Supplementary information for

Phase-Resolved Molecular Signatures Associated with Black Carbon across Biomass Fuels and Ignition Temperatures

Xiaocong Peng[1,#], Wei Song[1,2,3,#], Guohua Zhang[1,2,3,*], Qinhao Lin[4], Jianlin Cheng[1,5], Yang Li[1,5], Xinru Huang[1], Baoyi Chen[1,5], Jianzhong Song[1,2,3], Yingjun Chen[6], Gan Zhang[1,2,3], Ping'an Peng[1,2,3], Xinming Wang[1,2,3], Xinhui Bi[1,2,3]

[1] State Key Laboratory of Environmental Technology, Guangzhou Institute of Geochemistry, Chinese Academy of Sciences (CAS), Guangzhou 510640, PR China

[2] Guangdong Provincial Key Laboratory of Environmental Protection and Resources Utilization, Guangzhou Institute of Geochemistry, CAS, Guangzhou 510640, PR China

[3] Guangdong-Hong Kong-Macao Joint Laboratory for Environmental Pollution and Control, Guangzhou Institute of Geochemistry, CAS, Guangzhou 510640, PR China

[4] School of Environmental Science and Engineering, Institute of Environmental Health and Pollution Control, Guangdong University of Technology, Guangzhou 510006, PR China

[5] University of Chinese Academy of Sciences, Beijing 100049, PR China

[6] Shanghai Key Laboratory of Atmospheric Particle Pollution and Prevention (LAP3), Department of Environmental Science and Engineering, Fudan University, Shanghai 200438, PR China

#These authors contributed equally to this work.

Corresponding to Guohua Zhang (zhanggh@gig.ac.cn)

Contents

25

Text S1. Two-Level selection of core BC-linked molecular families and representative nominal-m/z markers

At the first level, all detected species are grouped into eight broad chemical classes based on molecular structure and functional group identity. Within each event block, defined by fuel, phase, replicate experiment, and ignition temperature, Spearman rank correlations were calculated between BC and the class-summed molecular signal. The heterogeneous residual category "others" was excluded a priori from mechanistic family selection.

Four classes passed the complete screen: oxygenated aromatic (median $\rho = 0.92$), single-ring aromatic (SRA; $\rho = 0.91$), polycyclic aromatic hydrocarbon (PAH; $\rho = 0.91$), and furans ($\rho = 0.86$). The association direction was positive in all particle-phase blocks for oxygenated aromatic, SRA, and PAHs, and in 90% of particle-phase blocks for furans. O-containing and CxHy signals also showed strong family-level coupling, but neither class contained at least two marker candidates that satisfied the complete-coverage and marker-level criteria. Carbonyl and N-containing classes did not meet the family-level correlation and/or directional-consistency requirements. The complete class-level evaluation is reported in Table S1.

A nominal-m/z feature was considered an eligible marker candidate when it (i) yielded valid block-level correlations in all 21 event blocks and represented all four fuel–phase conditions; (ii) showed a particle-phase median Spearman correlation of at least 0.70; and (iii) was positively correlated with BC in at least 75% of the 10 particle-phase blocks. Eligible candidates were ranked separately within each family using the equal-weight mean of their within-family percentile ranks for particle-phase median ρ, particle-phase positive-correlation fraction, and median particle-phase relative abundance. The three highest-ranked candidates in each family were retained.

This procedure yielded 12 representative markers: furan (m/z 69), 2(3H)-furanone (m/z 85), and 2-pentylfuran (m/z 139); phenol (m/z 95), methyl benzoate (m/z 137), and m-thymol (m/z 151); ethylbenzene (m/z 107), 4-tert-butyltoluene (m/z 149), and 3,3-dimethyl-2-phenylbutene (m/z 161); and acenaphthylene (m/z 153), acenaphthene (m/z 155), and 2,6-dimethylnaphthalene (m/z 157). Because PTR-MS does not distinguish structural isomers, these names are putative assignments to nominal-m/z features rather than unique molecular identifications. The full metrics and within-family rankings are reported in Table S2 and the associated candidate-audit table.

The stability of the family classification was examined over 189 combinations of the family-level correlation threshold, directional-consistency threshold, and minimum eligible-species requirement. Under the implemented threshold grid, the same four-family solution was recovered in all combinations, with a mean Jaccard similarity of 1.000 relative to the reference solution. This analysis evaluates sensitivity to the numerical family thresholds; it does not test alternative chemical assignments or differences in instrumental feature coverage.

Text S2. Ridge regression validation of phase-resolved molecular information in independently withheld combustion experiments

Grouped statistical-learning models were used to test whether the molecular profiles provided BC-relevant information beyond ignition temperature, fuel identity, and MCE. In particle-phase data, the Ridge baseline model yielded an out-of-experiment R^2 of 0.29 and an MAE of 0.86. Adding the four core-family signals increased R^2 to 0.56 and reduced MAE to 0.64 (Figure S5). The 12-marker model also improved prediction ($R^2 = 0.48$), whereas models based on total VOCs abundance, broad chemical classes, or all annotated m/z features performed less well. The core-family model reduced prediction error in each of the four independently withheld particle experiments. PLS regression produced a comparable improvement, with R^2 increasing from 0.19 to 0.55. In gas-phase data, the baseline model yielded an R^2 of −0.002, and the addition of the four core families reduced model performance to $R^2 = -0.077$. Thus, the predictive information associated with the selected molecular families was phase specific.

A within-block circular-shift permutation test was used to determine whether the particle-phase gain resulted solely from common temperature or combustion-stage trends. The observed improvement, $\Delta R^2=0.27$, exceeded the permuted null distribution ($p = 0.0099$). However, after aggregation to experiment–temperature medians, the core-family model no longer improved prediction. The statistical-learning results therefore validate the reproducibility of within-event particle-phase molecular–BC coupling but do not establish direct marker conversion or condition-level BC yield control.

Table S1. Multi-criteria evaluation of VOC classes for BC-associated marker selection. Eligible species count denotes nominal-m/z features that had valid correlations in all 21 event blocks and all four fuel–phase conditions, an unambiguous class assignment and putative name, particle-phase median $\rho \geq 0.70$, and positive association in at least 75% of particle-phase blocks. Effective blocks denotes the number of event blocks with valid marker-BC correlations.

Chemical Class	Particle pahase ρ	Gas phase ρ	Eligible species count	Effective blocks
Furans	**0.86**	0.44	**8**	**21/21**
Oxygenated aromatic	**0.92**	0.27	**7**	**21/21**
SRA	**0.91**	-0.1	**5**	**21/21**
PAHs	**0.91**	0.04	**4**	**21/21**
N-containing	0.121	0.43	-	21/21
O-containing	**0.906**	0.48	**0**	13/21
CxHy	**0.750**	0.46	**0**	13/21
Carbonyl	-0.231	0.51	-	21/21

Table S2. Twelve representative nominal-m/z markers selected from the four core BC-linked molecular families. All-block positive fractions denote the proportions of positive correlations across the 10 particle-phase blocks and all 21 blocks, respectively.

Class	Species	Effective blocks	Particle positive fraction %	Median ρ	All-block positive fraction %
Furans	Furan (*m/z* 69)	21	80	0.77	81
Furans	2(3H)-Furanone (*m/z* 85)	21	80	0.81	57
Furans	2-Pentylfuran (*m/z* 139)	21	100	0.81	76
Oxy-aromatic	Phenol (*m/z* 95)	21	100	0.75	90
Oxy-aromatic	*m*-Thymol (*m/z* 151)	21	100	0.85	76
Oxy-aromatic	Methyl benzoate (*m/z* 137)	21	100	0.82	76
PAH	Acenaphthylene (*m/z* 153)	21	100	0.83	86
PAH	Acenaphthene(*m/z* 155)	21	80	0.86	71
PAH	2,6-Dimethylnaphthalene (*m/z* 157)	21	100	0.84	76
SRA	Ethylbenzene (*m/z* 107)	21	100	0.80	86
SRA	4-*tert*-Butyltoluene (*m/z* 149)	21	100	0.89	67
SRA	3,3-Dimethyl-2-phenylbutene (*m/z* 161)	21	100	0.86	76

Figure S1. Absolute VOC class loading in gas- and particle-phase emissions from wheat straw and poplar wood across ignition temperatures. (a) Gas-phase cumulative VOC class loading and (b) particle-phase cumulative VOC class loading for wheat straw and poplar wood at 300, 500, and 700 °C. For each fuel–phase–temperature condition, VOC class loading was calculated by summing the signal of all retained ions within each chemical class across time-resolved observations within each replicate block, followed by taking the median across replicate blocks. Stacked bars show the absolute contributions of furans, oxygenated aromatic (O-arom.), single-ring aromatic (SRA), PAH, carbonyl, O-containing, CxHy, N-containing, and other compounds.

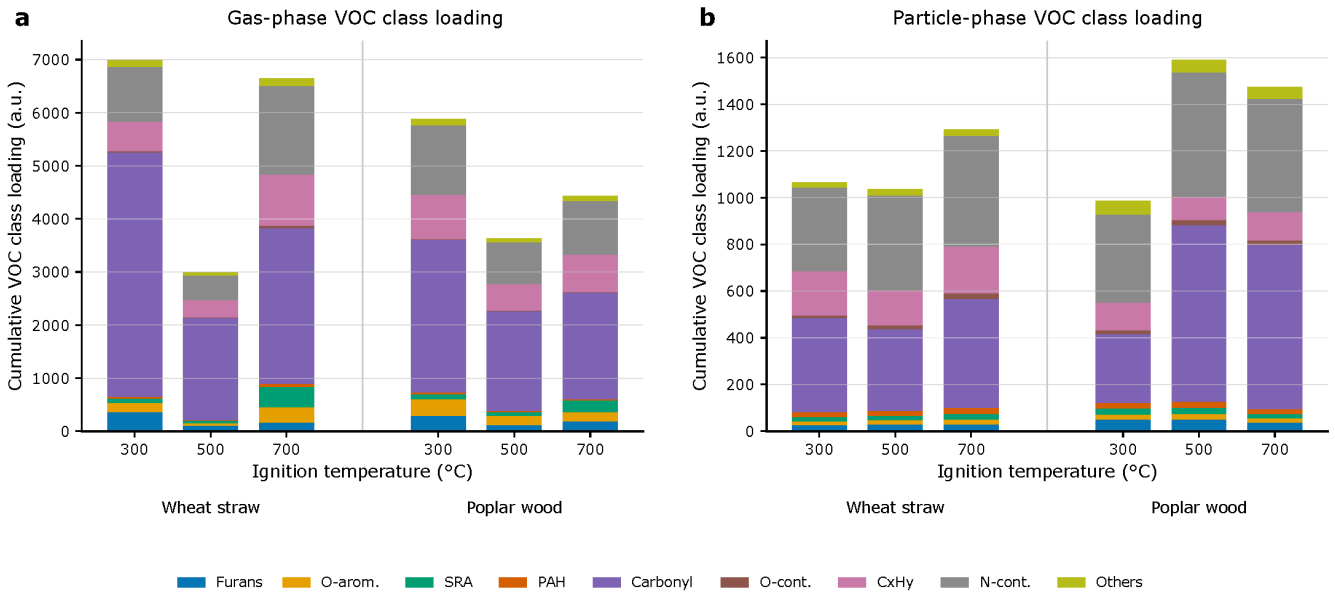

Figure S2. Gas-phase four-family core-marker composition across fuel type and ignition temperature. Cumulative gas-phase marker signals assigned to the four core marker families, including furans, oxygenated aromatic (O-arom.), single-ring aromatic (SRA), and polycyclic aromatic hydrocarbon (PAH), are shown for wheat straw and poplar wood at 300, 500, and 700 °C. For each fuel–temperature condition, signals were summed across rows within each replicate block and normalized to the summed four-family core-marker pool; bars show the median percentage composition across replicate blocks.

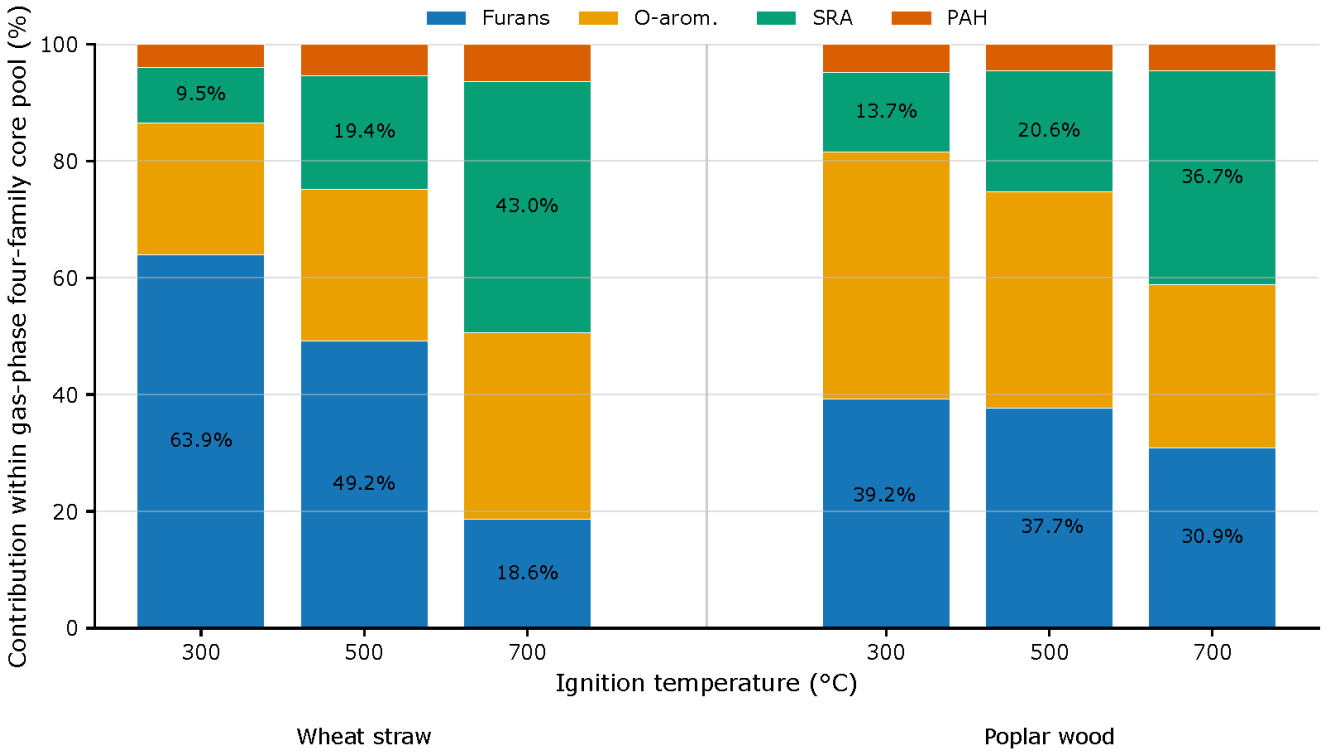

Figure S3. Particle-phase composition of the four-family core marker pool. Relative contributions of furans, O-arom., SRA, and PAH marker families to the summed particle-phase four-family core pool for wheat straw and poplar wood at 300, 500, and 700 °C. For each fuel–temperature condition, marker-family signals were first summed across rows within each replicate block and then normalized to the summed four-family core-marker signal; bars show the median percentage composition across replicate blocks.

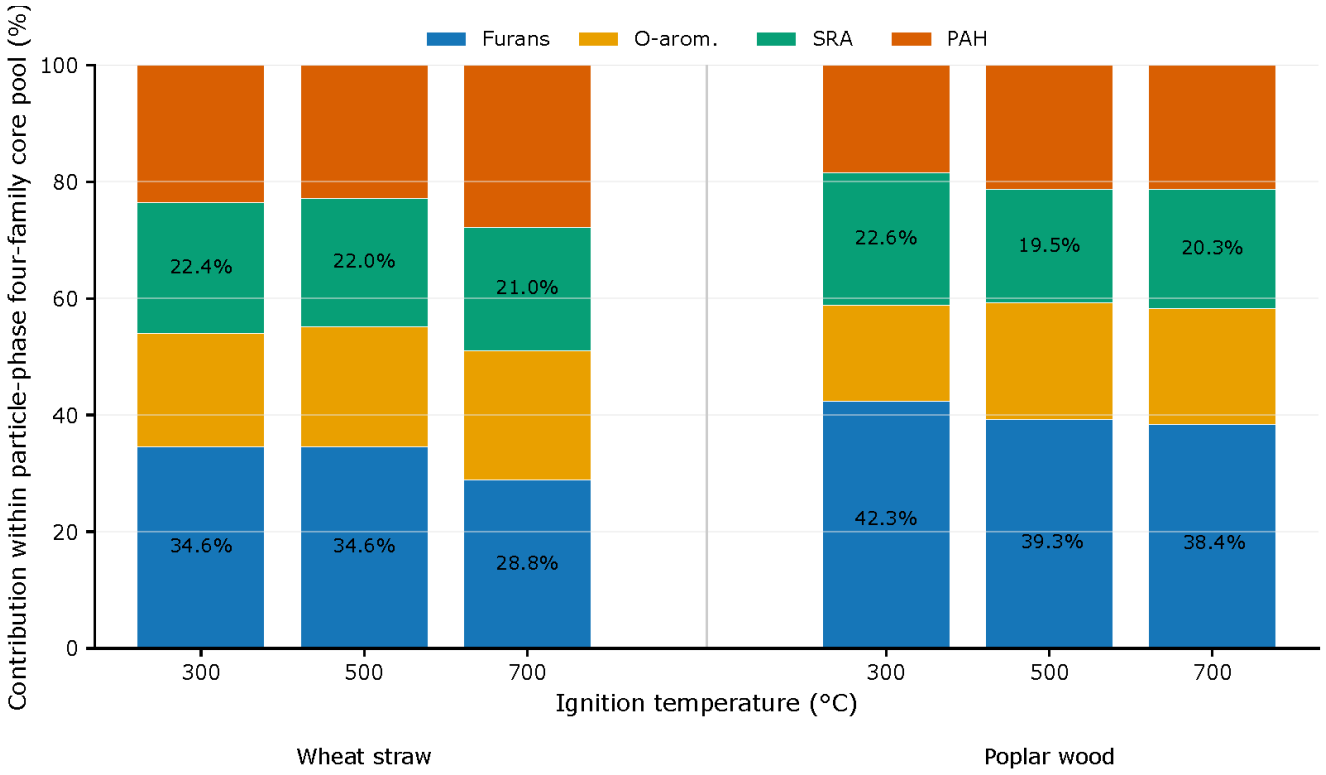

Figure S4. Signal coverage and correlation robustness of the 12 selected core markers. (a) Particle-phase median signal intensity of each selected marker. (b) Particle-phase median relative abundance of each marker, normalized to the total retained VOC signal. (c) Comparison of gas-phase and particle-phase median block-level Spearman correlations with BC for individual markers. The dashed line indicates equal gas- and particle-phase correlations. (d) Positive-correlation fraction and scaled block coverage for each marker. This figure evaluates whether the selected markers are supported by measurable signal strength, phase-resolved BC coupling, and cross-block consistency.

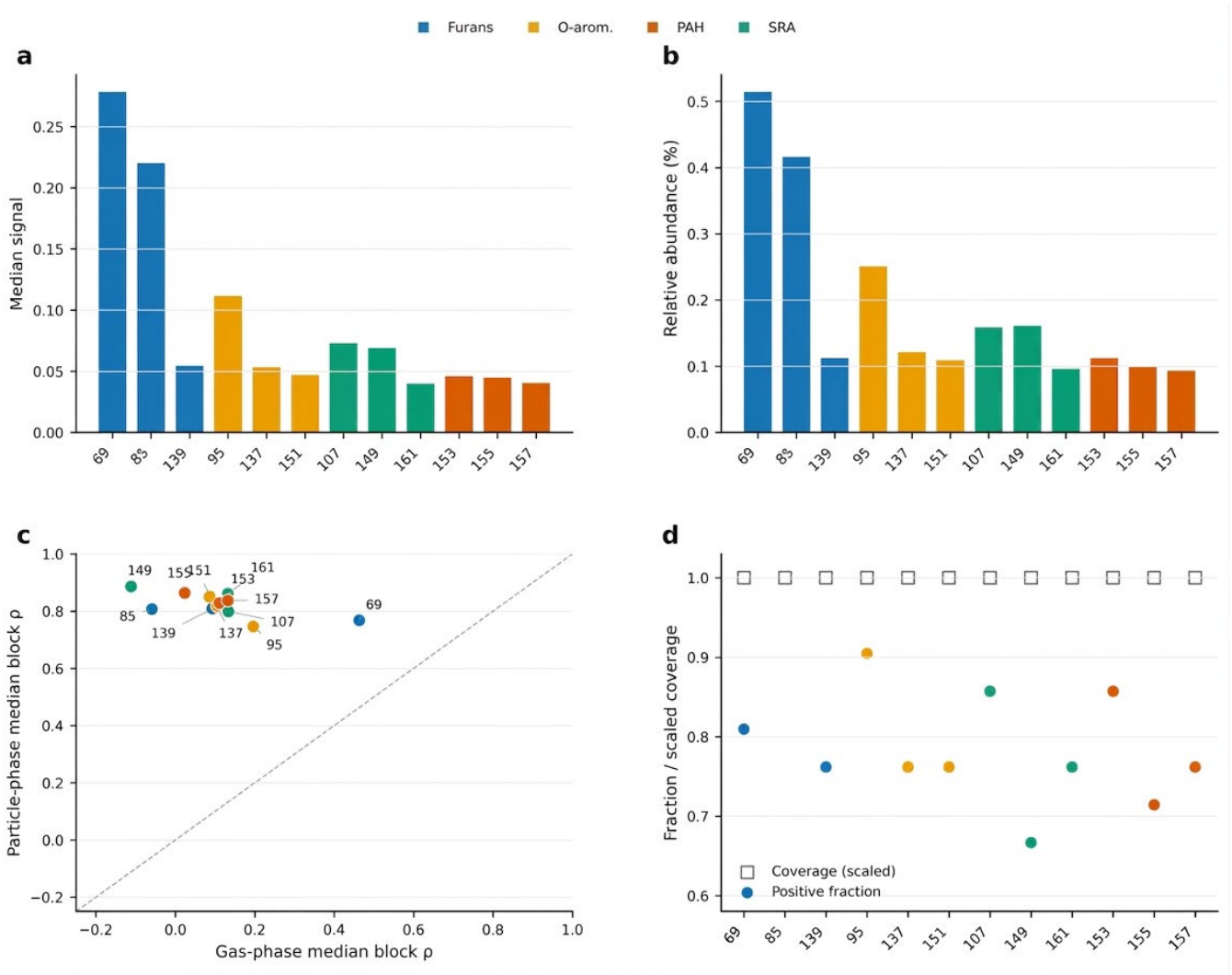

Figure S5. Model performance was evaluated using Ridge regression with nested leave-one-experiment-out cross-validation. Partial least-squares regression was used as an independent sensitivity analysis. (a) Out-of-experiment coefficients of determination (R^2) for Ridge regression models applied separately to gas- and particle-phase data. The baseline model included ignition temperature, fuel identity, and modified combustion efficiency; molecular models additionally included total VOCs signal, broad chemical classes, four core marker families, 12 selected markers, or all retained annotated ions. (b) Change in mean absolute prediction error after adding the four core-family signals to the baseline model for each independently withheld combustion experiment. Positive values indicate reduced prediction error. (c) Null distribution of the particle-phase ΔR^2 generated by circularly shifting the four core-family profiles within each experiment–temperature block. The dashed line indicates the observed ΔR^2 for the unshifted data.