Supplement of:

Molecular fingerprints reveal traffic $PM_{2.5}$ complexity beyond a controlled gasoline tailpipe molecular profile

Lijuan Li et al.

Correspondence to: Yuemei Han (yuemei.han@ieecas.cn); Junji Cao (jjcao@mail.iap.ac.cn)

Text S1. Reference guided screening of candidate formulas associated with tire wear compounds

Candidate tire wear related molecular formulas were screened from formula level HRMS peak area time series in the 28 tunnel samples. The reference formulas were N-(1,3-dimethylbutyl)-N′-phenyl-p-phenylenediamine (6PPD, $C_{18}H_{24}N_2$), 2-aminobenzothiazole (ABT, $C_7H_6N_2S$), and 2-phenylbenzothiazole (PhBT, $C_{13}H_9NS$) for the Primary TWP group, and 6PPD quinone ($C_{18}H_{22}N_2O_2$) and 2-hydroxybenzothiazole (2-OH-BTH, C_7H_5NOS) (Seiwert et al., 2022; Klöckner et al., 2020). Candidate selection was restricted to formula assignments retained after blank correction and detected in at least 10 tunnel samples. Candidates were required to correlate with at least one reference formula (Pearson $r \geq 0.80$ for the Primary TWP group or $r \geq 0.75$ for the Secondary TWP group) and to satisfy CH_2 based KMD coherence (KMD difference $< 3 \times 10^{-3}$; at least three members) and elemental composition criteria. The Primary TWP group required $O/C \leq 0.20$ and $H/C \leq 1.40$. The Secondary TWP group required $N \geq 1$ and $O/C > 0.05$. This procedure retained 62 candidate formulas associated with the Primary TWP references and 24 associated with the Secondary TWP references. These formula level associations were used for HCA annotation and to define candidates for targeted confirmation with authentic standards and MS/MS spectra.

Text S2. Volcano screening and multiple-testing checks

Volcano analysis was performed separately in ESI+ and ESI− to identify molecular formula assignments that differed between tunnel aerosol and the controlled single vehicle gasoline tailpipe profile and to define the input for HCA (Anders and Huber, 2010). Nondetected entries were set to zero before normalization. For each sample, blank-corrected peak areas were divided by the summed peak area of all assigned formulas in the same ionization mode to obtain within-mode relative peak areas (RA). Two-sided Welch's t tests were applied to $\log_{10}(RA + \varepsilon)$, where $\varepsilon = 10^{-9}$ (Welch, 1947), and $\log_2$ fold changes were calculated from the arithmetic group means of the untransformed RA values as

$$\log_2 \mathrm{FC} = \log_2 \mathrm{FC} \left(\frac{\mathrm{median_{tunnel}}}{\mathrm{median_{bench}}}\right) \tag{S2-1}$$

where positive values indicate higher mean within-mode relative peak areas in the tunnel samples. The fixed pseudocount ε was used solely to avoid logarithms and ratios involving zero; applying the same ε to both group means yielded finite $\log_2$FC values for T–O and CD–O assignments detected in only one dataset. Statistical testing was restricted to formula and ionization mode assignments detected in at least 10 of the 28 tunnel samples. T–O and CD–O were assigned from the occurrence classification and were evaluated at their calculated finite $\log_2 FC$ positions. Assignments with $p < 0.05$, $|\log_2 FC| \geq 2$, and detection in at least 10 tunnel samples defined the HCA input. Benjamini–Hochberg q values were calculated separately within each ionization mode (Benjamini and Hochberg, 1995). All 709 assignments used for HCA met $q < 0.10$, and 698 met $q < 0.05$ (Table S4). The 11 assignments with $0.05 \leq q < 0.10$ occurred in C1, while every assignment in C2 and C3 met $q < 0.05$. CD–O formulas did not enter HCA because they lacked tunnel temporal profiles. Occurrence based category labels were added after clustering and did not contribute to the distance matrix.

Text S3. Hierarchical clustering model selection and α sensitivity

HCA used complete linkage agglomeration and a composite distance matrix that combined temporal variation in tunnel peak area with molecular composition. Peak areas were normalized separately within each ESI mode using median ratio size factors (Anders and Huber, 2010). Nondetected values were set to zero after size factor estimation, and normalized peak areas were log1p transformed. For each formula and ionization mode assignment, transformed peak areas were averaged across the four 3 h sampling periods on each sampling day to construct a temporal profile containing seven daily values in chronological sampling order. The profile was smoothed using a moving average over three consecutive days with replicated endpoints, and temporal dissimilarity was calculated using cosine distance. The chemical descriptor matrix contained O/C, H/C, and carbon number, and chemical dissimilarity was calculated as cosine distance from the descriptor values on their native scales. Temporal and chemical distance vectors were independently rescaled to [0, 1] before weighting. The composite distance metric was constructed for this study, and its sensitivity to K, α, and descriptor scaling is reported in Tables S5 and S9 (Strehl et al., 2000; Han et al., 2022).

$$D = \alpha D_{time} + (1 - \alpha) D_{chem} \tag{S3-1}$$

where D_{time} is the temporal profile distance, D_{chem} is the molecular descriptor distance, and α controls the relative weight of temporal and chemical information.

Values of α from 0.20 to 0.90 and K from 3 to 6 were evaluated. Solutions were screened using a negative silhouette fraction ≤ 18% and a minimum domain size ≥ 21 assignments.(Rousseeuw, 1987) The K = 3, α = 0.80 solution produced domain sizes of 585, 74, and 50 assignments after relabeling to C1, C2, and C3, respectively. Its mean silhouette coefficient was 0.622, the median silhouette coefficient was 0.729, and the negative silhouette fraction was 4.2% (Table S5). At α = 0.70 and 0.90, 97.5%

and 97.6% of assignments, respectively, agreed with the $\alpha = 0.80$ partition after label matching, with adjusted Rand indices of 0.895 and 0.901. At $K = 4$ and $\alpha = 0.80$, C1 and C3 were retained, while C2 was divided into groups of 59 and 15 assignments. The 15 assignment group did not meet the prespecified minimum domain size of 21. The $K = 3$ solution was therefore retained as the higher level partition, and the $K = 4$ composition is reported in Table S8.

The primary chemical descriptor set contained O/C, H/C, and carbon number. Raw descriptor values were retained so that absolute carbon number remained part of the chemical distance. The analysis was repeated after column wise z score and min–max scaling (Table S9). At $K = 3$ and $\alpha = 0.80$, z score scaling gave a mean silhouette coefficient of 0.246, an adjusted Rand index of 0.463 relative to the raw descriptor partition, and 82.5% mapped assignment agreement. Min–max scaling gave a mean silhouette coefficient of 0.243, an adjusted Rand index of 0.499, and 83.9% mapped agreement. The raw descriptor formulation gave the highest silhouette coefficient and retained the absolute carbon number coordinate used in the structural analysis.

The influence of the dominant $C_4H_{11}N$ assignment on C3 was evaluated by removing this assignment and repeating HCA with the selected $K = 3$, $\alpha = 0.80$ specification (Table S10). $C_4H_{11}N$ contributed 41.3% of the summed normalized C3 peak area in ESI+. After its exclusion, all 49 remaining C3 assignments remained together. The repeated analysis gave mean and median silhouette coefficients of 0.620 and 0.724, respectively, with a negative silhouette fraction of 4.2%. The C3 membership was therefore retained after removal of the dominant low molecular mass CHN response.

Temporal variability in domain level HRMS response was summarized separately within ESI+ and ESI−. For each domain and ionization mode, normalized peak areas were summed across member assignments for each of the 28 tunnel samples. The coefficient of variation was calculated as $CV = 100 \times SD/mean$ from these sample level domain sums (Table S11). These values describe temporal variation in the normalized HRMS response and were not used for HCA model selection.

Text S4. Statistical evaluation of the persistent tunnel core and molecular profile variability

CoreFrac was calculated for each sample as the summed peak area of the fixed persistent core formulas divided by the summed peak area of all assigned formulas in the same ionization mode. The fixed core contained 156 formulas in ESI+ and 25 formulas in ESI−. Formula number, summed HRMS peak area, and CoreFrac were evaluated using ordinary one-way ANOVA. The statistical unit was one 3 h tunnel $PM_{2.5}$ filter sample. Sampling-period comparisons contained seven samples per group. Comparisons among the three sampled campaign days included four sampling-period samples from each of the rainy, cloudy, and sunny days. Because each condition was represented by only one campaign day, these tests quantify differences among the three sampled days rather than replicated weather-category effects. Median-centered Levene tests gave $p \geq 0.203$ for all comparisons in Tables S2 and S3, supporting use of the common-variance ANOVA for these balanced groups.

Table S1. Summary of formula numbers and molecular properties for emergent and attenuated formulas in each speed pair under ESI− and ESI+ modes. Values are mean ± standard deviation.

Mode	Speed comparison	Family	Category	Number	m/z	OSc	DBE/C
ESI-	20 vs 0 km/h	CHO	Emergent	8	221±104	-0.74±0.68	0.39±0.21
			Attenuated	57	216±79	-0.84±0.50	0.34±0.20
		CHN	Emergent	15	202±131	-1.10±0.35	0.80±0.21
			Attenuated	7	323±161	-1.04±0.36	0.71±0.33
		CHON	Emergent	73	202±82	-0.61±0.53	0.70±0.23
			Attenuated	48	296±108	-0.39±0.83	0.59±0.31
		CHOS	Emergent	9	358±131	-0.88±0.85	0.25±0.25
			Attenuated	45	273±88	-0.27±0.83	0.32±0.30
		CHONS	Emergent	18	338±75	-0.40±0.64	0.73±0.25
			Attenuated	45	422±107	-0.52±0.84	0.59±0.35
		Other	Emergent	8	372±94	-1.18±0.49	0.50±0.28
			Attenuated	9	412±38	-1.54±0.25	0.30±0.13
	40 vs 0 km/h	CHO	Emergent	19	272±120	-0.96±0.45	0.34±0.19
			Attenuated	39	222±81	-0.75±0.53	0.33±0.22
		CHN	Emergent	23	168±70	-1.14±0.30	0.79±0.20
			Attenuated	7	333±162	-1.05±0.36	0.70±0.34
		CHON	Emergent	164	199±80	-0.55±0.53	0.72±0.24
			Attenuated	41	320±112	-0.48±0.87	0.58±0.33
		CHOS	Emergent	16	323±90	-1.05±0.50	0.32±0.31
			Attenuated	43	271±95	-0.19±0.86	0.35±0.34
		CHONS	Emergent	32	344±99	-0.90±0.80	0.45±0.33
			Attenuated	56	403±97	-0.71±0.94	0.51±0.38
		Other	Emergent	14	334±130	-1.49±0.57	0.38±0.32
			Attenuated	13	400±106	-1.52±0.47	0.32±0.26
ESI+	20 vs 0 km/h	CHO	Emergent	46	283±111	-1.04±0.48	0.28±0.16
			Attenuated	80	287±110	-0.96±0.54	0.29±0.23
		CHN	Emergent	16	262±110	-1.33±0.31	0.45±0.17
			Attenuated	43	250±119	-1.53±0.48	0.38±0.25
		CHON	Emergent	83	323±142	-1.28±0.45	0.31±0.21
			Attenuated	205	342±129	-1.19±0.45	0.32±0.20
		CHOS	Emergent	30	359±137	-1.60±0.17	0.13±0.09

Mode	Speed comparison	Family	Category	Number	m/z	OSc	DBE/C
			Attenuated	72	349±120	-1.31±0.57	0.23±0.21
		CHONS	Emergent	63	363±124	-1.54±0.61	0.18±0.22
			Attenuated	78	418±131	-1.30±0.62	0.29±0.29
		Other	Emergent	45	259±103	-1.86±0.34	0.20±0.19
			Attenuated	33	305±126	-1.50±0.52	0.35±0.29
	40 vs 0 km/h	CHO	Emergent	84	321±121	-1.10±0.42	0.30±0.16
			Attenuated	72	290±117	-1.02±0.39	0.30±0.20
		CHN	Emergent	40	212±91	-1.46±0.36	0.44±0.20
			Attenuated	40	256±101	-1.54±0.47	0.36±0.24
		CHON	Emergent	157	306±144	-1.16±0.55	0.36±0.25
			Attenuated	171	348±108	-1.28±0.39	0.28±0.16
		CHOS	Emergent	72	343±124	-1.45±0.46	0.18±0.18
			Attenuated	51	334±102	-1.41±0.47	0.21±0.17
		CHONS	Emergent	112	367±130	-1.49±0.54	0.20±0.22
			Attenuated	68	404±118	-1.56±0.45	0.16±0.19
		Other	Emergent	74	244±99	-1.84±0.36	0.22±0.18
			Attenuated	37	314±102	-1.57±0.47	0.30±0.27

Table S2. Descriptive statistics (mean ± SD) and ordinary one-way ANOVA results (F and p) for the core fraction (CoreFrac = core peak area / total peak area) across four diurnal periods and three weather conditions, reported separately for ESI+ and ESI− modes.

Campaign day	Mode		Category means ± SD	ANOVA F	p
Time-of-Day	ESI+	M	0.398 ± 0.058	1.59	0.219
		N	0.415 ± 0.040		
		AF	0.404 ± 0.036		
		T	0.367 ± 0.036		
	ESI−	M	0.286 ± 0.026	1.77	0.18
		N	0.275 ± 0.053		
		AF	0.310 ± 0.042		
		T	0.246 ± 0.077		
Weather	ESI+	Rainy	0.392 ± 0.042	0.57	0.585
		Cloudy	0.388 ± 0.046		
		Sunny	0.414 ± 0.012		
	ESI−	Rainy	0.254 ± 0.063	1.44	0.286
		Cloudy	0.286 ± 0.036		
		Sunny	0.311 ± 0.037		

Table S3. Formula number and summed blank-corrected HRMS peak area by sampling period and sampled campaign day. Values are mean ± SD; F and p are from ordinary one-way ANOVA. The statistical unit was one 3 h tunnel $PM_{2.5}$ filter sample. Sampling-period comparisons included seven samples per period; comparisons among the sampled campaign days included four samples from each of the rainy, cloudy, and sunny days.

Factor	**Mode**	**Category means ± SD**	**ANOVA F**	**p**
		Formula No		
Time-of-Day	ESI+	M: 935 ± 196 N: 1109 ± 126 AF: 1064 ± 65 T: 1134 ± 133	2.87	0.057
	ESI−	M: 304 ± 98 N: 373 ± 109 AF: 341 ± 110 T: 433 ± 113	1.80	0.174
Weather	ESI+	Rainy: 1026 ± 119 Cloudy: 1064 ± 196 Sunny: 1170 ± 46	1.21	0.343
	ESI−	Rainy: 309 ± 65 Cloudy: 379 ± 78 Sunny: 284 ± 69	1.93	0.201
		Total peak area		
Time-of-Day	ESI+ ($\times10^{10}$)	M: 1.1 ± 0.5 N: 1.4 ± 0.3 AF: 1.3 ± 0.2 T: 1.3 ± 0.2	0.82	0.497
	ESI− ($\times10^{8}$)	M: 7.4 ± 3.0 N: 8.8 ± 3.0 AF: 7.6 ± 3.1 T: 10.8 ± 5.9	1.08	0.375
Weather	ESI^{+} ($\times10^{10}$)	Rainy: 1.1 ± 0.2 Cloudy: 1.1 ± 0.2 Sunny: 1.7 ± 0.2	10.27	0.005
	ESI− ($\times10^{8}$)	Rainy: 6.0 ± 1.0 Cloudy: 8.5 ± 0.8 Sunny: 6.7 ± 1.4	5.67	0.026

Table S4. Volcano selection and Benjamini–Hochberg false discovery rate evaluation for the 709 formula and ionization mode assignments used in HCA.

Assessment	**Criterion**	**ESI+ assignments**	**ESI− assignments**	**Total assignments**	**FDR result**
HCA input	$p < 0.05$, $\lvert \log_2 FC \rvert \geq 2$, detected in ≥10 tunnel samples	626	83	709	All $q < 0.10$; 698 $q < 0.05$
Assignments with $0.05 \leq q < 0.10$	Formula and ionization mode assignments retained by the volcano rule	10	1	11	All occurred in C1

Table S5. Sensitivity of HCA diagnostics to K and α using O/C, H/C, and carbon number as chemical descriptors, cosine distance for the temporal and chemical profiles, and complete linkage.

K	α	Mean silhouette	Negative silhouette (%)	Minimum cluster size	Feasible
3	0.2	0.575	3.8	16	No
3	0.3	0.714	2.4	10	No
3	0.4	0.655	2.7	10	No
3	0.5	0.537	9.3	10	No
3	0.6	0.253	9.9	20	No
3	0.7	0.605	3.9	55	Yes
3	0.8	0.622	4.2	50	Yes
3	0.9	0.647	4.7	59	Yes
4	0.2	0.566	3.1	3	No
4	0.3	0.469	7.8	10	No
4	0.4	0.506	6.6	10	No
4	0.5	0.545	4.5	10	No
4	0.6	0.252	23.1	20	No
4	0.7	0.550	6.9	46	Yes
4	0.8	0.616	4.2	15	No
4	0.9	0.643	4.8	12	No
5	0.2	0.554	4.1	3	No
5	0.3	0.469	8.0	3	No
5	0.4	0.503	7.3	3	No
5	0.5	0.540	6.8	10	No
5	0.6	0.232	12.6	20	No
5	0.7	0.551	7.3	19	No
5	0.8	0.404	11.1	15	No
5	0.9	0.617	4.2	12	No
6	0.2	0.417	9.6	3	No
6	0.3	0.424	11.6	3	No
6	0.4	0.500	9.9	3	No
6	0.5	0.529	8.2	3	No
6	0.6	0.217	19.6	20	No
6	0.7	0.552	6.5	15	No
6	0.8	0.408	11.6	15	No
6	0.9	0.400	18.8	12	No

Table S6. Formula assignment number and structural summary of the K = 3 HCA domains shown in Figure 6. Source related and elemental family fractions are calculated from assignment number; peak areas are not combined across ESI modes.

Description	**Shared CHO/CHON baseline**	**Sulfur rich tunnel retained**	**Low mass CHN and candidate TWP**
Assignments (ESI+/ESI−)	585 (518/67)	74 (61/13)	50 (47/3)
TCC + CCO assignments (%)	27.7	17.6	18
T–O assignments (%)	64.3	74.3	62
Dominant family assignments (%)	CHO + CHON 60.9	CHOS + CHONS 52.7	CHO + CHON 72.0
Candidate TWP assignments (%)	8	8.1	20
Median $\overline{\mathrm{OS}}_{\mathrm{c}}$	-1.07	-1.19	-1.09
Median AI_{mod}	0.29	0.22	0.35

Table S7. Top molecular formula assignments ranked by within domain peak area separately in each ionization mode. Peak area fractions are calculated within the corresponding domain and ESI mode and are not compared across modes.

Mode	Domain	Rank	Formula	m/z	Category	Within domain peak area (%)
ESI+	C1	1	$C_{12}H_{14}N_2O_2$	219.1128	CCO	11.10
	C1	2	$C_6H_8O_2$	113.0597	T–O	10.13
	C1	3	$C_{12}H_{23}NO$	198.1852	CCO	9.76
	C1	4	C_4H_8O	73.0648	CCO	6.50
	C1	5	$C_5H_{12}N_2O$	117.1022	CCO	4.35
	C1	6	$C_{12}H_{17}NO$	192.1383	CCO	3.98
	C1	7	$C_{10}H_{14}N_2$	163.1230	CCO	3.48
	C1	8	$C_{10}H_{16}O_5$	217.1071	CCO	1.79
	C1	9	$C_{10}H_{22}O_4$	207.1591	CCO	1.56
	C1	10	$C_{12}H_{16}N_2O_2$	221.1285	CCO	1.48
	C2	1	$C_{11}H_{12}N_2OS$	221.0743	T–O	6.30
	C2	2	$C_{20}H_{26}O_3$	315.1955	T–O	4.03
	C2	3	$C_{39}H_{50}N_2O$	563.3996	T–O	3.62
	C2	4	$C_{28}H_{43}N$	394.3468	T–O	3.52
	C2	5	$C_{10}H_{19}NO_7S$	298.0955	T–O	3.23
	C2	6	$C_7H_4N_2O_6S$	244.9863	T–O	3.02
	C2	7	$C_{12}H_{17}NO_7S$	320.0798	T–O	2.97
	C2	8	$C_{13}H_{24}N_2O_2S$	273.1631	T–O	2.77
	C2	9	$C_{21}H_{39}NO_{15}S_2$	610.1834	T–O	2.72
	C2	10	$C_{15}H_{32}N_2OS$	289.2308	T–O	2.39
	C3	1	$C_4H_{11}N$	74.0964	T–O	41.29
	C3	2	$C_4H_{10}N_2$	87.0917	T–O	21.69
	C3	3	$C_{17}H_{22}O$	243.1743	Primary TWP	4.70
	C3	4	$C_{16}H_{20}O$	229.1587	Primary TWP	4.04
	C3	5	$C_{18}H_{22}O$	255.1743	T–O	2.89
	C3	6	$C_{19}H_{30}O$	275.2369	TCC	2.15
	C3	7	$C_{19}H_{28}O$	273.2213	T–O	2.08
	C3	8	$C_{19}H_{34}N_2O_2$	323.2693	T–O	1.09
	C3	9	$C_{32}H_{30}N_2O_2$	475.2380	Primary TWP	1.09
	C3	10	$C_{16}H_{29}NO_6$	332.2068	T–O	0.95

Mode	Domain	Rank	Formula	m/z	Category	Within domain peak area (%)
ESI−	C1	1	$C_{14}H_{21}NO_3$	250.1449	TCC	29.36
	C1	2	$C_8H_9NO_3$	166.0510	CCO	9.34
	C1	3	$C_{18}H_{34}O_3$	297.2435	TCC	6.37
	C1	4	$C_9H_{18}O_5S$	237.0802	T–O	5.10
	C1	5	$C_{16}H_{30}O_4$	285.2071	CCO	4.39
	C1	6	$C_{11}H_{15}NO_3$	208.0979	CCO	3.73
	C1	7	$C_{18}H_{32}O_3$	295.2279	CCO	3.70
	C1	8	$C_{19}H_{30}O_4$	321.2071	TCC	2.78
	C1	9	$C_{12}H_{14}N_2O_3$	233.0932	TCC	2.16
	C1	10	$C_{12}H_{14}N_2O_2$	217.0983	TCC	2.14
	C2	1	$C_9H_{19}NOS$	188.1115	TCC	33.28
	C2	2	C_7H_5NOS	150.0019	Secondary TWP	15.33
	C2	3	$C_{10}H_8O_4$	191.0350	TCC	7.15
	C2	4	$C_{17}H_{30}O_3$	281.2122	TCC	6.36
	C2	5	$C_{18}H_{34}O_2$	281.2486	TCC	6.24
	C2	6	$C_{20}H_{26}O_3$	313.1809	T–O	5.54
	C2	7	$C_{12}H_{10}N_2O_2$	213.0670	T–O	5.10
	C2	8	$C_{18}H_{32}O_4$	311.2228	TCC	4.84
	C2	9	$C_{15}H_{10}O_5$	269.0455	T–O	3.90
	C2	10	$C_{18}H_{30}O_4$	309.2071	TCC	3.44
	C3	1	$C_{18}H_{28}O_4$	307.1915	T–O	45.66
	C3	2	$C_{12}H_{20}O_4$	227.1289	TCC	36.64
	C3	3	$C_{18}H_7NO_6$	332.0201	T–O	17.71

Table S8. Composition of the K = 4, α = 0.80 sensitivity solution using the same all 28 sample temporal descriptor and raw chemical descriptors as the selected K = 3 solution. C1 and C3 were retained, while C2 was divided into 59 and 15 assignment groups; the smaller group did not meet the minimum domain size criterion.

K = 4 domain	**C1**	**C2a**	**C2b**	**C3**
Assignments	585	59	15	50
K = 3 origin	C1	C2	C2	C3
T–O assignments (%)	64.3	78.0	60.0	62.0
Sulfur containing assignments (%)	26.0	52.5	53.3	20.0
CHN assignments (%)	9.7	8.5	6.7	6.0
Candidate TWP assignments (%)	8.0	6.8	13.3	20.0
Median OS_C	-1.07	-1.20	-1.17	-1.09
Median AI_mod	0.29	0.26	0.10	0.35

Table S9. Sensitivity of the K = 3, α = 0.80 partition to column wise scaling of O/C, H/C, and carbon number before calculating chemical cosine distance. Agreement was calculated after optimal label matching to the raw descriptor solution.

Descriptor scaling	Raw	Column z-score	Column min–max
Mean silhouette	0.622	0.246	0.243
Negative silhouette (%)	4.2	10.7	19.0
Minimum domain size	50	32	43
Adjusted Rand index	1.000	0.463	0.499
Mapped agreement (%)	100.0	82.5	83.9

Table S10. Influence of $C_4H_{11}N$ in ESI+ on C3 membership and HCA diagnostics. HCA was repeated after excluding $C_4H_{11}N$ using the same descriptors, distance construction, K, and α.

Metric	**Original K = 3 analysis**	**After $C_4H_{11}N$ exclusion**
Total HCA assignments	709	708
C3 assignments	50	49
$C_4H_{11}N$ contribution to ESI+ C3 peak area (%)	41.3	Excluded
Original C3 assignments retained together	50/50	49/49
Retained together (%)	100	100
Mean silhouette	0.622	0.620
Median silhouette	0.729	0.724
Negative silhouette (%)	4.2	4.2

Table S11. Temporal variability of normalized domain peak areas calculated separately within ESI+ and ESI− across all 28 tunnel samples.

Domain	Ionization mode	Assignments	Samples	Mean normalized domain peak area	SD	CV (%)
C1	ESI+	518	28	6.94e+09	1.69e+09	24.3
C1	ESI−	67	28	1.23e+08	3.22e+07	26.2
C2	ESI+	61	28	3.95e+07	2.99e+07	75.7
C2	ESI−	13	28	5.03e+06	5.48e+06	108.9
C3	ESI+	47	28	8.62e+07	7.14e+07	82.8
C3	ESI−	3	28	9.52e+05	7.17e+05	75.3

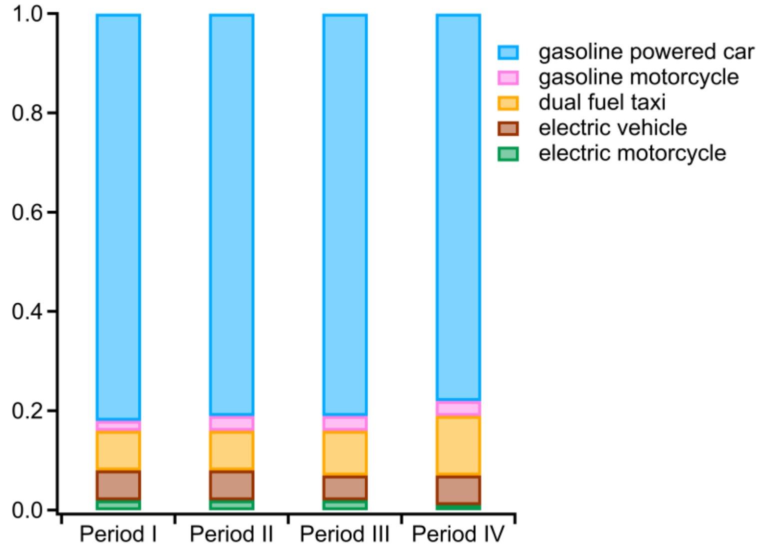

Figure S1. Vehicle type composition observed during the four tunnel sampling periods.

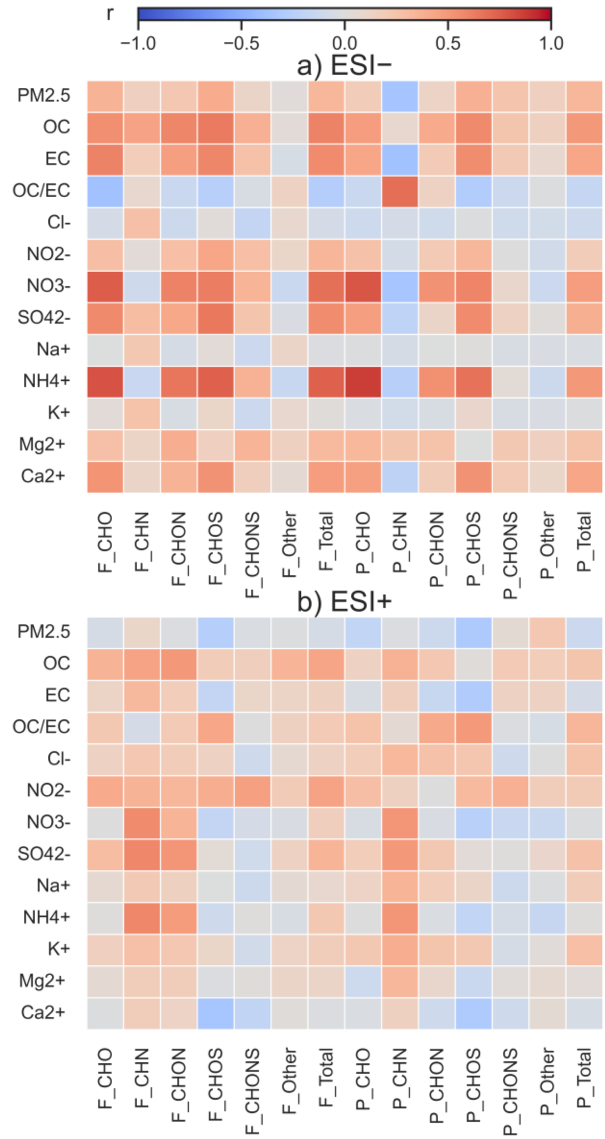

Figure S2. Heatmaps of Pearson correlation coefficients (r) between HRMS molecular-class metrics and co-located bulk aerosol variables ($PM_{2.5}$, OC, EC, OC/EC, and inorganic ions) in (a) ESI− and (b) ESI+. F_* denotes the number of assigned formulas (formula count), and P_* denotes the total peak area.

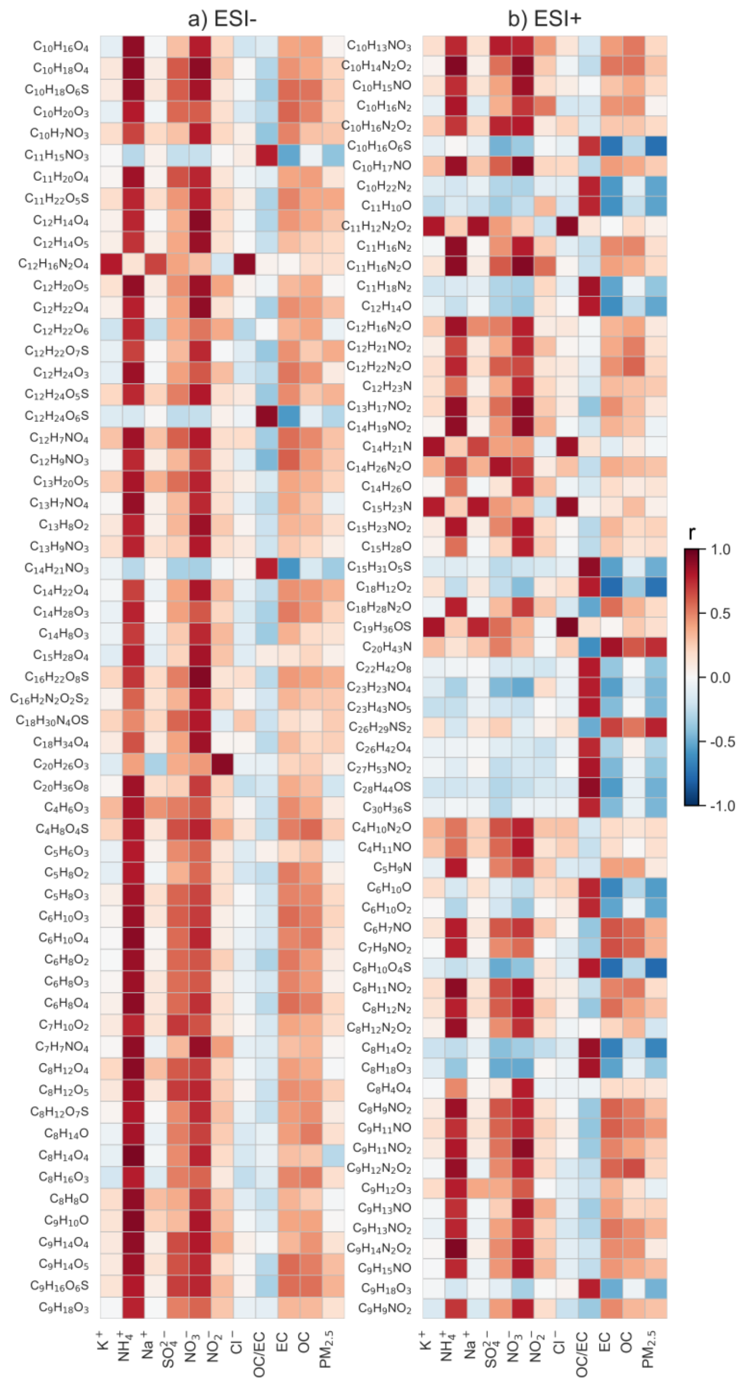

Figure S3. Heatmaps of Pearson correlation coefficients (r) between individual molecular formula assignments and co-located bulk aerosol measurements ($PM_{2.5}$, OC, EC, OC/EC, and inorganic ions) in (a) ESI− and (b) ESI+. Only assignments with at least 17 paired observations and $|r| \geq 0.75$ for at least one bulk measurement are displayed.

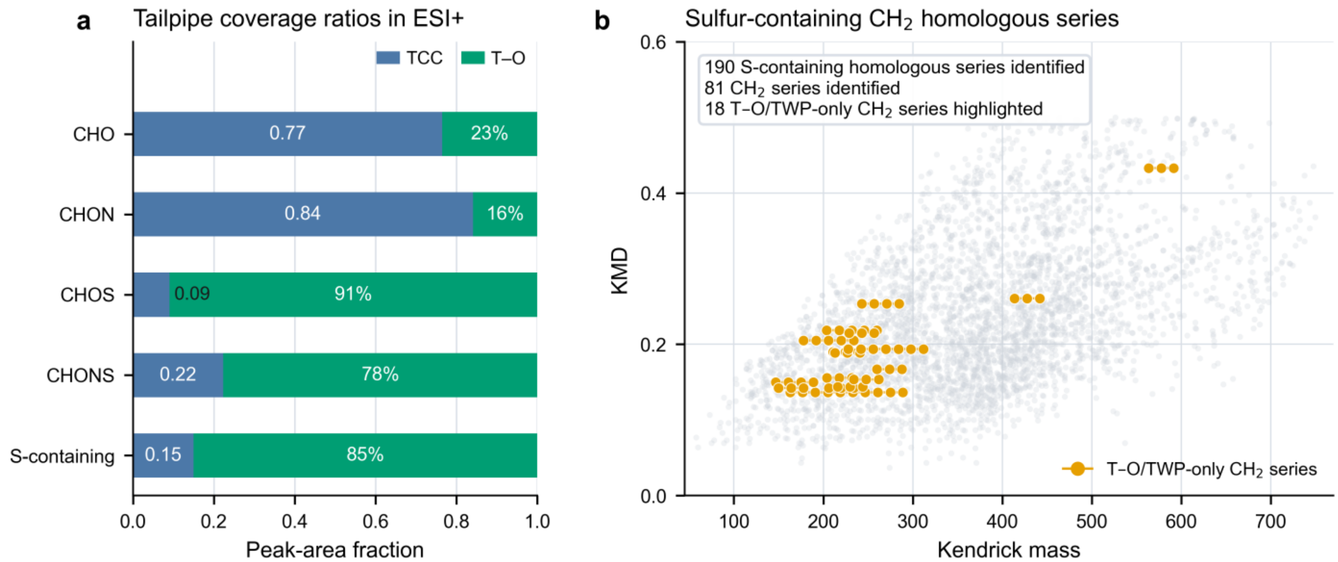

Figure S4. (a) Tailpipe coverage ratios (TCRs) for major elemental classes in ESI+. TCR was calculated from pooled blank-corrected peak areas across the 28 tunnel samples as ΣA_TCC/[ΣA_TCC + ΣA_T–O]. Blue and green segments denote the TCC and T–O fractions, respectively; labels in these segments show TCR and the complementary T–O percentage. (b) CH_2-based Kendrick mass defect (KMD) diagram for sulfur-containing formula–ionization-mode assignments from ESI+ and ESI−. Gray points show assignments within the displayed KMD range, whereas orange connected points highlight the 18 T–O/TWP-only CH_2 series selected from 81 CH_2 series in the full network of 190 sulfur-containing homologous series.

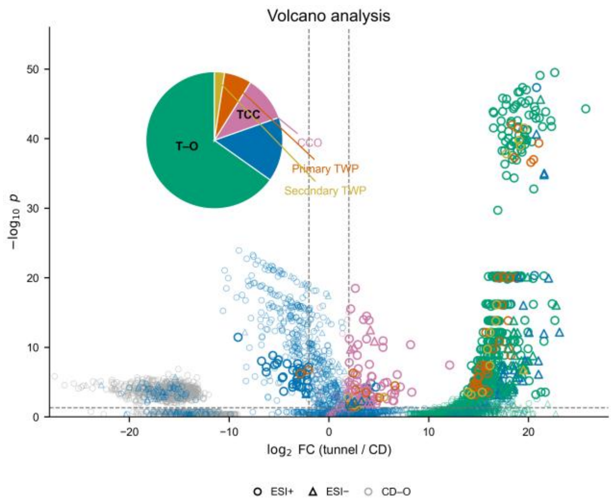

Figure S5. Volcano plot comparing tunnel aerosol with the controlled single-vehicle chassis-dynamometer (CD) tailpipe molecular profile. Analyses were performed separately for ESI+ and ESI−, and each point represents one formula–ionization-mode assignment; all 5666 tested assignments are shown. The x axis shows $\log_2$FC (tunnel/CD) based on group-mean within-mode relative peak areas, and the y axis shows $-\log_{10}p$ from two-sided Welch's t tests. Dashed lines mark $p = 0.05$ and $|\log_2 FC| = 2$. The 709 non-CD–O assignments detected in at least 10 of the 28 tunnel samples and meeting both thresholds were retained for HCA and are shown with larger symbols and thicker outlines; the inset summarizes their category composition. Outline colors denote T–O, TCC, CCO, candidate Primary TWP, candidate Secondary TWP, and CD–O (gray); circles and triangles denote ESI+ and ESI−, respectively. Normalization, transformation, and fold-change calculations are detailed in Sect. S2.

REFERENCES

Anders, S. and Huber, W.: Differential expression analysis for sequence count data, Genome Biology, 11, R106, 10.1186/gb-2010-11-10-r106, 2010.

Benjamini, Y. and Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing, Journal of the Royal Statistical Society. Series B (Methodological), 57, 289–300, 10.1111/j.2517-6161.1995.tb02031.x, 1995.

Han, J., Pei, J., and Tong, H.: Data Mining: Concepts and Techniques, 4th edn., Morgan Kaufmann, Cambridge, MA, USA, 2022.

Klöckner, P., Seiwert, B., Eisentraut, P., Braun, U., Reemtsma, T., and Wagner, S.: Characterization of tire and road wear particles from road runoff indicates highly dynamic particle properties, Water Research, 185, 116262, 10.1016/j.watres.2020.116262, 2020.

Rousseeuw, P. J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, 20, 53–65, 10.1016/0377-0427(87)90125-7, 1987.

Seiwert, B., Nihemaiti, M., Troussier, M., Weyrauch, S., and Reemtsma, T.: Abiotic oxidative transformation of 6-PPD and 6-PPD quinone from tires and occurrence of their products in snow from urban roads and in municipal wastewater, Water Research, 212, 118122, 10.1016/j.watres.2022.118122, 2022.

Strehl, A., Ghosh, J., and Mooney, R.: Impact of similarity measures on web-page clustering, AAAI 2000 Workshop on Artificial Intelligence for Web Search, 58–64,

Welch, B. L.: The generalization of 'Student's' problem when several different population variances are involved, Biometrika, 34, 28–35, 10.1093/biomet/34.1-2.28, 1947.