Supplementary information

Table S1: Comparison between IUPAC and Kendrick mass of a series of aromatic and aliphatic compounds, commonly present on soot particles.

Chemical formula	IUPAC mass	Kendrick mass
Aliphatic compounds		
C_8H_{18}	114.14085	114.01339
C_9H_{20}	128.15650	128.01339
$C_{10}H_{22}$	142.17215	142.01339
$C_{11}H_{24}$	156.18780	156.01339
$C_{12}H_{26}$	170.20345	170.01339
$C_{13}H_{28}$	184.21910	184.01339
Aromatic compounds		
$C_{12}H_{10}$	154.07825	153.90620
$C_{13}H_{10}$	166.07825	165.89280
$C_{14}H_{10}$	178.07825	177.87940
$C_{15}H_{10}$	190.07825	189.86600
$C_{16}H_{10}$	202.07825	201.85260
$C_{18}H_{10}$	226.07825	225.82581

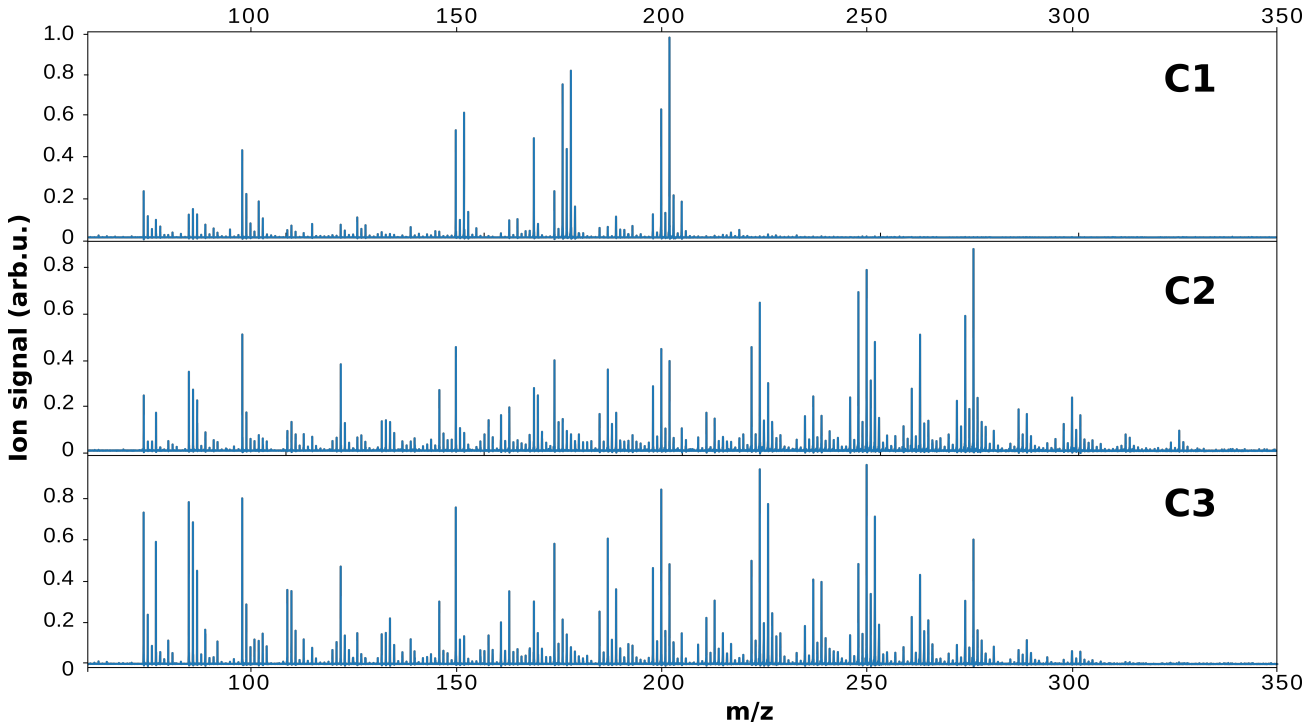

Figure S1: Comparison of mass spectra of miniCAST soot samples obtained with L2MS. All the spectra were normalised to the highest intensity peak, for visual purposes only.

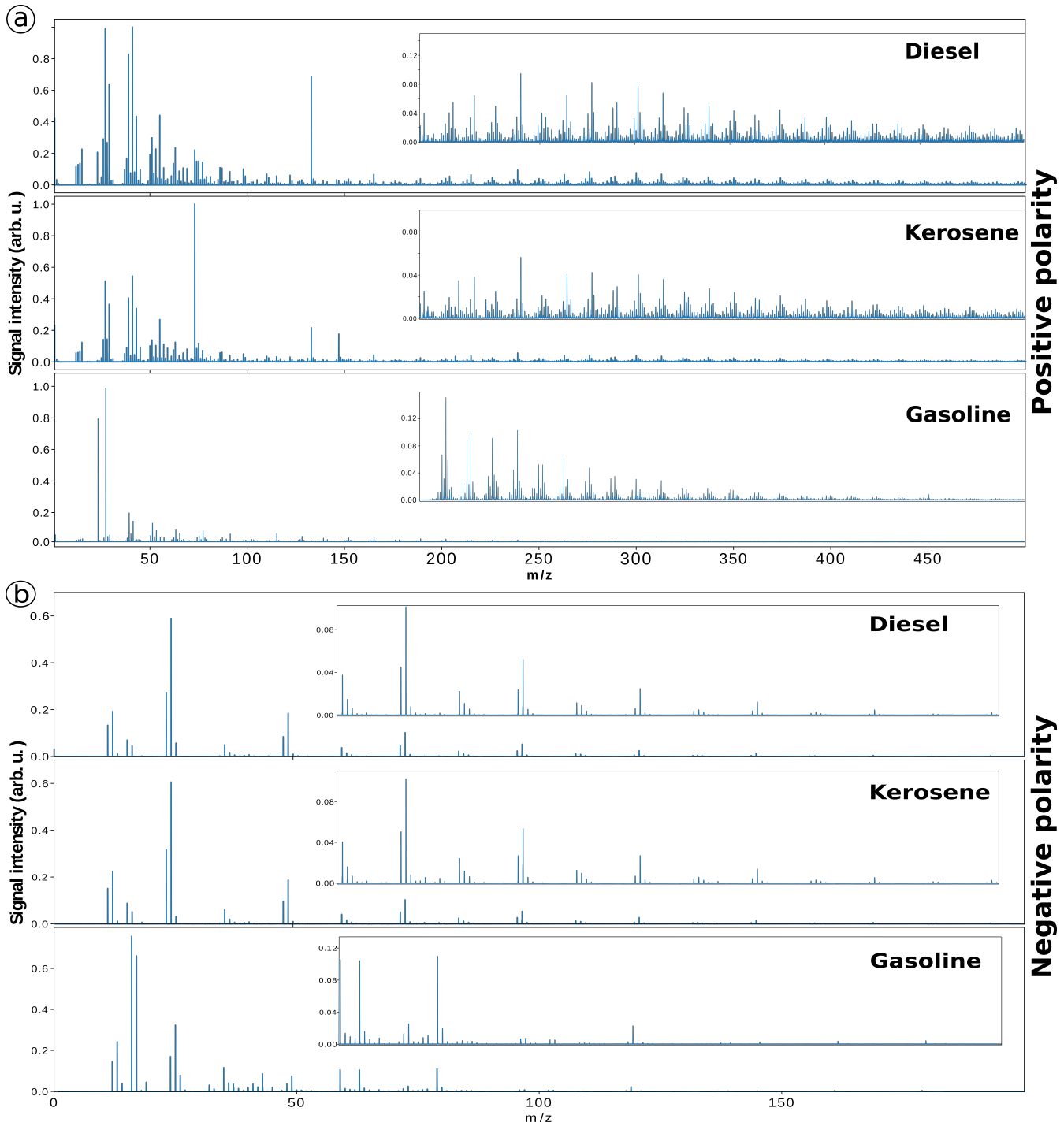

Figure S2: Comparison between mass spectra of diesel (D2), kerosene (K2) and gasoline (GOM1) soot obtained with SIMS in positive (a) and negative polarity (b). All the spectra were normalised to the highest intensity peak, for visual purpose only.

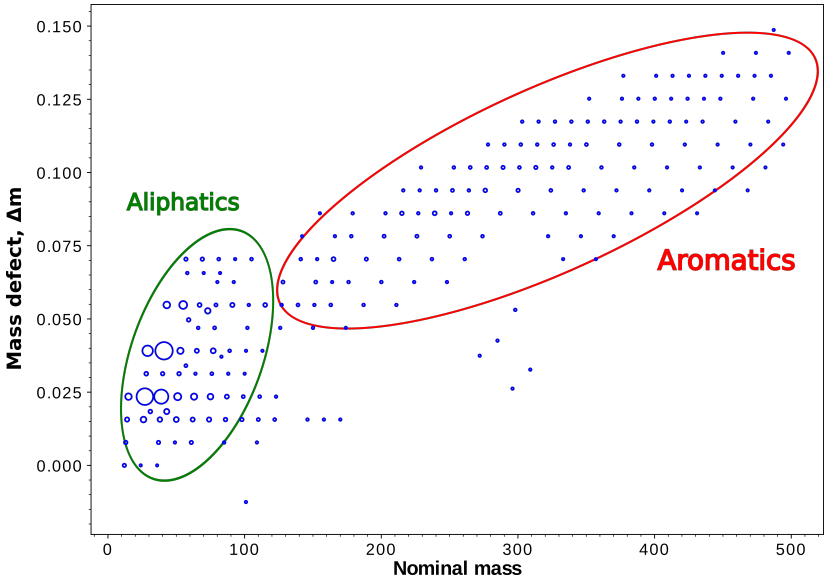

Figure S3: Mass defect plot obtained from the mass spectrum of soot generated by laboratory diesel flame. The size of the data points is proportional to the corresponding peak area, normalised to total ion count. Mass spectra obtained from SIMS contain many fragment ions (mainly in the green ellipse). Fragment ions typically cluster on recognisable patterns on the mass defect plot. On the other hand, the red (aromatics) and green (aliphatics and fragments) ellipses in clearly exhibit different slopes, showcasing the helpfulness of this method in sorting chemical classes in the overcrowded SIMS spectra.

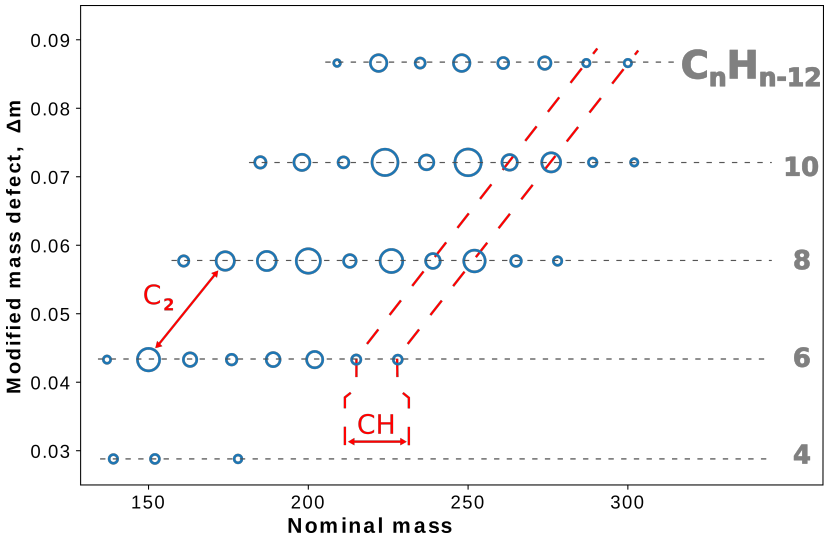

Figure S4: Kendrick mass defect plot obtained from the mass spectrum of miniCAST soot, *C2* sample. *CH* was taken as the base unit. For the ease of visualisation, only hydrocarbons are shown. The size of the data points is proportional to the corresponding peak area normalised to the total ion count.

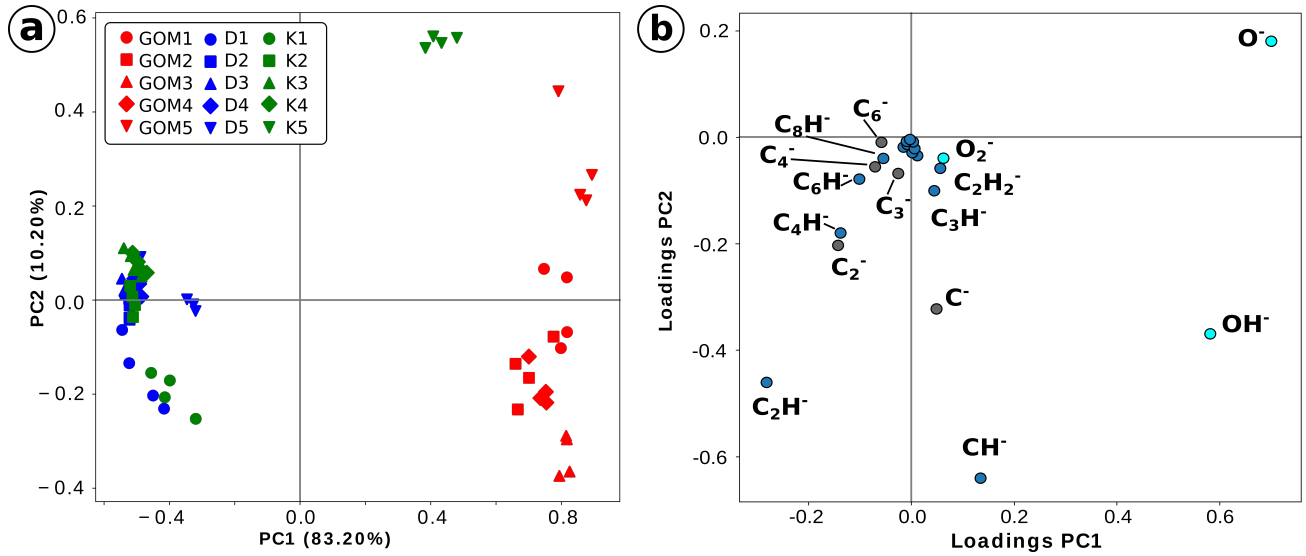

Figure S5: Score plot of PC2 vs PC1 for negative ions of soot samples obtained from gasoline engine and laboratory flames – (a). Corresponding loadings plot of the first two principal components – (b). For sample description see Table 1.

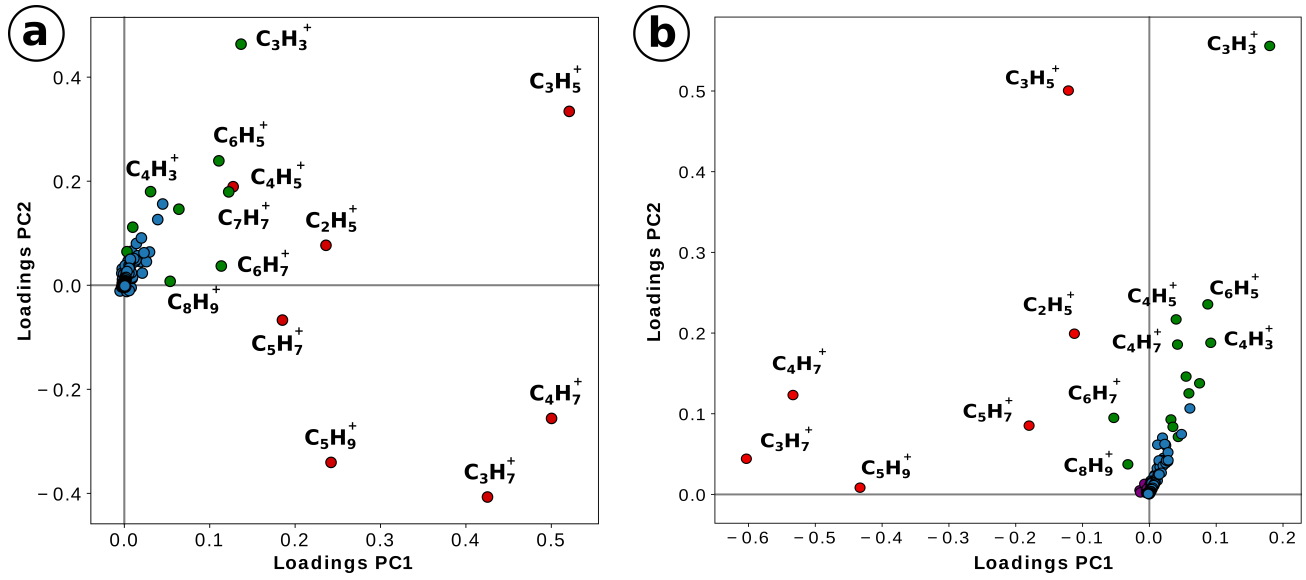

Figure S6: Loadings corresponding to the contribution of different species to PC1 and PC2 derived from mass spectra of soot samples produced by a single cylinder engine. (a) – Analysis of GOM, GOH, GEF and GEO samples, (b) – GEF and GEO samples. Aliphatic species are shown in red, fragments - in green, aromatic species - in blue and high mass high H/C compounds in purple.

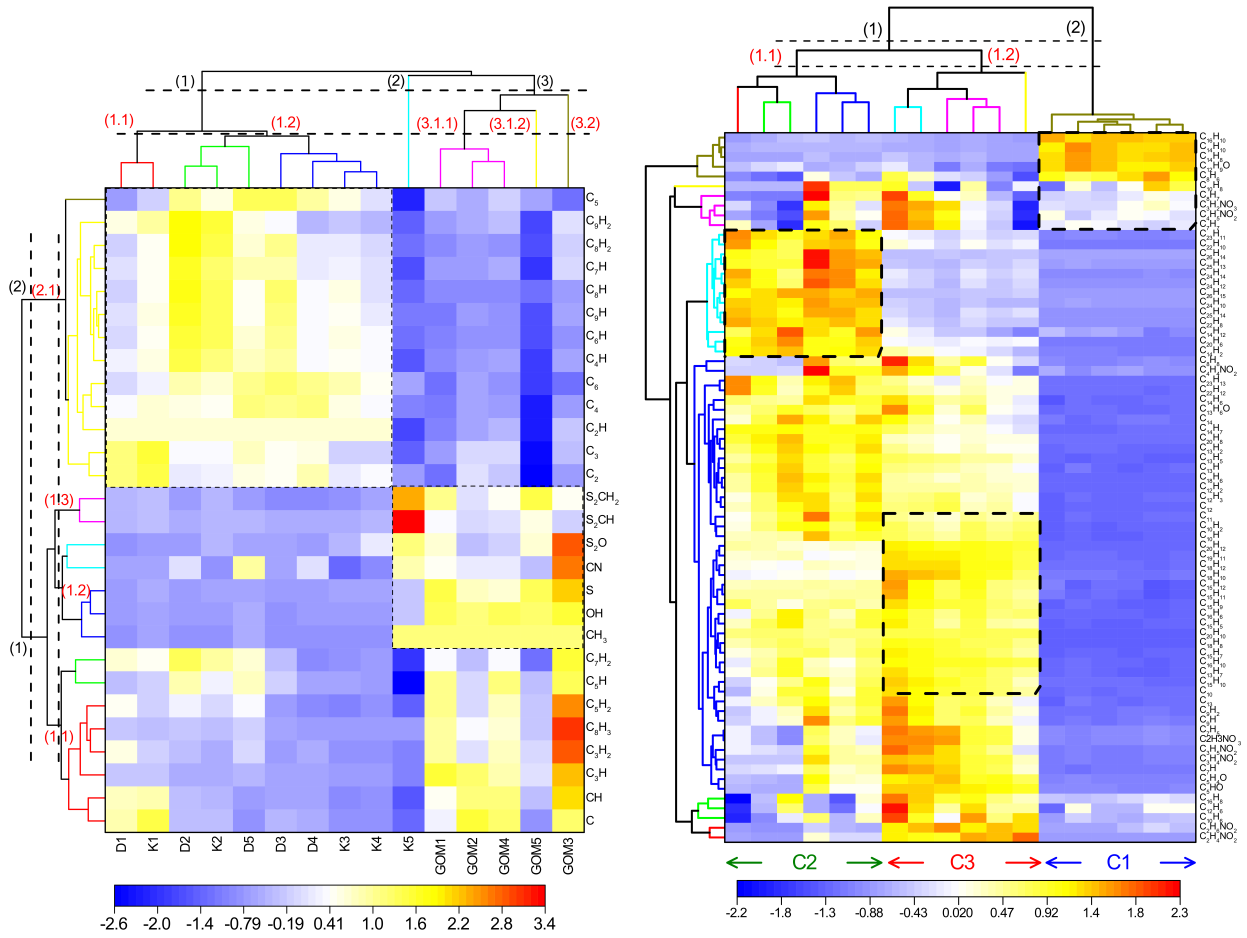

Figure S7: Two-way hierarchical clustering heat-map for negative ions of gasoline, diesel and kerosene soot obtained with SIMS (left panel) and of positive L2MS ions of CAST soot sample (right panel). Each column corresponds to the averaged mass spectra obtained for a soot sample. The contribution of each mass in individual samples is expressed as relative value and is represented by the cell colour.

HCA is also applied to L2MS and SIMS negative polarity data while using the same mathematical parameters (Figure S7). The HCA on the negative polarity SIMS finds dissimilarities between the soot collected from the flame and the one from the gasoline engine, with indicators to a higher organic content for the flame soot and a higher content of oxygen and sulphur containing compounds for the soot particles emitted by the gasoline engine. The heatmap obtained for the L2MS (Figure S7) is grouping the CAST samples in three clusters specific to each sample. The classification is mostly the same as obtained with PCA, where C1 samples are dominated by low mass PAHs (from two to four aromatic rings), C3 samples have the maximum contribution of five number rings aromatics and C2 samples have the highest contribution of aromatics with high m/z.