. n/m vs. m/z diagrams. Only $C_mH_n^+$ ions are plotted. Similarly to Fig. 3, rows represent different flames and columns represent different flame regions. Data obtained from the condensable gas and soot are shown in the top and bottom panel, respectively. The black dashed lines represent the limits of benzene oligomers (top line, high n/m) and pericondensed PAHs (bottom line, low n/m) from D'Anna et al. [50].

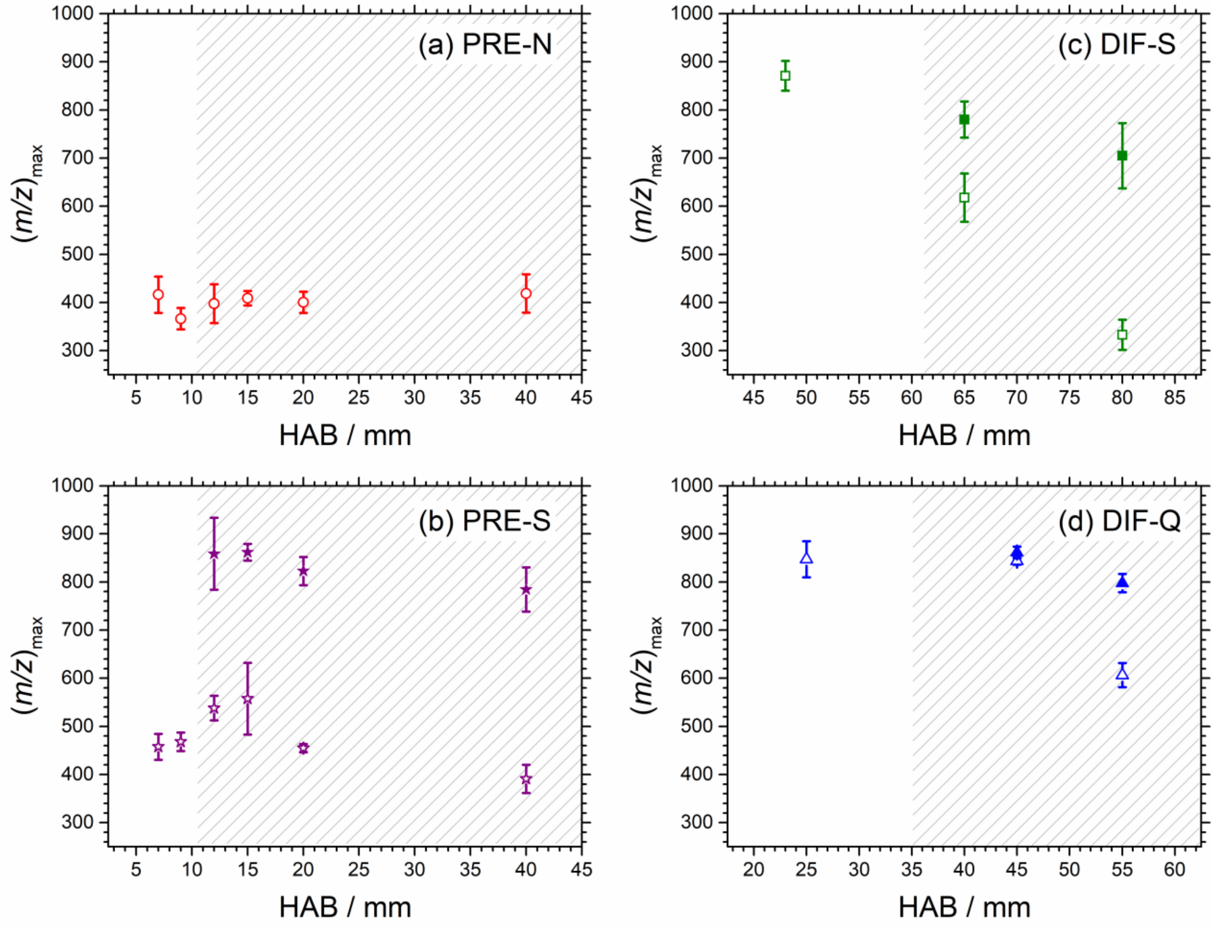


Fig. 6. $(m/z)_{\text{max}}$ against HAB, sorted by flames. The error bars represent the standard deviation calculated on three samples. The gray patterns mark the soot region determined by LII */LIF measurements*.

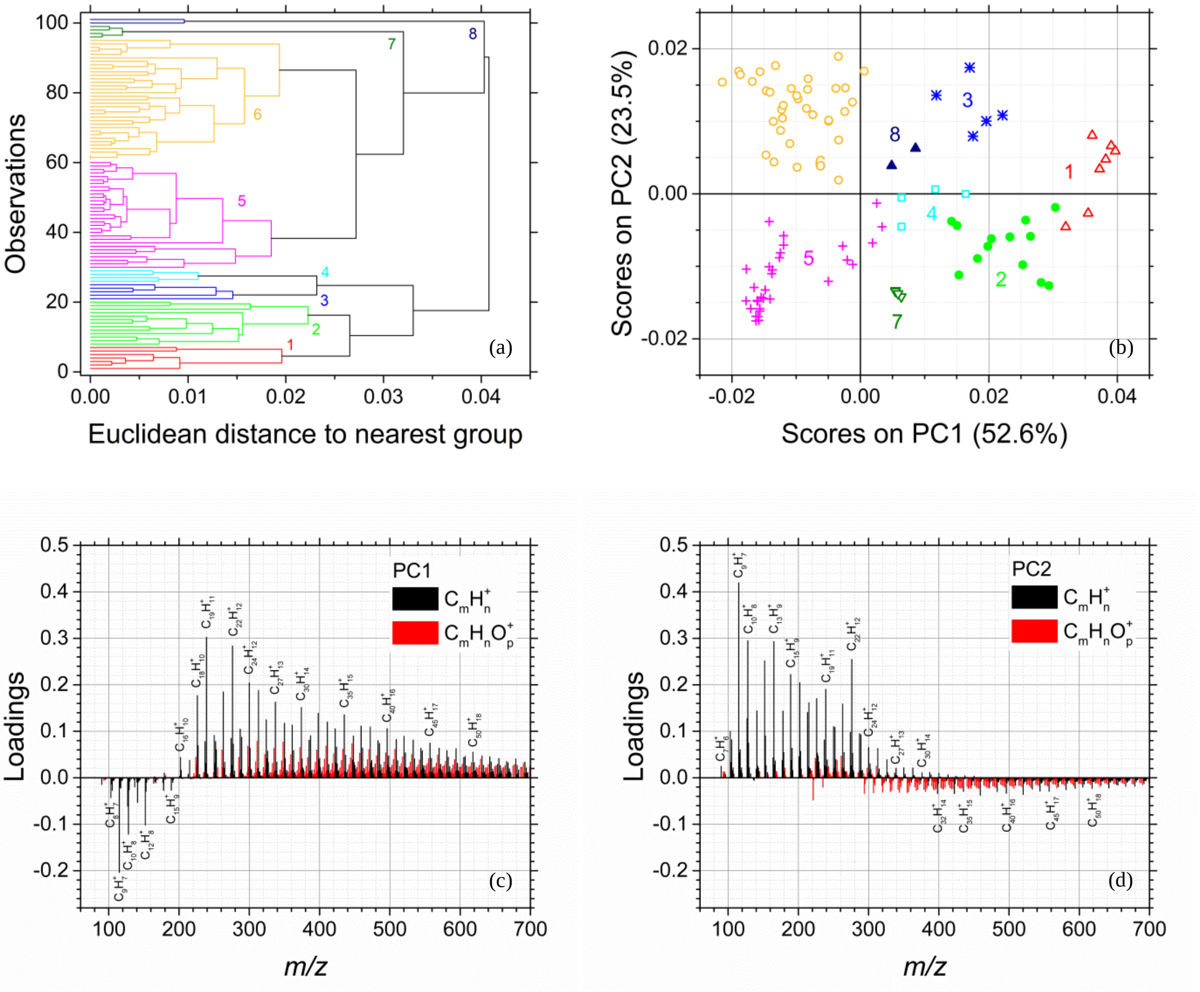


Fig. 7. Samples classification obtained by HCA and PCA. In the top row, the HCA dendrogram (a) and the PC2 vs. PC1 score plot (b) are compared. The colors and cluster numbers match in the two representations. In the bottom row, the loading plots of PC1 and PC2 against *m*/*z* are shown.

Data cluster	PRE-N	PRE-S	DIF-S	DIF-Q
1	/	/	48	25
2	/	/	**65***	**45*,55***
3	/	**12***	/	/
4	/	**15***	/	/
5	7,9,12,15,20	40,**40***	80	55
6	15,40	7,9,12,15,20	/	/
7	/	**20***	/	/
8	/	/	**80***	/

Table 3. Classification of samples by HCA and PCA clusters (rows) and flames (columns). Each cell contains the sampled HAB. Bold starred and normal text mark the samples obtained from the analysis of soot and the condensable gas phase, respectively.

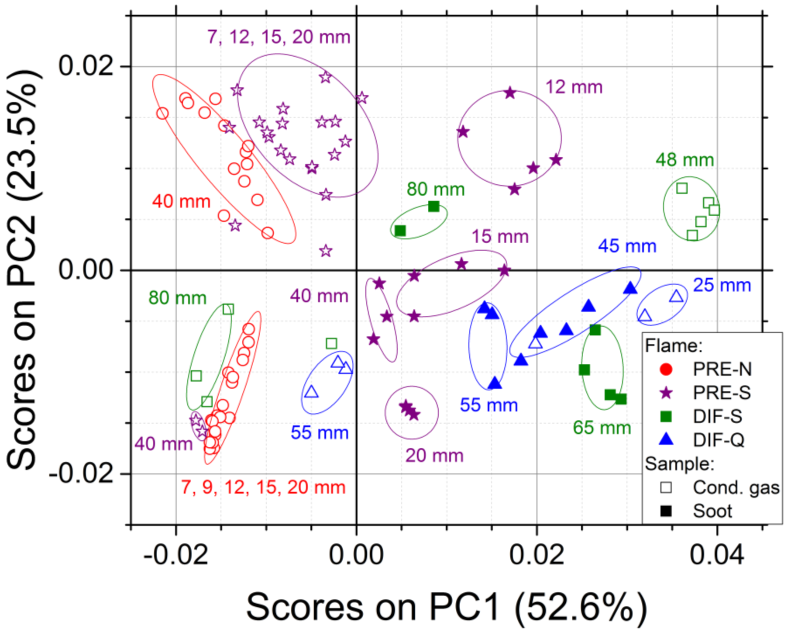


Fig. 8. PC2 vs. PC1 score plot with samples classified by flame (data point shape and color) and analysis of soot or condensable gas (data point interior).

6. References

[1] M. Frenklach, Reaction mechanism of soot formation in flames, Phys. Chem. Chem. Phys. 4 (2002) 2028–2037. doi:10.1039/B110045A.

[2] A. D’Anna, Combustion-formed nanoparticles, Proc. Combust. Inst. 32 (2009) 593–613. doi:10.1016/j.proci.2008.09.005.

[3] H. Wang, Formation of nascent soot and other condensed-phase materials in flames, Proc. Combust. Inst. 33 (2011) 41–67. doi:10.1016/j.proci.2010.09.009.

[4] A.E. Karataş, Ö.L. Gülder, Soot formation in high pressure laminar diffusion flames, Prog. Energy Combust. Sci. 38 (2012) 818–845. doi:10.1016/j.pecs.2012.04.003.

[5] N.A. Eaves, S.B. Dworkin, M.J. Thomson, Assessing relative contributions of PAHs to soot mass by reversible heterogeneous nucleation and condensation, Proc. Combust. Inst. 36 (2017) 935–945. doi:10.1016/j.proci.2016.06.051.

[6] M. Commodo, G. De Falco, A. Bruno, C. Borriello, P. Minutolo, A. D’Anna, Physicochemical evolution of nascent soot particles in a laminar premixed flame: from nucleation to early growth, Combust. Flame. 162 (2015) 3854–3863. doi:10.1016/j.combustflame.2015.07.022.

[7] K.O. Johansson, T. Dillstrom, P. Elvati, M.F. Campbell, P.E. Schrader, D.M. Popolan-Vaida, N.K. Richards-Henderson, K.R. Wilson, A. Violi, H.A. Michelsen, Radical-radical reactions, pyrene nucleation, and incipient soot formation in combustion, Proc. Combust. Inst. 36 (2017) 799–806. doi:10.1016/j.proci.2016.07.130.

[8] P. Desgroux, A. Faccinetto, X. Mercier, T. Mouton, D. Aubagnac Karkar, A. El Bakali, Comparative study of the soot formation process in a nucleation and a sooting low pressure premixed methane flame, Combust. Flame. 184 (2017) 153–166. doi:10.1016/j.combustflame.2017.05.034.

[9] J. Cain, A. Laskin, M.R. Kholghy, M.J. Thomson, H. Wang, Molecular characterization of organic content of soot along the centerline of a coflow diffusion flame, Phys. Chem. Chem. Phys. 16 (2014) 25862–25875. doi:10.1039/C4CP03330B.

[10] P.D. Teini, D.M.A. Karwat, A. Atreya, Observations of nascent soot: molecular deposition and particle morphology, Combust. Flame. 158 (2011) 2045–2055. doi:10.1016/j.combustflame.2011.03.005.

[11] F. Carbone, S. Moslih, A. Gomez, Probing gas-to-particle transition in a moderately sooting atmospheric pressure ethylene/air laminar premixed flame. Part II: Molecular clusters and nascent soot particle size distributions, Combust. Flame. 181 (2017) 329–341. doi:10.1016/j.combustflame.2017.02.021.

[12] P. Desgroux, X. Mercier, K.A. Thomson, Study of the formation of soot and its precursors in flames using optical diagnostics, Proc. Combust. Inst. 34 (2013) 1713–1738. doi:10.1016/j.proci.2012.09.004.

[13] H.A. Michelsen, C. Schulz, G.J. Smallwood, S. Will, Laser-induced incandescence: particulate diagnostics for combustion, atmospheric, and industrial applications, Prog. Energy Combust. Sci. 51 (2015) 2–48. doi:10.1016/j.pecs.2015.07.001.

[14] C. Betrancourt, F. Liu, P. Desgroux, X. Mercier, A. Faccinetto, M. Salamanca, L. Ruwe, K. Kohse-Höinghaus, D. Emmrich, A. Beyer, A. Gölzhäuser, T. Tritscher, Investigation of the size of the incandescent incipient soot particles in premixed sooting and nucleation flames of n-butane using LII, HIM, and 1 nm-SMPS, Aerosol Sci. Technol. 51 (2017) 916–935. doi:10.1080/02786826.2017.1325440.

[15] R.L. Vander Wal, A.J. Tomasek, Soot nanostructure: dependence upon synthesis conditions, Combust. Flame. 136 (2004) 129–140. doi:10.1016/j.combustflame.2003.09.008.

[16] M.R. Kholghy, A. Veshkini, M.J. Thomson, The core–shell internal nanostructure of soot – A criterion to model soot maturity, Carbon. 100 (2016) 508–536. doi:10.1016/j.carbon.2016.01.022.

[17] B. Apicella, P. Pré, M. Alfè, A. Ciajolo, V. Gargiulo, C. Russo, A. Tregrossi, D. Deldique, J.N. Rouzaud, Soot nanostructure evolution in premixed flames by high resolution electron transmission microscopy (HRTEM), Proc. Combust. Inst. 35 (2015) 1895–1902. doi:10.1016/j.proci.2014.06.121.

[18] S.J. Harris, I.M. Kennedy, The coagulation of soot particles with Van der Waals forces, Combust. Sci. Technol. 59 (1988) 443–454. doi:10.1080/00102208808947110.

[19] M. Frenklach, H. Wang, Detailed modeling of soot particle nucleation and growth, Proc. Combust. Inst. 23 (1990) 1559–1566. doi:10.1016/S0082-0784(06)80426-1.

[20] J.S. Lowe, J.Y.W. Lai, P. Elvati, A. Violi, Towards a predictive model for polycyclic aromatic hydrocarbon dimerization propensity, Proc. Combust. Inst. 35 (2015) 1827–1832. doi:10.1016/j.proci.2014.06.142.

[21] A. Violi, A.F. Sarofim, G.A. Voth, Kinetic Monte Carlo molecular dynamics approach to model soot inception, Combust. Sci. Technol. 176 (2004) 991–1005. doi:10.1080/00102200490428594.

[22] Q. Mao, A.C.T. van Duin, K.H. Luo, Formation of incipient soot particles from polycyclic aromatic hydrocarbons: A ReaxFF molecular dynamics study, Carbon. 121 (2017) 380–388. doi:10.1016/j.carbon.2017.06.009.

[23] S.E. Stein, A. Fahr, High-temperature stabilities of hydrocarbons, J. Phys. Chem. 89 (1985) 3714–3725. doi:10.1021/j100263a027.

[24] B. Öktem, M.P. Tolocka, B. Zhao, H. Wang, M.V. Johnston, Chemical species associated with the early stage of soot growth in a laminar premixed ethylene–oxygen–argon flame, Combust. Flame. 142 (2005) 364–373. doi:10.1016/j.combustflame.2005.03.016.

[25] R.A. Dobbins, R.A. Fletcher, B.A. Benner, S. Hoeft, Polycyclic aromatic hydrocarbons in flames, in diesel fuels, and in diesel emissions, Combust. Flame. 144 (2006) 773–781. doi:10.1016/j.combustflame.2005.09.008.

[26] M.M. Maricq, Physical and chemical comparison of soot in hydrocarbon and biodiesel fuel diffusion flames: A study of model and commercial fuels, Combust. Flame. 158 (2011) 105–116. doi:10.1016/j.combustflame.2010.07.022.

[27] A. Faccinetto, P. Desgroux, M. Ziskind, E. Therssen, C. Focsa, High-sensitivity detection of polycyclic aromatic hydrocarbons adsorbed onto soot particles using laser desorption/laser ionization/time-of-flight mass spectrometry: an approach to studying the soot inception process in low-pressure flames, Combust. Flame. 158 (2011) 227–239. doi:10.1016/j.combustflame.2010.08.012.

[28] P. Elvati, A. Violi, Thermodynamics of poly-aromatic hydrocarbon clustering and the effects of substituted aliphatic chains, Proc. Combust. Inst. 34 (2013) 1837–1843. doi:10.1016/j.proci.2012.07.030.

[29] A. Raj, M. Celnik, R. Shirley, M. Sander, R. Patterson, R. West, M. Kraft, A statistical approach to develop a detailed soot growth model using PAH characteristics, Combust. Flame. 156 (2009) 896–913. doi:10.1016/j.combustflame.2009.01.005.

[30] H. Richter, J.B. Howard, Formation of polycyclic aromatic hydrocarbons and their growth to soot—a review of chemical reaction pathways, Prog. Energy Combust. Sci. 26 (2000) 565–608. doi:10.1016/S0360-1285(00)00009-5.

[31] M.M. Maricq, An examination of soot composition in premixed hydrocarbon flames via laser ablation particle mass spectrometry, J. Aerosol Sci. 40 (2009) 844–857. doi:10.1016/j.jaerosci.2009.07.001.

[32] F. Schulz, M. Commodo, K. Kaiser, G. De Falco, P. Minutolo, G. Meyer, A. D`Anna, L. Gross, Insights into incipient soot formation by atomic force microscopy, Proc. Combust. Inst. (2018). doi:10.1016/j.proci.2018.06.100.

[33] P. Elvati, A. Violi, Homo-dimerization of oxygenated polycyclic aromatic hydrocarbons under flame conditions, Fuel. 222 (2018) 307–311. doi:10.1016/j.fuel.2018.02.095.

[34] K.O. Johansson, T. Dillstrom, M. Monti, F. El Gabaly, M.F. Campbell, P.E. Schrader, D.M. Popolan-Vaida, N.K. Richards-Henderson, K.R. Wilson, A. Violi, H.A. Michelsen, Formation and emission of large furans and oxygenated hydrocarbons from flames, Proc. Natl. Acad. Sci. 113 (2016) 8374–8379. doi:10.1073/pnas.1604772113.

[35] P. Elvati, V.T. Dillstrom, A. Violi, Oxygen driven soot formation, Proc. Combust. Inst. 36 (2017) 825–832. doi:10.1016/j.proci.2016.09.019.

[36] R.L. Vander Wal, A. Yezerets, N.W. Currier, D.H. Kim, C.M. Wang, HRTEM Study of diesel soot collected from diesel particulate filters, Carbon. 45 (2007) 70–77. doi:10.1016/j.carbon.2006.08.005.

[37] G.A. Kelesidis, E. Goudeli, S.E. Pratsinis, Morphology and mobility diameter of carbonaceous aerosols during agglomeration and surface growth, Carbon. 121 (2017) 527–535. doi:10.1016/j.carbon.2017.06.004.

[38] P. Desgroux, X. Mercier, B. Lefort, R. Lemaire, E. Therssen, J.-F. Pauwels, Soot volume fraction measurement in low-pressure methane flames by combining laser-induced incandescence and cavity ring-down spectroscopy: Effect of pressure on soot formation, Combust. Flame. 155 (2008) 289–301. doi:10.1016/j.combustflame.2008.05.016.

[39] J. Yon, R. Lemaire, E. Therssen, P. Desgroux, A. Coppalle, K.F. Ren, Examination of wavelength dependent soot optical properties of diesel and diesel/rapeseed methyl ester mixture by extinction spectra analysis and LII measurements, Appl. Phys. B. 104 (2011) 253–271. doi:10.1007/s00340-011-4416-4.

[40] A.D. Abid, N. Heinz, E.D. Tolmachoff, D.J. Phares, C.S. Campbell, H. Wang, On evolution of particle size distribution functions of incipient soot in premixed ethylene-oxygen-argon flames, Combust. Flame. 154 (2008) 775–788. doi:10.1016/j.combustflame.2008.06.009.

[41] M. Schenk, S. Lieb, H. Vieker, A. Beyer, A. Gölzhäuser, H. Wang, K. Kohse-Höinghaus, Imaging nanocarbon materials: soot particles in flames are not structurally homogeneous, ChemPhysChem. 14 (2013) 3248–3254. doi:10.1002/cphc.201300581.

[42] H.-H. Grotheer, H. Pokorny, K.-L. Barth, M. Thierley, M. Aigner, Mass spectrometry up to 1 million mass units for the simultaneous detection of primary soot and of soot precursors (nanoparticles) in flames, Chemosphere. 57 (2004) 1335–1342. doi:10.1016/j.chemosphere.2004.08.054.

[43] B. Apicella, A. Carpentieri, M. Alfè, R. Barbella, A. Tregrossi, P. Pucci, A. Ciajolo, Mass spectrometric analysis of large PAH in a fuel-rich ethylene flame, Proc. Combust. Inst. 31 (2007) 547–553. doi:10.1016/j.proci.2006.08.014.

[44] R.A. Dobbins, R.A. Fletcher, H.-C. Chang, The evolution of soot precursor particles in a diffusion flame, Combust. Flame. 115 (1998) 285–298. doi:10.1016/S0010-2180(98)00010-8.

[45] C. Irimiea, A. Faccinetto, Y. Carpentier, I.-K. Ortega, N. Nuns, E. Therssen, P. Desgroux, C. Focsa, A comprehensive protocol for chemical analysis of flame combustion emissions by secondary ion mass spectrometry, Rapid Commun. Mass Spectrom. 32 (2018) 1015–1025. doi:10.1002/rcm.8133.

[46] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. doi:10.1039/C3AY41907J.

[47] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos. Transact. A Math. Phys. Eng. Sci. 374 (2016) 20150202. doi:10.1098/rsta.2015.0202.

[48] T. Mouton, X. Mercier, M. Wartel, N. Lamoureux, P. Desgroux, Laser-induced incandescence technique to identify soot nucleation and very small particles in low-pressure methane flames, Appl. Phys. B. 112 (2013) 369–379. doi:10.1007/s00340-013-5446-x.

[49] D. Aubagnac-Karkar, A. El Bakali, P. Desgroux, Soot particles inception and PAH condensation modelling applied in a soot model utilizing a sectional method, Combust. Flame. 189 (2018) 190–206. doi:10.1016/j.combustflame.2017.10.027.

[50] A. D'Anna, M. Sirignano, J. Kent, A model of particle nucleation in premixed ethylene flames, Combust. Flame. 157 (2010) 2106–2115. doi:10.1016/j.combustflame.2010.04.019.

[51] H. Bladh, N.-E. Olofsson, T. Mouton, J. Simonsson, X. Mercier, A. Faccinetto, P.-E. Bengtsson, P. Desgroux, Probing the smallest soot particles in low-sooting premixed flames using laser-induced incandescence, Proc. Combust. Inst. 35 (2015) 1843–1850. doi:10.1016/j.proci.2014.06.001.

[52] J.C. Biordi, C.P. Lazzara, J.F. Papp, Molecular beam mass spectrometry applied to determining the kinetics of reactions in flames. I. Empirical characterization of flame perturbation by molecular beam sampling probes, Combust. Flame. 23 (1974) 73–82. doi:10.1016/S0010-2180(74)80028-3.

[53] H. Hepp, K. Siegmann, Mapping of soot particles in a weakly sooting diffusion flame by aerosol techniques, Combust. Flame. 115 (1998) 275–283. doi:10.1016/S0010-2180(97)00346-5.

[54] K. Siegmann, K. Sattler, H.C. Siegmann, Clustering at high temperatures: carbon formation in combustion, J. Electron Spectrosc. Relat. Phenom. 126 (2002) 191–202. doi:10.1016/S0368-2048(02)00152-4.

[55] H.A. Michelsen, Probing soot formation, chemical and physical evolution, and oxidation: A review of in situ diagnostic techniques and needs, Proc. Combust. Inst. 36 (2017) 717–735. doi:10.1016/j.proci.2016.08.027.

[56] C. Schoemaecker Moreau, E. Therssen, X. Mercier, J.-F. Pauwels, P. Desgroux, Two-color laser-induced incandescence and cavity ring-down spectroscopy for sensitive and quantitative imaging of soot and PAHs in flames, Appl. Phys. B. 78 (2004) 485–492. doi:10.1007/s00340-003-1370-9.

[57] S. Bejaoui, R. Lemaire, E. Therssen, Analysis of laser-induced fluorescence spectra obtained in spray flames of diesel and rapeseed methyl ester using the multiple-excitation wavelength laser-induced incandescence technique with IR, UV, and visible excitations, Combust. Sci. Technol. 187 (2015) 906–924. doi:10.1080/00102202.2014.973949.

[58] M. Wartel, J.-F. Pauwels, P. Desgroux, X. Mercier, Pyrene measurements in sooting low pressure methane flames by jet-cooled laser-induced fluorescence, J. Phys. Chem. A. 115 (2011) 14153–14162. doi:10.1021/jp206970t.

[59] H.J. Heger, R. Zimmermann, R. Dorfner, M. Beckmann, H. Griebel, A. Kettrup, U. Boesl, On-line emission analysis of polycyclic aromatic hydrocarbons down to pptv concentration levels in the flue gas of an incineration pilot plant with a mobile resonance-enhanced multiphoton ionization time-of-flight mass spectrometer, Anal. Chem. 71 (1999) 46–57. doi:10.1021/ac980611y.

[60] A.C. Aiken, P.F. DeCarlo, J.L. Jimenez, Elemental analysis of organic species with electron ionization high-resolution mass spectrometry, Anal. Chem. 79 (2007) 8350–8358. doi:10.1021/ac071150w.

[61] L.M. Russell, R. Bahadur, L.N. Hawkins, J. Allan, D. Baumgardner, P.K. Quinn, T.S. Bates, Organic aerosol characterization by complementary measurements of chemical bonds and molecular fragments, Atmos. Environ. 43 (2009) 6100–6105. doi:10.1016/j.atmosenv.2009.09.036.

[62] D.S. Liscinsky, Z. Yu, B. True, J. Peck, A.C. Jennings, H.-W. Wong, J. Franklin, S.C. Herndon, R.C. Miake-Lye, Measurement of naphthalene uptake by combustion soot particles, Environ. Sci. Technol. 47 (2013) 4875–4881. doi:10.1021/es304912d.

[63] M.T. Timko, S.E. Albo, T.B. Onasch, E.C. Fortner, Z. Yu, R.C. Miake-Lye, M.R. Canagaratna, N.L. Ng, D.R. Worsnop, Composition and sources of the organic particle emissions from aircraft engines, Aerosol Sci. Technol. 48 (2014) 61–73. doi:10.1080/02786826.2013.857758.

[64] K.O. Johansson, J. Zádor, P. Elvati, M.F. Campbell, P.E. Schrader, N.K. Richards-Henderson, K.R. Wilson, A. Violi, H.A. Michelsen, Critical assessment of photoionization efficiency measurements for characterization of soot-precursor species, J. Phys. Chem. A. 121 (2017) 4475–4485. doi:10.1021/acs.jpca.7b02992.

[65] M.D. Hays, R.J. Lavrich, Developments in direct thermal extraction gas chromatography-mass spectrometry of fine aerosols, Trend Anal Chem. 26 (2007) 88–102.

[66] J. Gil-Moltó, M. Varea, N. Galindo, J. Crespo, Application of an automatic thermal desorption-gas chromatography-mass spectrometry system for the analysis of polycyclic aromatic hydrocarbons in airborne particulate matter, J. Chromatogr. A. 1216 (2009) 1285–1289. doi:10.1016/j.chroma.2008.12.080.

[67] C. Mihesan, M. Ziskind, E. Therssen, P. Desgroux, C. Focsa, IR laser resonant desorption of polycyclic aromatic hydrocarbons, Chem. Phys. Lett. 423 (2006) 407–412. doi:10.1016/j.cplett.2006.04.032.

[68] O.P. Haefliger, R. Zenobi, New sample preparation for quantitative laser desorption mass spectrometry and optical spectroscopy, Rev. Sci. Instrum. 69 (1998) 1828–1832.

[69] J.W.G. Bentz, J. Goschnick, J. Schuricht, H.J. Ache, J. Zehnpfennig, A. Benninghoven, Analysis and classification of individual outdoor aerosol particles with SIMS time-of-flight mass spectrometry, Fresenius J. Anal. Chem. 353 (1995) 603–608. doi:10.1007/BF00321333.

[70] O.B. Popovicheva, C. Irimiea, Y. Carpentier, I.K. Ortega, E.D. Kireeva, N.K. Shonija, J. Schwarz, M. Vojtíšek-Lom, C. Focsa, Chemical composition of diesel/biodiesel particulate exhaust by FTIR spectroscopy and mass spectrometry: impact of fuel and driving cycle, Aerosol Air Qual. Res. 17 (2017) 1717–1734. doi:10.4209/aaqr.2017.04.0127.

[71] T. Stephan, E.K. Jessberger, C.H. Heiss, D. Rost, TOF-SIMS analysis of polycyclic aromatic hydrocarbons in Allan Hills 84001, Meteorit. Planet. Sci. 38 (2003) 109–116. doi:10.1111/j.1945-5100.2003.tb01049.x.

[72] R.E. Peterson, B.J. Tyler, Analysis of organic and inorganic species on the surface of atmospheric aerosol using time-of-flight secondary ion mass spectrometry (TOF-SIMS), Atmos. Environ. 36 (2002) 6041–6049. doi:10.1016/S1352-2310(02)00686-6.

[73] L. Pei, G. Jiang, B.J. Tyler, L.L. Baxter, M.R. Linford, Time-of-flight secondary ion mass spectrometry of a range of coal samples: A chemometrics (PCA, cluster, and PLS) analysis, Energy Fuels. 22 (2008) 1059–1072. doi:10.1021/ef7003199.

[74] P. Reitz, S.R. Zorn, S.H. Trimborn, A.M. Trimborn, A new, powerful technique to analyze single particle aerosol mass spectra using a combination of OPTICS and the fuzzy c-means algorithm, J. Aerosol Sci. 98 (2016) 1–14. doi:10.1016/j.jaerosci.2016.04.003.

[75] H. Abdi, L. Williams, Principal component analysis, Wiley Interdiscip Rev Comput Stat. 2 (2010) 433–459. doi:10.1002/wics.101.

[76] R. Baigorri, A.M. Zamarreño, M. Fuentes, G. González-Gaitano, J.M. García-Mina, G. Almendros, F.J. González-Vila, Multivariate statistical analysis of mass spectra as a tool for the classification of the main humic substances according to their structural and conformational features, J. Agric. Food Chem. 56 (2008) 5480–5487. doi:10.1021/jf800507u.

[77] O.P. Haefliger, R. Zenobi, Laser mass spectrometric analysis of polycyclic aromatic hydrocarbons with wide wavelength range laser multiphoton ionization spectroscopy, Anal. Chem. 70 (1998) 2660–2665. doi:10.1021/ac971264f.

[78] L. Sleno, The use of mass defect in modern mass spectrometry, J. Mass Spectrom. 47 (2012) 226–236. doi:10.1002/jms.2978.

[79] A. Keller, R. Kovacs, K.-H. Homann, Large molecules, ions, radicals and small soot particles in fuel-rich hydrocarbon flames. Part IV. Large polycyclic aromatic hydrocarbons and their radicals in a fuel-rich benzene-oxygen flame, Phys. Chem. Chem. Phys. 2 (2000) 1667–1675. doi:10.1039/A908190I.

[80] D. Chen, Z. Zainuddin, E. Yapp, J. Akroyd, S. Mosbach, M. Kraft, A fully coupled simulation of PAH and soot growth with a population balance model, Proc. Combust. Inst. 34 (2013) 1827–1835. doi:10.1016/j.proci.2012.06.089.

[81] K.O. Johansson, M.P. Head-Gordon, P.E. Schrader, K.R. Wilson, H.A. Michelsen, Resonance-stabilized hydrocarbon-radical chain reactions may explain soot inception and growth, Science. 361 (2018) 997. doi:10.1126/science.aat3417.

[82] H.-B. Zhang, X. You, H. Wang, C.K. Law, Dimerization of polycyclic aromatic hydrocarbons in soot nucleation, J. Phys. Chem. A. 118 (2014) 1287–1292. doi:10.1021/jp411806q.

[83] M.R. Kholghy, G.A. Kelesidis, S.E. Pratsinis, Reactive polycyclic aromatic hydrocarbon dimerization drives soot nucleation, Phys. Chem. Chem. Phys. 20 (2018) 10926–10938. doi:10.1039/C7CP07803J.

[84] C.A. Schuetz, M. Frenklach, Nucleation of soot: Molecular dynamics simulations of pyrene dimerization, Proc. Combust. Inst. 29 (2002) 2307–2314. doi:10.1016/S1540-7489(02)80281-4.

[85] M. Commodo, A. D’Anna, G. De Falco, R. Larciprete, P. Minutolo, Illuminating the earliest stages of the soot formation by photoemission and Raman spectroscopy, Combust. Flame. 181 (2017) 188–197. doi:10.1016/j.combustflame.2017.03.020.

[86] F.-X. Ouf, P. Parent, C. Laffon, I. Marhaba, D. Ferry, B. Marcillaud, E. Antonsson, S. Benkoula, X.-J. Liu, C. Nicolas, E. Robert, M. Patanen, F.-A. Barreda, O. Sublemontier, A. Coppalle, J. Yon, F. Miserque, T. Mostefaoui, T.Z. Regier, J.-B.A. Mitchell, C. Miron, First in-flight synchrotron X-ray absorption and photoemission study of carbon soot nanoparticles, Sci. Rep. 6 (2016) 36495. doi:10.1038/srep36495.

Supporting information

1. SI. Definition of terms used in mass spectrometry

This section contains the IUPAC definitions of some terms related to mass spectrometry and used in the paper [87].

- *Accurate mass.* Experimentally determined mass of an ion of known charge. Can be used to determine elemental composition to within limits defined by both the accuracy and precision of the measurement.
- *Exact mass.* Calculated mass of an ion or molecule with specified isotopic composition.
- *Mass calibration.* Means of determining *m*/*z* values of ions from experimentally detected signals using a theoretical or empirical relational equation. Most commonly, this is accomplished using a computer-based data system and a calibration file obtained from a mass spectrum of a compound that produces ions of known *m*/z values.
- *Mass defect*. Difference between the nominal mass and the monoisotopic mass of an atom, molecule, or ion. Mass defect can be a positive or negative value dependent upon the elemental composition.
- *Monoisotopic mass.* Exact mass of an ion or molecule calculated using the mass of the most abundant isotope of each element.
- *m*/*z*. Abbreviation representing the dimensionless quantity formed by dividing the ratio of the mass of an ion to the unified atomic mass unit, by its charge number (regardless of sign). The abbreviation is written in italicized lowercase letters with no spaces. *m*/*z* is recommended as an abbreviation to represent the dimensionless quantity that is used almost universally as the independent variable in a mass spectrum. The abbreviation m/z should not be used as a variable in a mathematical equation. Instead, the variables m (in kg) and q (in C) should be used to denote mass and charge
- *Nominal mass*. Mass of a molecular ion or molecule calculated using the isotope mass of the most abundant constituent element isotope of each element rounded to the nearest integer value and multiplied by the number of atoms of each element.

- *Resolution.* In a mass spectrum, the observed *m*/*z* value divided by the smallest difference Δ(*m*/*z*) for two ions that can be separated: (*m*/*z*)/Δ(*m*/*z*). The *m*/*z* value at which the measurement was made should be reported. The definition and method of measurement of Δ(*m*/*z*) should be reported. Commonly, this is performed using peak width measured at a specified percentage of peak height. Alternatively, Δ(*m*/*z*) is defined as the separation between two adjacent equal magnitude peaks such that the valley between them is a specified fraction of the peak height, for example as measured by peak matching.
- *Resolving power.* Measure of the ability of a mass spectrometer to provide a specified value of mass resolution.
- *Unified atomic mass unit (u)*. Non-SI unit of mass defined as one-twelfth of the mass of one atom of ^{12}C at rest in its ground state and equal to $1.660538921(73)\cdot 10^{-27}$ kg where the digits in parentheses indicate the estimated uncertainty in the final two digits of the value. Equivalent to the Dalton (Da) unit. The abbreviation amu for atomic mass unit is deprecated; it has been used to denote atomic masses measured relative to a single atom of ^{16}O, or to the isotope-averaged mass of an oxygen atom, or to a single atom of ^{12}C.

2. SI. Local factors affecting the structure of the database

The complete analyzed database consists of 152 samples and contains mass spectrometry data obtained from depositions of both soot and condensable gas on three different substrates: (1) binder-free, glass microfiber Whatman filters (21 mm diameter, 0.26 mm thickness, 52 g m^{-2}, 0.7 µm smallest pore size purchased from Fisher Scientific), (2) Silicon (100) wafers (10x10x0.5 mm optically polished on one side purchased from PI-KEM Ltd) and (3) grade 5 titanium alloy wafers (90% Ti, 6% Al, 4% V, 10x10x0.5 mm optically polished on one side purchased from Optics Concept).

Fig. S1a shows the HCA output dendrogram, while the other three panels show the PCA output: Fig. S1b contains the PC1 vs. PC2 score plot and Fig. S1c and S1d contain the coefficient plots of the first two PCs vs. *m*/*z*. The database is reasonably well described with the first two PCs that together explain 67.3% of the total variance. In Fig. S1b the shape of the data points represents the deposition substrate: glass microfiber filters (round empty dots), silicon wafers (crosses) and titanium wafers (square solid dots). The clusters of samples identified by HCA are color-coded to match the data in the score plot. A satisfying agreement of HCA and PCA is found by following the data reduction procedure proposed in [73] that results in eight data clusters well separated and adjacent to each others in the score plot. The list of samples in each cluster can be found in Table S1. The analysis of the PC2 vs. PC1 score plot reveals that all depositions on glass fiber filters but three (clusters 1, 2 and 8) and the depositions contaminated with organosilicon compounds (cluster 3) all appear at PC1 < 0, thus the deposition substrate and the presence of organosilicon contaminants rather than the nature of the samples give the main contribution (PC1, 43.8%) to the total variance of the database. More in detail:

- a statistically significant difference exists between samples deposited on microfiber glass filters and on wafers. As shown in Fig. S1b, all samples obtained by deposition on microfiber glass filters (clusters 1, 2 and 8) appear at PC1 < 0. On the other hand, no statistically significant difference exists between samples deposited on silicon and titanium wafers once the mass spectra are normalized and the background peaks ($Si_xO_y^+$ and organosilicon ions in Si wafers, $Ti_xO_y^+$ and organotitanium ions in Ti wafers) are removed. To verify the equivalence of silicon and titanium wafers, ten samples are prepared in the same experimental conditions, five by deposition on silicon and five on titanium wafers. The intensity of the signals in the peak list are built into a scatter matrix and used to correlate each sample with all the samples obtained on the other substrate. The high correlation between intensities in the two groups (Pearson's coefficient variability range 0.81-0.98) and the absence of a significant separation in the score plots enforce the operational equivalence of silicon and titanium wafers as substrates for soot deposition;
- significant amounts of organosilicon compounds were occasionally detected and attributed to uncontrolled thermal desorption from the conductive tubing for aerosol transport used in the early implementation of the sampling line [45]. This is a well known issue [88,89]. In our case, all the contaminated samples are grouped in cluster 3. Some of the depositions obtained from flame PRE-S before replacing the tubing were heavily contaminated and had to be discarded (underlined in Table S1). These samples were easy to recognize by visual inspection of the wafer surface due to the polysiloxane distinctive light

polarization effect [45]. However, lighter contaminations are more difficult to identify. To this regard, PCA proved to be so effective that some other depositions were recognized as slightly contaminated only after PCA classified them in cluster 3.

This situation is less than ideal since high negative scores on PC1 (43.8%) correspond to samples affected by analytical and local factors, while the contribution of the hydrocarbons is relegated back to PC2 (23.5%) and even PC3 (11.8%). Therefore, to improve the quality of the data reduction, we removed from the database all the depositions on microfiber glass filters and the samples heavily contaminated with organosilicon compounds, and then re-analyzed the reduced data set (101 samples). The results of this second analysis are shown and discussed in the main text. Although at this stage it is rather difficult to obtain detailed information from the score plot, it is noteworthy that PC2 > 0 is mostly associated to PAH masses, as shown by the coefficients plot in Fig. S1d.

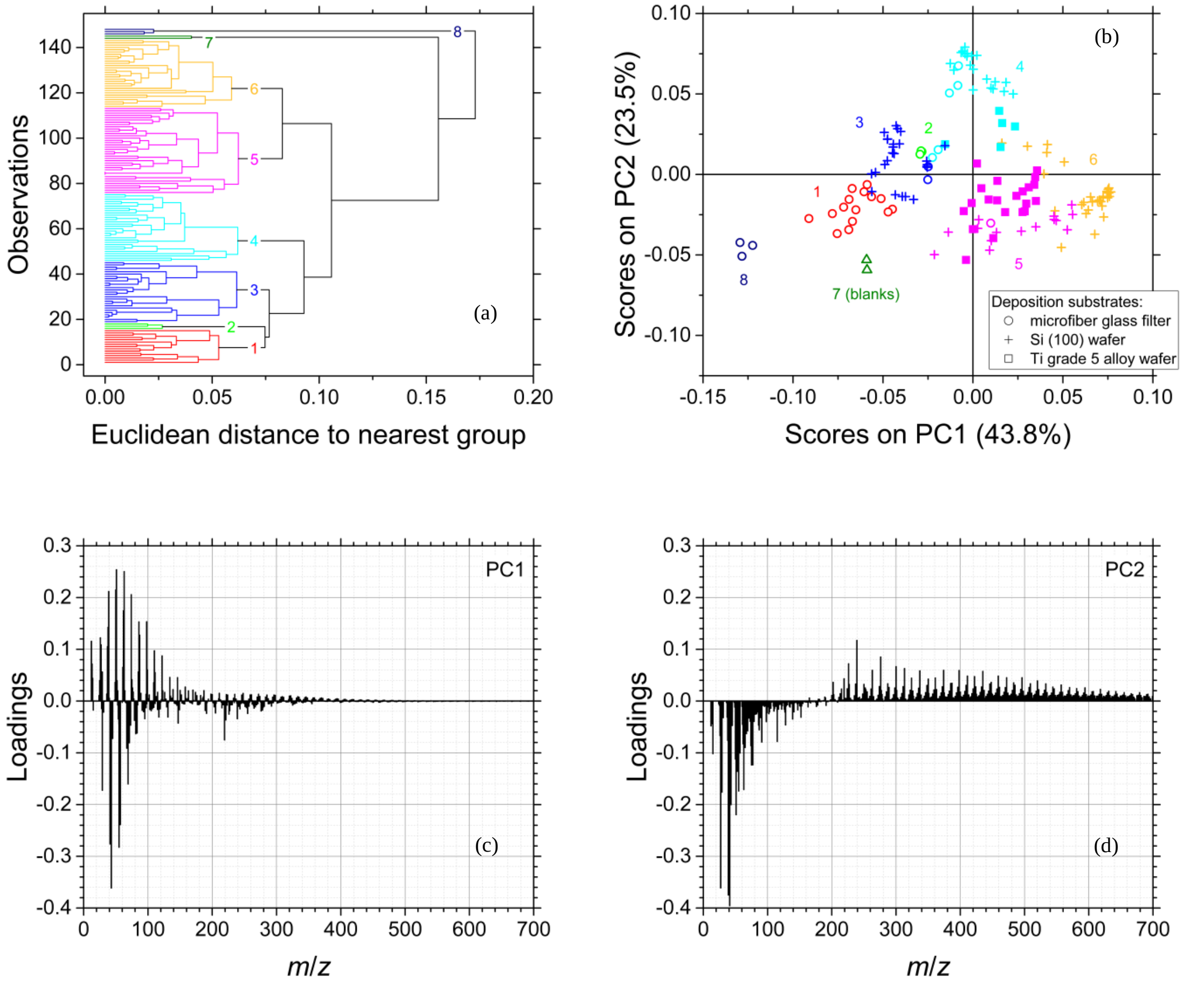


Fig. S1. Comparison of HCA and PCA. HCA dendrogram (a), score plots of the first two PCs (b), coefficients of PC1 (c) and PC2 (d). The colors and numbers match the clusters in the two representations. In the score plot, round dots represent depositions on microfiber glass filters, while crosses and square dots represent depositions on Si and Ti wafers, respectively.

Cluster number	PRE-N	PRE-S	DIF-Q	DIF-S
1	7 gf, 12 gf, 15 gf, 20 gf	7 gf	/	/
2	/	15 gf	/	/
3	/	7 Si, 9 Si, 12 Si, 15 Si **20 Si**, 20 gf, 40 gf	/	/
4	15 gf, 40 gf	9 gf, 12 gf, 15 gf, **12 Ti**	25 Si, **45 Si**, **55 Si**	48 Si, **65 Si**
5	7 Si, 9 Si, 12 Si, 15 Si, 15 gf, 40 Si, 40 Ti	12 Ti, 15 Ti, **15 Ti**, 20 Si, 40 Si, **40 Si**	/	/
6	7 Si, 9 Si, 12 Si, 15 Si, 20 Si	40 Si	55 Si	65 Si, 80 Si, **80 Si**
7 (blanks)	/	/	/	/
8	40 gf		/	/

Table S1. Classification of samples based on HCA and PCA. The flames are listed in the columns while the cluster numbers in the rows. Each cell contains the sampled HAB followed by a two letter group indicating the deposition substrate (gf: microfiber glass filters, Si and Ti: wafers). Bold, normal and underlined text mark the samples obtained from the analysis of the soot spot (if present), the wafer surface and siloxane-contaminated, respectively.

3. SI. Database of identified ions

The four tables in this section list the 771 signals built into the main peak list and sorted by *m/z* by HCA. The peak list includes identified fragment ions (Table S2), signals that could not be reliably distinguished from the blank (Table S3) and hydrocarbons detected on soot and condensable gas with PC1 < 0 (nucleation in premixed flames, Table S4) and PC1 > 0 (nucleation in diffusion flames and surface growth, Table S5).

m/z	Formula	*m/z*	Formula	*m/z*	Formula	*m/z*	Formula	*m/z*	Formula
12.000	C^+	62.016	$C_5H_2^+$	121.008	$C_{10}H^+$	170.016	$C_{14}H_2^+$	235.040	/
13.008	CH^+	63.023	$C_5H_3^+$	122.016	$C_{10}H_2^+$	171.023	$C_{14}H_3^+$	246.016	/
14.016	CH_2^+	64.031	$C_5H_4^+$	123.023	$C_{10}H_3^+$	173.039	$C_{14}H_5^+$	247.024	/
15.023	CH_3^+	73.008	C_6H^+	124.031	$C_{10}H_4^+$	174.047	$C_{14}H_6^+$	248.047	/
24.000	C_2^+	74.016	$C_6H_2^+$	125.039	$C_{10}H_5^+$	175.055	$C_{14}H_7^+$	257.009	/
25.008	C_2H^+	75.023	$C_6H_3^+$	126.047	$C_{10}H_6^+$	176.063	$C_{14}H_8^+$	258.016	/
26.016	$C_2H_2^+$	76.031	$C_6H_4^+$	134.016	$C_{11}H_2^+$	182.016	$C_{15}H_2^+$	259.040	/
27.023	$C_2H_3^+$	85.008	C_7H^+	135.023	$C_{11}H_3^+$	183.023	$C_{15}H_3^+$	260.047	/
29.003	CHO^+	86.016	$C_7H_2^+$	136.031	$C_{11}H_4^+$	185.039	$C_{15}H_5^+$	270.016	/
36.000	C_3^+	87.023	$C_7H_3^+$	137.039	$C_{11}H_5^+$	186.032	$C_{11}H_6O_3^+$	271.024	/
37.008	C_3H^+	88.031	$C_7H_4^+$	138.047	$C_{11}H_6^+$	187.055	$C_{15}H_7^+$	272.047	/
38.016	$C_3H_2^+$	89.039	$C_7H_5^+$	139.055	$C_{11}H_7^+$	188.063	$C_{15}H_8^+$	281.009	/
39.023	$C_3H_3^+$	97.008	C_8H^+	145.008	$C_{12}H^+$	194.016	$C_{16}H_2^+$	282.016	/
40.031	$C_3H_4^+$	98.016	$C_8H_2^+$	146.016	$C_{12}H_2^+$	198.032	$C_{12}H_6O_3^+$	283.024	/
41.003	C_2HO^+	99.023	$C_8H_3^+$	147.023	$C_{12}H_3^+$	199.040	$C_{12}H_7O_3^+$	284.032	/
42.011	$C_2H_2O^+$	100.031	$C_8H_4^+$	148.031	$C_{12}H_4^+$	200.063	$C_{16}H_8^+$	296.032	/
48.000	C_4^+	101.039	$C_8H_5^+$	150.047	$C_{12}H_6^+$	209.024	/	305.01	/
49.008	C_4H^+	102.047	$C_8H_6^+$	151.055	$C_{12}H_7^+$	210.032	/	306.02	/
50.016	$C_4H_2^+$	109.008	C_9H^+	158.016	$C_{13}H_2^+$	211.040	/	328.99	/
51.023	$C_4H_3^+$	110.016	$C_9H_2^+$	159.023	$C_{13}H_3^+$	212.047	/	343.01	/
52.031	$C_4H_4^+$	111.023	$C_9H_3^+$	160.031	$C_{13}H_4^+$	222.032	/		
53.003	C_3HO^+	112.031	$C_9H_4^+$	161.039	$C_{13}H_5^+$	223.040	/		
60.000	C_5^+	113.039	$C_9H_5^+$	162.047	$C_{13}H_6^+$	233.009	/		
61.008	C_5H^+	114.047	$C_9H_6^+$	163.055	$C_{13}H_7^+$	234.016	/		

Table S2. Peaks attributed to fragment ions by HCA.

m/z	Formula	*m/z*	Formula	*m/z*	Formula	*m/z*	Formula	*m/z*	Formula
28.031	$C_2H_4^+$	55.055	$C_4H_7^+$	69.070	$C_5H_9^+$	82.042	$C_5H_6O^+$	97.065	$C_6H_9O^+$
29.039	$C_2H_5^+$	56.063	$C_4H_8^+$	70.042	$C_4H_6O^+$	82.078	$C_6H_{10}^+$	98.073	$C_6H_{10}O^+$
31.018	CH_3O^+	57.034	$C_3H_5O^+$	70.078	$C_5H_{10}^+$	83.086	$C_6H_{11}^+$	109.065	$C_7H_9O^+$
33.034	CH_5O^+	57.070	$C_4H_9^+$	71.050	$C_4H_7O^+$	84.058	$C_5H_8O^+$	109.102	$C_8H_{13}^+$
41.039	$C_3H_5^+$	65.039	$C_5H_5^+$	71.086	$C_5H_{11}^+$	84.094	$C_6H_{12}^+$	111.081	$C_7H_{11}O^+$
42.047	$C_3H_6^+$	66.047	$C_5H_6^+$	77.039	$C_6H_5^+$	91.018	$C_6H_3O^+$	115.018	$C_8H_3O^+$
43.018	$C_2H_3O^+$	67.018	$C_4H_3O^+$	78.047	$C_6H_6^+$	91.055	$C_7H_7^+$	117.034	$C_8H_5O^+$
43.055	$C_3H_7^+$	67.055	$C_5H_7^+$	79.018	$C_5H_3O^+$	92.026	$C_6H_4O^+$	131.050	$C_9H_7O^+$
45.034	$C_2H_5O^+$	68.026	$C_4H_4O^+$	79.055	$C_6H_7^+$	92.063	$C_7H_8^+$	131.086	$C_{10}H_{11}^+$
53.039	$C_4H_5^+$	68.063	$C_5H_8^+$	80.063	$C_6H_8^+$	93.070	$C_7H_9^+$	149.039	$C_{12}H_5^+$
54.047	$C_4H_6^+$	68.998	$C_3HO_2^+$	81.034	$C_5H_5O^+$	94.078	$C_7H_{10}^+$	155.086	$C_{12}H_{11}^+$
55.018	$C_3H_3O^+$	69.034	$C_4H_5O^+$	81.070	$C_6H_9^+$	95.086	$C_7H_{11}^+$		

Table S3. Peaks that could not be reliably distinguished from the blank.

m/z	Formula	*m/z*	Formula	*m/z*	Formula	*m/z*	Formula	*m/z*	Formula
90.047	$C_7H_6^+$	108.058	$C_7H_8O^+$	141.070	$C_{11}H_9^+$	170.094	$C_9H_{14}O_3^+$	219.081	$C_{16}H_{11}O^+$
93.034	$C_6H_5O^+$	108.094	$C_8H_{12}^+$	142.078	$C_{11}H_{10}^+$	177.070	$C_{14}H_9^+$	220.089	$C_{16}H_{12}O^+$
94.042	$C_6H_6O^+$	115.055	$C_9H_7^+$	143.086	$C_{11}H_{11}^+$	178.078	$C_{14}H_{10}^+$	220.183	$C_{15}H_{24}O^+$
95.050	$C_6H_7O^+$	116.063	$C_9H_8^+$	144.094	$C_{11}H_{12}^+$	180.094	$C_{14}H_{12}^+$	230.110	$C_{18}H_{14}^+$
96.058	$C_6H_8O^+$	117.070	$C_9H_9^+$	145.065	$C_{10}H_9O^+$	182.073	$C_{13}H_{10}O^+$	231.102	/
96.094	$C_7H_{12}^+$	118.042	$C_8H_6O^+$	146.073	$C_{10}H_{10}O^+$	189.070	$C_{15}H_9^+$	234.104	/
103.018	$C_7H_3O^+$	118.078	$C_9H_{10}^+$	152.063	$C_{12}H_8^+$	190.078	$C_{15}H_{10}^+$	242.094	/
103.055	$C_8H_7^+$	119.050	$C_8H_7O^+$	153.070	$C_{12}H_9^+$	191.086	$C_{15}H_{11}^+$	243.102	/
104.026	$C_7H_4O^+$	119.086	$C_9H_{11}^+$	154.078	$C_{12}H_{10}^+$	192.094	$C_{15}H_{12}^+$	244.110	/
104.063	$C_8H_8^+$	120.094	$C_9H_{12}^+$	156.094	$C_{12}H_{12}^+$	204.094	$C_{16}H_{12}^+$	245.118	/
105.034	$C_7H_5O^+$	127.055	$C_{10}H_7^+$	157.065	$C_{11}H_9O^+$	206.073	$C_{15}H_{10}O^+$	256.089	/
105.070	$C_8H_9^+$	128.063	$C_{10}H_8^+$	164.063	$C_{13}H_8^+$	214.078	$C_{17}H_{10}^+$	257.097	/
106.042	$C_7H_6O^+$	129.070	$C_{10}H_9^+$	166.078	$C_{13}H_{10}^+$	216.094	$C_{17}H_{12}^+$	258.104	/
106.078	$C_8H_{10}^+$	130.078	$C_{10}H_{10}^+$	167.086	$C_{13}H_{11}^+$	217.102	$C_{17}H_{13}^+$	268.110	/
107.050	$C_7H_7O^+$	132.094	$C_{10}H_{12}^+$	168.079	$C_9H_{12}O_3^+$	218.094	$C_{13}H_{14}O_3^+$		
107.086	$C_8H_{11}^+$	140.063	$C_{11}H_8^+$	169.086	$C_9H_{13}O_3^+$	218.167	$C_{15}H_{22}O^+$		

Table S4. Peaks attributed to soot and condensable gas by HCA and PCA (nucleation in premixed flames, PC1 < 0, see the loadings plots in Fig. 7).

m/*z*	Formula	*m*/*z*	Formula	*m*/*z*	Formula	*m*/*z*	Formula	*m*/*z*	Formula
165.070	$C_{13}H_9^+$	325.10	$C_{26}H_{13}^+$	414.14	$C_{33}H_{18}^+$	508.13	$C_{41}H_{16}^+$	601.0	/
179.086	$C_{14}H_{11}^+$	326.11	$C_{26}H_{14}^+$	415.01	/	509.13	$C_{41}H_{17}^+$	602.0	/
181.065	$C_{13}H_9O^+$	327.12	$C_{26}H_{15}^+$	415.13	/	510.14	$C_{41}H_{18}^+$	603.1	/
201.070	$C_{16}H_9^+$	328.13	$C_{26}H_{16}^+$	416.02	/	511.15	$C_{41}H_{19}^+$	604.1	/
202.078	$C_{16}H_{10}^+$	329.12	/	416.14	/	512.16	$C_{41}H_{20}^+$	605.1	$C_{49}H_{17}^+$
203.086	$C_{16}H_{11}^+$	330.00	/	417.03	/	513.16	$C_{41}H_{21}^+$	606.1	$C_{49}H_{18}^+$
205.065	$C_{15}H_9O^+$	330.13	/	418.03	/	514.02	/	607.1	$C_{49}H_{19}^+$
213.070	$C_{17}H_9^+$	331.01	/	419.04	/	515.03	/	608.2	$C_{49}H_{20}^+$
215.086	$C_{17}H_{11}^+$	331.13	/	420.05	/	516.03	/	609.2	$C_{49}H_{21}^+$
217.159	$C_{15}H_{21}O^+$	332.02	/	421.07	/	517.04	/	610.2	$C_{49}H_{22}^+$
219.175	$C_{15}H_{23}O^+$	333.04	/	422.11	$C_{34}H_{14}^+$	518.08	/	611.2	$C_{49}H_{23}^+$
221.118	/	334.05	/	423.12	$C_{34}H_{15}^+$	520.13	$C_{42}H_{16}^+$	612.0	/
222.126	/	335.07	/	424.13	$C_{34}H_{16}^+$	521.13	$C_{42}H_{17}^+$	612.2	$C_{49}H_{24}^+$
223.097	/	335.18	$C_{26}H_{23}^+$	425.13	$C_{34}H_{17}^+$	522.14	$C_{42}H_{18}^+$	613.0	/
224.047	/	336.08	/	426.14	$C_{34}H_{18}^+$	523.15	$C_{42}H_{19}^+$	614.0	/
225.055	/	337.10	$C_{27}H_{13}^+$	427.01	/	524.16	$C_{42}H_{20}^+$	615.1	/
226.078	$C_{18}H_{10}^+$	338.11	$C_{27}H_{14}^+$	427.15	$C_{34}H_{19}^+$	525.01	/	616.1	/
227.086	$C_{18}H_{11}^+$	339.12	$C_{27}H_{15}^+$	428.02	/	525.16	$C_{42}H_{21}^+$	617.1	/
228.094	$C_{18}H_{12}^+$	340.13	$C_{27}H_{16}^+$	428.16	$C_{34}H_{20}^+$	526.02	/	618.1	$C_{50}H_{18}^+$
229.102	$C_{18}H_{13}^+$	341.12	/	429.03	/	526.17	$C_{42}H_{22}^+$	619.1	$C_{50}H_{19}^+$
236.047	/	342.00	/	430.03	/	527.03	/	620.2	$C_{50}H_{20}^+$
237.070	$C_{19}H_9^+$	342.13	/	431.04	/	528.03	/	621.2	$C_{50}H_{21}^+$
238.078	$C_{19}H_{10}^+$	343.13	/	432.05	/	529.06	/	622.2	$C_{50}H_{22}^+$
239.086	$C_{19}H_{11}^+$	344.02	/	433.07	/	530.06	/	623.0	/
240.094	$C_{19}H_{12}^+$	345.04	/	435.12	$C_{35}H_{15}^+$	531.10	/	623.2	$C_{50}H_{23}^+$
241.102	$C_{19}H_{13}^+$	345.15	/	436.13	$C_{35}H_{16}^+$	532.11	/	624.0	/
246.104	$C_{18}H_{14}O^+$	346.03	/	437.13	$C_{35}H_{17}^+$	533.13	$C_{43}H_{17}^+$	624.2	$C_{50}H_{24}^+$
247.112	/	346.17	$C_{27}H_{22}^+$	438.14	$C_{35}H_{18}^+$	534.14	$C_{43}H_{18}^+$	625.0	/
249.055	/	347.06	/	439.15	$C_{35}H_{19}^+$	535.15	$C_{43}H_{19}^+$	626.0	/
250.078	$C_{20}H_{10}^+$	348.08	/	440.02	/	536.16	$C_{43}H_{20}^+$	627.0	/
251.086	$C_{20}H_{11}^+$	349.09	/	442.03	/	537.16	$C_{43}H_{21}^+$	629.1	/
252.094	$C_{20}H_{12}^+$	350.11	$C_{28}H_{14}^+$	443.04	/	538.02	/	630.1	/
253.102	$C_{20}H_{13}^+$	351.12	$C_{28}H_{15}^+$	444.05	/	538.17	$C_{43}H_{22}^+$	631.1	$C_{51}H_{19}^+$
254.110	$C_{20}H_{14}^+$	352.13	$C_{28}H_{16}^+$	445.06	/	539.03	/	632.2	$C_{51}H_{20}^+$
255.081	/	353.13	$C_{28}H_{17}^+$	446.06	/	539.18	$C_{43}H_{23}^+$	633.2	$C_{51}H_{21}^+$
261.055	/	354.14	$C_{28}H_{18}^+$	447.12	$C_{36}H_{15}^+$	540.03	/	634.2	$C_{51}H_{22}^+$
262.063	/	354.29	$C_{25}H_{38}O^+$	448.13	$C_{36}H_{16}^+$	541.04	/	635.2	$C_{51}H_{23}^+$
262.230	$C_{18}H_{30}O^+$	355.01	/	449.13	$C_{36}H_{17}^+$	542.05	/	636.0	/
263.086	$C_{21}H_{11}^+$	355.13	/	450.14	$C_{36}H_{18}^+$	543.06	/	636.2	$C_{51}H_{24}^+$
264.094	$C_{21}H_{12}^+$	355.30	$C_{25}H_{39}O^+$	451.01	/	544.13	$C_{44}H_{16}^+$	637.0	/
264.245	$C_{18}H_{32}O^+$	356.02	/	451.15	$C_{36}H_{19}^+$	545.13	$C_{44}H_{17}^+$	637.2	$C_{51}H_{25}^+$
265.102	$C_{21}H_{13}^+$	357.03	/	452.14	/	546.14	$C_{44}H_{18}^+$	638.0	/
265.253	$C_{18}H_{33}O^+$	358.05	/	453.01	/	547.15	$C_{44}H_{19}^+$	639.0	/

266.110	$C_{21}H_{14}^+$	359.04	/	454.02	/	548.16	$C_{44}H_{20}^+$	640.1	/
267.117	$C_{21}H_{15}^+$	360.06	/	455.03	/	549.16	$C_{44}H_{21}^+$	641.1	/
269.118	/	361.10	$C_{29}H_{13}^+$	456.03	/	550.17	$C_{44}H_{22}^+$	642.1	$C_{52}H_{18}^+$
270.126	/	362.11	$C_{29}H_{14}^+$	457.06	/	551.03	/	643.1	$C_{52}H_{19}^+$
271.112	/	363.12	$C_{29}H_{15}^+$	458.06	/	551.18	$C_{44}H_{23}^+$	644.2	$C_{52}H_{20}^+$
273.055	/	364.13	$C_{29}H_{16}^+$	459.12	$C_{37}H_{15}^+$	552.03	/	645.2	$C_{52}H_{21}^+$
274.078	$C_{22}H_{10}^+$	365.13	$C_{29}H_{17}^+$	460.13	$C_{37}H_{16}^+$	552.19	$C_{44}H_{24}^+$	646.2	$C_{52}H_{22}^+$
275.086	$C_{22}H_{11}^+$	366.00	/	461.13	$C_{37}H_{17}^+$	553.04	/	647.2	$C_{52}H_{23}^+$
276.094	$C_{22}H_{12}^+$	366.13	/	462.14	$C_{37}H_{18}^+$	554.05	/	648.2	$C_{52}H_{24}^+$
277.102	$C_{22}H_{13}^+$	367.01	/	463.15	$C_{37}H_{19}^+$	555.06	/	649.0	/
278.110	$C_{22}H_{14}^+$	368.02	/	464.02	/	556.06	/	651.1	/
278.261	$C_{19}H_{34}O^+$	369.03	/	464.16	$C_{37}H_{20}^+$	557.13	$C_{45}H_{17}^+$	652.1	/
279.117	$C_{22}H_{15}^+$	370.03	/	465.03	/	558.14	$C_{45}H_{18}^+$	653.1	$C_{53}H_{17}^+$
279.269	$C_{19}H_{35}O^+$	371.04	/	465.16	$C_{37}H_{21}^+$	559.15	$C_{45}H_{19}^+$	654.1	$C_{53}H_{18}^+$
285.055	/	372.06	/	466.02	/	560.16	$C_{45}H_{20}^+$	655.1	$C_{53}H_{19}^+$
286.063	/	373.07	/	467.03	/	561.16	$C_{45}H_{21}^+$	656.2	$C_{53}H_{20}^+$
287.086	$C_{23}H_{11}^+$	374.11	$C_{30}H_{14}^+$	468.03	/	562.02	/	657.2	$C_{53}H_{21}^+$
287.180	$C_{22}H_{23}^+$	375.12	$C_{30}H_{15}^+$	469.07	/	562.17	$C_{45}H_{22}^+$	658.2	$C_{53}H_{22}^+$
288.094	$C_{23}H_{12}^+$	376.13	$C_{30}H_{16}^+$	470.09	/	563.03	/	659.2	$C_{53}H_{23}^+$
289.102	$C_{23}H_{13}^+$	377.13	$C_{30}H_{17}^+$	471.10	/	563.18	$C_{45}H_{23}^+$	660.0	/
290.110	$C_{23}H_{14}^+$	378.14	$C_{30}H_{18}^+$	472.13	$C_{38}H_{16}^+$	564.03	/	660.2	$C_{53}H_{24}^+$
290.261	$C_{20}H_{34}O^+$	379.01	/	473.13	$C_{38}H_{17}^+$	565.04	/	662.0	/
291.117	$C_{23}H_{15}^+$	379.13	/	474.14	$C_{38}H_{18}^+$	566.05	/	664.1	/
292.277	$C_{20}H_{36}O^+$	380.03	/	475.15	$C_{38}H_{19}^+$	567.06	/	665.1	/
293.118	/	381.03	/	476.16	$C_{38}H_{20}^+$	568.08	/	666.1	$C_{54}H_{18}^+$
293.284	$C_{20}H_{37}O^+$	382.03	/	477.01	/	569.09	/	667.1	$C_{54}H_{19}^+$
294.016	/	383.06	/	477.16	$C_{38}H_{21}^+$	570.14	$C_{46}H_{18}^+$	668.2	$C_{54}H_{20}^+$
294.126	/	384.06	/	478.02	/	571.15	$C_{46}H_{19}^+$	669.2	$C_{54}H_{21}^+$
295.024	/	385.10	$C_{31}H_{13}^+$	478.17	$C_{38}H_{22}^+$	572.16	$C_{46}H_{20}^+$	670.2	$C_{54}H_{22}^+$
295.133	/	386.11	$C_{31}H_{14}^+$	479.03	/	573.16	$C_{46}H_{21}^+$	671.2	$C_{54}H_{23}^+$
297.040	/	387.12	$C_{31}H_{15}^+$	481.06	/	574.17	$C_{46}H_{22}^+$	672.2	$C_{54}H_{24}^+$
298.063	/	388.13	$C_{31}H_{16}^+$	482.06	/	575.03	/	675.1	/
299.071	/	389.13	$C_{31}H_{17}^+$	483.12	$C_{39}H_{15}^+$	575.18	$C_{46}H_{23}^+$	677.1	/
300.094	$C_{24}H_{12}^+$	390.00	/	484.13	$C_{39}H_{16}^+$	576.03	/	679.1	$C_{55}H_{19}^+$
301.10	$C_{24}H_{13}^+$	390.13	/	485.13	$C_{39}H_{17}^+$	577.04	/	680.2	$C_{55}H_{20}^+$
302.11	$C_{24}H_{14}^+$	391.01	/	486.14	$C_{39}H_{18}^+$	578.05	/	681.2	$C_{55}H_{21}^+$
303.12	$C_{24}H_{15}^+$	391.13	/	487.15	$C_{39}H_{19}^+$	579.07	/	682.2	$C_{55}H_{22}^+$
304.13	$C_{24}H_{16}^+$	392.02	/	488.02	/	580.08	/	683.2	$C_{55}H_{23}^+$
305.12	/	393.03	/	488.16	$C_{39}H_{20}^+$	581.13	$C_{47}H_{17}^+$	686.0	/
306.13	/	394.03	/	489.03	/	582.14	$C_{47}H_{18}^+$	688.0	/
307.02	/	395.04	/	489.16	$C_{39}H_{21}^+$	583.15	$C_{47}H_{19}^+$	690.1	/
307.13	/	396.06	/	490.02	/	584.16	$C_{47}H_{20}^+$	691.1	/
308.03	/	397.07	/	491.04	/	585.16	$C_{47}H_{21}^+$	692.2	$C_{56}H_{20}^+$
308.14	/	398.11	$C_{32}H_{14}^+$	492.05	/	586.17	$C_{47}H_{22}^+$	693.2	$C_{56}H_{21}^+$

309.04	/	399.12	$C_{32}H_{15}^+$	493.06	/	587.03	/	694.2	$C_{56}H_{22}^+$
310.06	/	400.13	$C_{32}H_{16}^+$	494.06	/	587.18	$C_{47}H_{23}^+$	695.2	$C_{56}H_{23}^+$
311.07	/	401.13	$C_{32}H_{17}^+$	495.09	/	588.03	/	696.2	$C_{56}H_{24}^+$
312.09	$C_{25}H_{12}^+$	402.14	$C_{32}H_{18}^+$	496.13	$C_{40}H_{16}^+$	589.04	/	699.1	/
313.10	$C_{25}H_{13}^+$	403.01	/	497.13	$C_{40}H_{17}^+$	590.05	/	701.1	/
314.11	$C_{25}H_{14}^+$	403.15	$C_{32}H_{19}^+$	498.14	$C_{40}H_{18}^+$	591.06	/	703.1	$C_{57}H_{19}^+$
315.12	$C_{25}H_{15}^+$	404.02	/	499.15	$C_{40}H_{19}^+$	592.08	/	704.2	$C_{57}H_{20}^+$
316.13	$C_{25}H_{16}^+$	405.03	/	500.16	$C_{40}H_{20}^+$	593.09	/	705.2	$C_{57}H_{21}^+$
317.13	$C_{25}H_{17}^+$	406.03	/	501.03	/	594.14	$C_{48}H_{18}^+$	706.2	$C_{57}H_{22}^+$
318.13	/	407.04	/	501.16	$C_{40}H_{21}^+$	595.15	$C_{48}H_{19}^+$	707.2	$C_{57}H_{23}^+$
319.01	/	408.05	/	502.03	/	596.16	$C_{48}H_{20}^+$	716.2	$C_{58}H_{20}^+$
319.13	/	409.09	/	502.17	$C_{40}H_{22}^+$	597.16	$C_{48}H_{21}^+$	717.2	$C_{58}H_{21}^+$
320.02	/	410.09	/	503.03	/	598.17	$C_{48}H_{22}^+$	718.2	$C_{58}H_{22}^+$
321.03	/	411.12	$C_{33}H_{15}^+$	504.03	/	599.03	/	729.2	$C_{59}H_{21}^+$
322.05	/	412.13	$C_{33}H_{16}^+$	505.06	/	599.18	$C_{48}H_{23}^+$	731.2	$C_{59}H_{23}^+$
323.06	/	413.13	$C_{33}H_{17}^+$	506.06	/	600.03	/		
324.09	$C_{26}H_{12}^+$	414.00	/	507.10	/	600.2	$C_{48}H_{24}^+$		

Table S5. Peaks attributed to soot and condensable gas by HCA and PCA (nucleation in premixed flames and surface growth, PC1 > 0, see the loadings plots in Fig. 7).

4. SI. References

[87] K.K. Murray, R.K. Boyd, M.N. Eberlin, G.J. Langley, L. Li, Y. Naito, Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013), Pure Appl Chem. 85 (2013) 1515–1609.

[88] M.T. Timko, Z. Yu, J. Kroll, J.T. Jayne, D.R. Worsnop, R.C. Miake-Lye, T.B. Onasch, D. Liscinsky, T.W. Kirchstetter, H. Destaillats, A.L. Holder, J.D. Smith, K.R. Wilson, Sampling artifacts from conductive silicone tubing, Aerosol Sci. Technol. 43 (2009) 855–865. doi:10.1080/02786820902984811.

[89] Y. Yu, M. Liz Alexander, V. Perraud, E.A. Bruns, S.N. Johnson, M.J. Ezell, B.J. Finlayson-Pitts, Contamination from electrically conductive silicone tubing during aerosol chemical analysis, Atmos. Environ. 43 (2009) 2836–2839. doi:10.1016/j.atmosenv.2009.02.014.