Supplementary Information for "Constraining the response factors of an extractive electrospray ionization mass spectrometer for near-molecular aerosol speciation"

Dongyu S. Wang, Chuan Ping Lee, Jordan E. Krechmer, Francesca Majluf, Yandong Tong, Manjula R. Canagaratna, Julia Schmale, André S.H. Prévôt, Urs Baltensperger, Josef Dommen, Imad El Haddad, Jay G. Slowik, and David M. Bell

Section S1. Condensation sink estimation

The condensation sink, *CS* (Lehtinen et al., 2003; Dal Maso et al., 2002) in s^{-1} is calculated as

$$CS = 2\pi D \int_0^\infty d_p \cdot \beta_M(d_p) \cdot N(d_p) \cdot dd_p \quad \text{Eq. (S1)}$$

where D is the vapor diffusivity, d_P is the particle diameter, the $N(d_P)$ is the number of particle of diameter d_p, and $\beta_M(d_P)$ is the Fuchs-Sutugin correction factor for gas-phase diffusion over particles in the transition regime. Using a discrete particle size distribution as measured by the SMPS, we calculate *CS* using an approximation of the integral, namely

$$CS = 2\pi D \sum_i \beta_i d_{pi} N_i \quad \text{Eq. (S2)}$$

The lifetime for gaseous condensation in the presence of a *CS* is (Markku Kulmala and Wagner 2001)

$$\tau_{cond} = \frac{1}{CS} \quad \text{Eq. (S3)}$$

As an approximation, D can be assumed to be 6 to 7 x 10^{-6} m^2 s^{-1} for condensable organic vapors (Palm et al., 2016; Krechmer et al., 2017). A more nuanced estimation is described below. The Fuchs-Sutugin correction factor β, is calculated

$$\beta = \frac{K_n+1}{0.377 \cdot K_n + \frac{4}{3} \cdot \alpha^{-1} \cdot K_n + \frac{4}{3} \cdot \alpha^{-1} \cdot {K_n}^2 + 1} \quad \text{Eq. (S4)}$$

where α is the mass accommodation coefficient. In lieu of empirical values, unity is assumed for α (Markku et al., 2001). Recent experimental results support this unity assumption (Krechmer et al., 2017; Liu et al., 2019). K_n is the Knudsen number,

$$K_n = \frac{2\lambda}{d_p} \quad \text{Eq. (S5)}$$

where the particle radius (d_p / 2) is used as the characteristic length; λ is the effective free mean path of vapor molecules. The mean free path in dry air varies slightly in the literature, e.g. 6.53 to 6.673 x 10^{-8} m (Jennings 1988). The mean free path of any organic compound can be calculated if its gas-phase diffusion coefficient (at the bath gas pressure), D_{Pr}, and average molecular speed c, are known,

$$\lambda = \frac{3D_{pr}}{c} \quad \text{Eq. (S6)}$$

$$c = \sqrt{\frac{8RT}{\pi MW}} \quad \text{Eq. (S7)}$$

where R is the ideal gas constant (8.314 J mol^{-1} K^{-1}), T is the temperature in K, and MW is the molar mass (kg mol^{-1}). Note that D_{pr} is a function of bath gas pressure, P (Torr), and the gas diffusivity, D (Torr cm^{-2} s^{-1})

$$D_{pr} = \frac{D}{P} \quad \text{Eq. (S8)}$$

For a trace gas A in a bath gas B, the gas diffusivity could be estimated using Fuller's method (Fuller et al., 1966; Tang et al., 2015),

$$D(A,B) = \frac{1.0868 \times T^{1.75}}{\sqrt{m(A,B)}\left(\sqrt[3]{V_A}+\sqrt[3]{V_B}\right)^2} \quad \text{Eq. (S9)}$$

where V_A and V_B are dimensionless diffusion volumes of A, and B; $m(A,B)$ is the reduced mass of the A-B pair and can be calculated based on the molecular masses (g mol^{-1}) of A and B, m_A and m_B, respectively

$$m(A,B) = \frac{2}{(1/m_A+1/m_B)} \quad \text{Eq. (S10)}$$

V_A may be estimated from the molecular formula of the trace gas

$$V = \sum n_i V_i \quad \text{Eq. (S11)}$$

where n_i is the number of atoms with diffusion volume of V_i, which is 15.9 for C, 2.31 for H, 6.11 for O, and 4.54 for N (Reid et al., 1987). Subtracting 18.3 from the total diffusion volume accounts for the effect of the aromatic ring. For compounds containing multiple aromatic rings, it maybe be best to correct only for independent aromatic rings, based on limited experimental data (Tang et al., 2015). Alicyclic rings are not expected to have an effect on the diffusion volume (Tang et al., 2015). Diffusion volumes of common bath gasses are known instead of estimated: N_2 (18.5), O_2 (19.7), H_2O (13.1). For inorganic and slightly oxygenated organic compounds, the mean free path of condensable vapors may be quite uniform (within 20%), where the Knudsen number can be estimated based on pressure and particle diameter alone (Tang et al., 2015),

$$K_n = \frac{2}{d_P} \times \frac{\lambda_P}{P} \quad \text{Eq. (S12)}$$

where P is the pressure of air in atm, and λ_P is the pressure normal mean free path equal to 100 nm atm. The deviation of K_n estimated using Eq. S12 for a 100 nm particle (i.e. K_n = 2) with respect to that estimated using Eq. S5 for selected compounds is shown in Table S1 with the corresponding gas diffusivity D, estimated using Eq. S9. All compounds are assumed to be non-aromatic unless indicated otherwise. For C_5 to C_{10} VOCs (e.g. isoprene, monoterpenes) and their oxidation products (e.g. C_5 to C_{10} monomers and C_{20} dimers), the estimated diffusivities differ less than a factor of 2 from $6.5 \cdot 10^{-6}$ cm^2 s^{-1}. Diffusion volume correction for (single) aromatic rings results in minor differences (< 5%) of the estimated D values. The estimated Knudsen numbers agree within 15%, as do the estimated Fuchs-Sutugin correction factors, β, between the simplified and the more rigorous estimation methods, assuming either a mass accommodation coefficient of 1 ($3.08 \cdot 10^{-1}$ for all compounds) or 0.1 ($3.67 \cdot 10^{-2}$ for all compounds), estimated using Eq. S4 and Eq. S12.

Table S1. Knudsen number and gas diffusivity

Gas	Kn	%Diff[a]	Diffusivity (m^2 s^{-1})	β (α=1)[b]	β (α=0.1)[b]
C_3H_6	1.83	9.30	1.18×10^{-5}	3.29×10^{-1}	4.00×10^{-2}
$C_3H_6O_2$	2.05	-2.44	9.97×10^{-6}	3.02×10^{-1}	3.58×10^{-2}
$C_3H_6O_4$	2.20	-9.26	8.96×10^{-6}	2.85×10^{-1}	3.34×10^{-2}
$C_3H_6O_6$	2.32	-13.77	8.27×10^{-6}	2.73×10^{-1}	3.18×10^{-2}
C_5H_8	1.77	13.02	8.98×10^{-6}	3.38×10^{-1}	4.13×10^{-2}
$C_5H_8O_2$	1.94	3.00	8.13×10^{-6}	3.15×10^{-1}	3.78×10^{-2}
$C_5H_8O_4$	2.07	-3.55	7.55×10^{-6}	2.99×10^{-1}	3.54×10^{-2}
$C_5H_8O_6$	2.18	-8.19	7.12×10^{-6}	2.88×10^{-1}	3.38×10^{-2}
$C_5H_8O_8$	2.26	-11.67	6.77×10^{-6}	2.79×10^{-1}	3.25×10^{-2}
$C_7H_8O_1$[c]	1.94	2.88	7.83×10^{-6}	3.14×10^{-1}	3.77×10^{-2}
$C_7H_8O_1$	1.83	9.41	7.36×10^{-6}	3.30×10^{-1}	4.00×10^{-2}
$C_7H_8O_2$	1.89	5.55	7.12×10^{-6}	3.21×10^{-1}	3.87×10^{-2}
$C_7H_8O_4$	2.01	-0.45	6.73×10^{-6}	3.06×10^{-1}	3.65×10^{-2}
$C_7H_8O_6$	2.10	-4.91	6.42×10^{-6}	2.96×10^{-1}	3.49×10^{-2}
$C_7H_8O_8$	2.18	-8.37	6.16×10^{-6}	2.87×10^{-1}	3.37×10^{-2}
$C_7H_8O_{10}$	2.25	-11.12	5.93×10^{-6}	2.80×10^{-1}	3.27×10^{-2}
C_9H_{12}[d]	1.81	10.46	6.92×10^{-6}	3.32×10^{-1}	4.04×10^{-2}
C_9H_{12}	1.72	16.08	6.58×10^{-6}	3.44×10^{-1}	4.24×10^{-2}
$C_9H_{12}O_2$	1.84	8.65	6.25×10^{-6}	3.28×10^{-1}	3.98×10^{-2}
$C_9H_{12}O_4$	1.94	3.13	5.98×10^{-6}	3.15×10^{-1}	3.78×10^{-2}
$C_9H_{12}O_6$	2.02	-1.13	5.76×10^{-6}	3.05×10^{-1}	3.63×10^{-2}
$C_9H_{12}O_8$	2.10	-4.54	5.57×10^{-6}	2.97×10^{-1}	3.51×10^{-2}
$C_9H_{12}O_{10}$	2.16	-7.33	5.40×10^{-6}	2.90×10^{-1}	3.41×10^{-2}
$C_{10}H_{16}$	1.70	17.36	6.11×10^{-6}	3.47×10^{-1}	4.29×10^{-2}
$C_{10}H_{16}O_2$	1.81	10.41	5.85×10^{-6}	3.32×10^{-1}	4.04×10^{-2}
$C_{10}H_{16}O_4$	1.90	5.13	5.63×10^{-6}	3.20×10^{-1}	3.85×10^{-2}
$C_{10}H_{16}O_6$	1.98	0.96	5.44×10^{-6}	3.10×10^{-1}	3.70×10^{-2}
$C_{10}H_{16}O_8$	2.05	-2.42	5.28×10^{-6}	3.02×10^{-1}	3.58×10^{-2}
$C_{10}H_{16}O_{10}$	2.11	-5.21	5.13×10^{-6}	2.95×10^{-1}	3.48×10^{-2}
$C_{10}H_{16}O_{12}$	2.16	-7.57	5.00×10^{-6}	2.89×10^{-1}	3.40×10^{-2}
$C_{10}H_{16}O_{14}$	2.21	-9.58	4.88×10^{-6}	2.84×10^{-1}	3.33×10^{-2}
$C_{10}H_{16}O_{16}$	2.26	-11.32	4.77×10^{-6}	2.80×10^{-1}	3.26×10^{-2}
$C_{20}H_{32}O_6$	1.83	9.56	3.98×10^{-6}	3.30×10^{-1}	4.01×10^{-2}
$C_{20}H_{32}O_8$	1.87	6.86	3.92×10^{-6}	3.24×10^{-1}	3.91×10^{-2}
$C_{20}H_{32}O_{10}$	1.91	4.49	3.85×10^{-6}	3.18×10^{-1}	3.83×10^{-2}
$C_{20}H_{32}O_{12}$	1.95	2.38	3.80×10^{-6}	3.13×10^{-1}	3.75×10^{-2}
$C_{20}H_{32}O_{14}$	1.99	0.49	3.74×10^{-6}	3.09×10^{-1}	3.69×10^{-2}
$C_{20}H_{32}O_{16}$	2.02	-1.21	3.69×10^{-6}	3.10×10^{-1}	3.63×10^{-2}

(a). Percent difference of K_n=2, estimated using Eq. S12, with respect to the K_n estimated using Eq. S5. (b). Fuchs-Sutugin correction factors estimated using Eq. S4 assuming different values for mass accommodation coefficients; the K_n used here was estimated using Eq. S5 (c) *o*-Cresol (d) 1,2,4-trimethylbenzene

Section S2. Oxidation flow reactor schematic

A schematic of the experiment setup is shown in Figure S1 along with the physical dimensions of the oxidation flow reactor (OFR). VOC precursor and seed particles are injected near the entrance region of the OFR, whereas O_3 is injected coaxially in the direction of the flow through a 6 mm outer diameter stainless-steel tubing about 61 cm downstream of the entrance region. Instruments sampled from near the exit region of the OFR. The cross-sectional area of the OFR is approximately $4.3 \cdot 10^{-3}$ m^2. At 12 L min^{-1}, the plug flow velocity is roughly $4.65 \cdot 10^{-2}$ m s^{-1}. The residence time within the oxidation region (i.e. 39 cm) is roughly 8.38s, or an effective dilution rate of 0.12 s^{-1}.

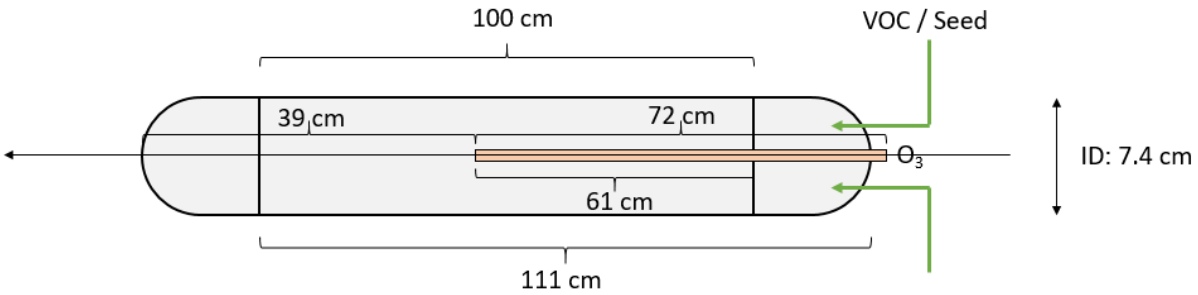


Figure S1. Flow tube dimension.

Section S3. Vocus-PTR Mass transmission

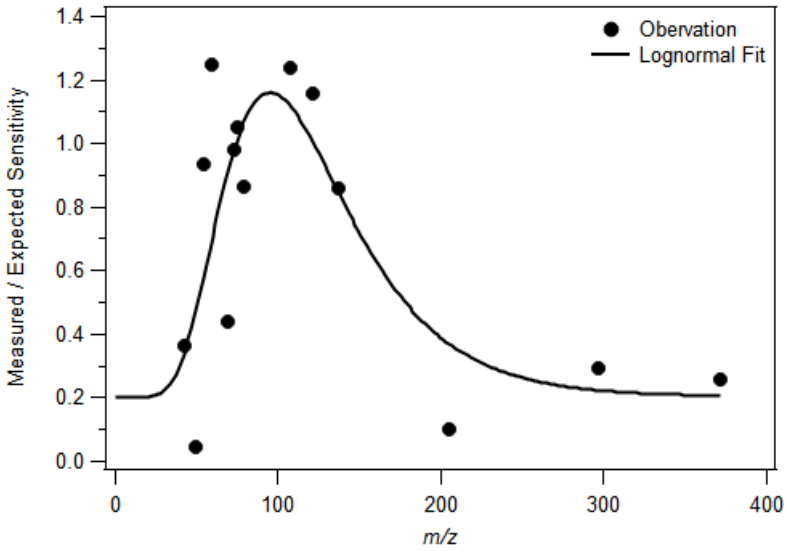


Figure S2. Vocus-PTR mass transmission efficiency

The mass transmission efficiency curve for Vocus-PTR is fitted using a lognormal function. Calibration of the mass transmission efficiency for PTR is described in details elsewhere (Holzinger et al., 2019).

$$MT = 0.20 + 0.96 \times \exp\left(-\frac{ln\left(\frac{m/z}{95.98}\right)}{0.58^2}\right) \quad \text{Eq. (S13)}$$

Section S4. AMS Vaporizer artifact correction

The high-resolution aerosol mass spectrometer (AMS) determines the aerosol composition in terms of NO_3, NH_4, SO_4, Chl, and Organics (OA). All experiments were conducted under low-NO_x conditions using NH_4NO_3 seed particles. Therefore, all NH_4^+ and NO_3^- observed are attributed to NH_4NO_3. Due to the high inorganic concentrations used (up to 11.6 mg m^{-3}), caution needs to be taken to account for vaporizer artifacts, where NO_x^+ ions generated from nitrate particles during the electron impact ionization process could oxidize organic residues on the vaporizer surface, producing CO_2^+ ions that are falsely attributed to organic aerosols (Pieber et al., 2016). The extent of this artifact is determined by injecting NH_4NO_3 seed particles into the OFR in the absence of any organic oxidation products. As shown in Figure S3a below, the correlation of the organic vaporizer artifact, $Org_{artifact,}$ can be described by an exponential function of the NH_4NO_3 (i.e. combined mass concentrations of NO_3^- and NH_4^+). This correlation is used to correct for $Org_{artifact}$ for all runs, as shown in Figure S3b to Figure S3d. Note that this correlation could change with the vaporizer history (Pieber et al., 2016). Here, the vaporizer artifact was characterized in the midst of the campaign.

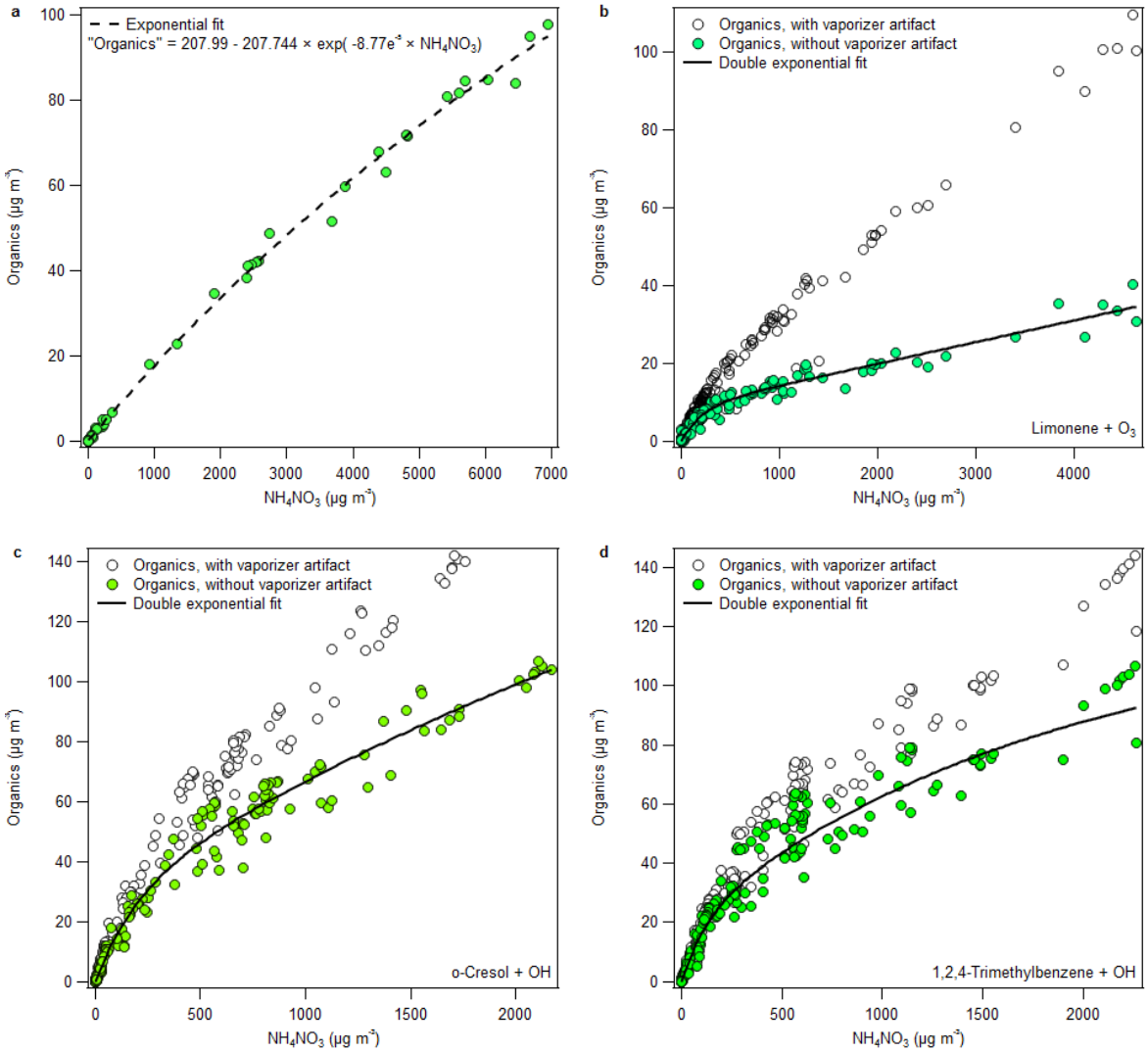


Figure S3. Inorganic salt-induced vaporizer artifact

(a) Artefact organics concentration observed by the AMS when sampling nebulized NH_4NO_3 in the absence of any organic oxidation products. An exponential function of NH_4NO_3 concentration is used to estimate the organic signal attributable to the vaporizer artifact. The

organic concentrations with and without applying this correction are shown in (b) for limonene ozonolysis, in (c) for the OH oxidation of *o*-cresol, and in (d) for the OH oxidation of 1,2,4-trimethylbenzene. The correlation between condensed organics and NH_4NO_3 seed concentrations can be roughly described by a double exponential function.

Section S5. Oxidation flow reactor model

The organic vapor wall loss may be estimated from the OFR dimension and the gas-diffusivities as proposed by McMurry and Grosjean (1995),

$$k_{wall} = \frac{1}{\tau_{wall}} = \frac{A}{V} \cdot \frac{2}{\pi}\sqrt{k_e D} \qquad \text{Eq. (S14)}$$

when the vapor wall accommodation coefficient is greater than 10^{-5}, i.e. eddy diffusion dominates. This is the case for oxidation flow reactors (OFR) of similar dimensions to the one used in this study (Brune 2019; George et al., 2007). *A* and *V* are the surface area (1.02×10^{-1} m^2) and volume (1.72×10^{-3} m^3) of the OFR, respectively. k_e is the coefficient of Eddy diffusion, which may be estimated as a function of the enclosure volume (Krechmer et al., 2016),

$$k_e(s^{-1}) = 0.004 + (5.6 \times 10^{-3})(V)^{0.74} \qquad \text{Eq. (S15)}$$

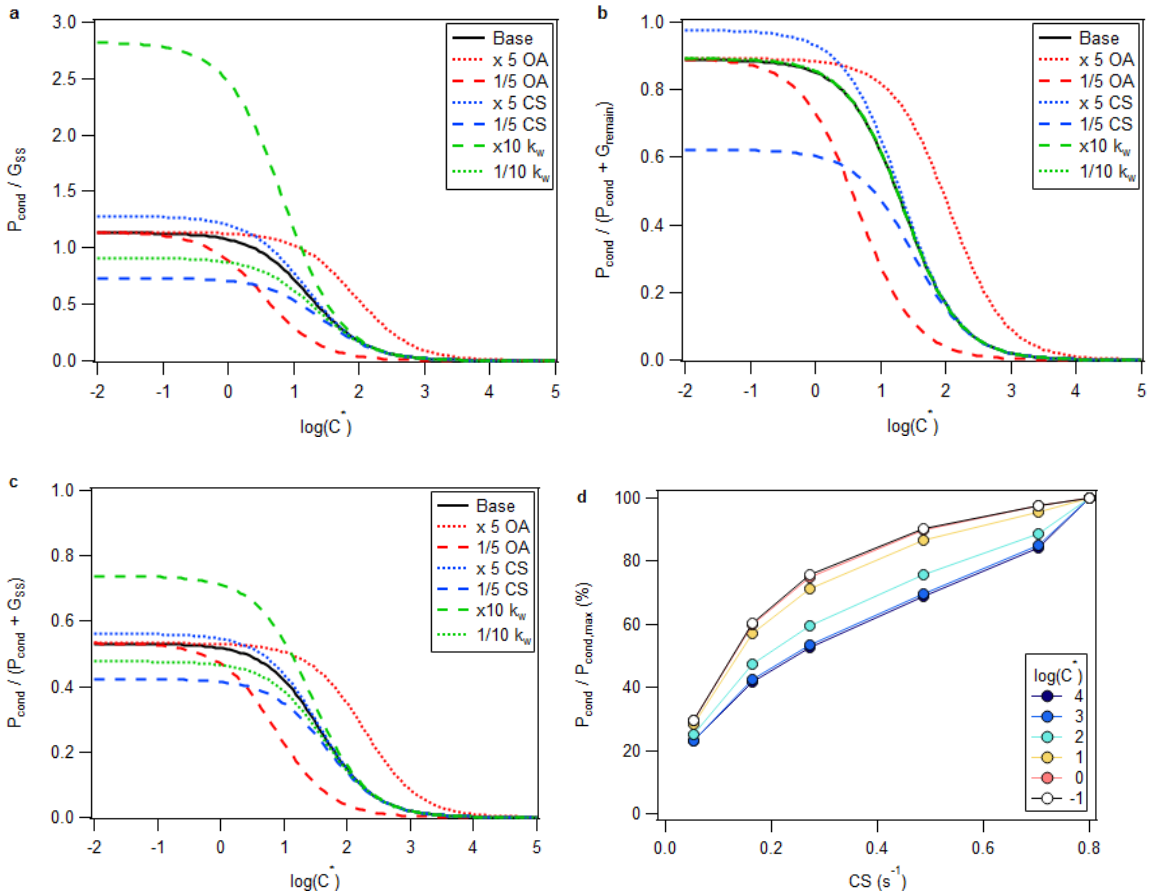
which is 4.05×10^{-3} s^{-1}. Due to their relatively small enclosure volume (relative to that of a typical smog chamber, k_e would be close to $4 \cdot 10^{-3}$ s^{-1} for most OFR designs. For estimated gas diffusivity, *D* ranging from $3.69 \cdot 10^{-6}$ ($C_{20}H_{32}O_{16}$) to $1.18 \cdot 10^{-5}$ (C_3H_6) m^2 s^{-1}, the corresponding k_{wall} ranges from $4.60 \cdot 10^{-3}$ s^{-1} to $8.22 \cdot 10^{-3}$ s^{-1}, resulting in a wall loss timescale, τ_{wall} between 122 and 218 s. Two different vapor wall loss experiments conducted using a PTR-TOF and an acetate atmospheric pressure interface chemical ionization TOF-MS indicate a 50% vapor wall loss rate at 10 L min^{-1} flow rate, which suggest a τ_{wall} similar to that of the dilution lifetime, i.e. 27 seconds, meaning that the actual k_{wall} is close to $3.7 \cdot 10^{-2}$ s^{-1}, roughly 4 to 8 times higher than Eq. S14 and Eq. S15 would suggest. For simplicity, a k_w value of 0.04 s^{-1} is used as the base case scenario. The effects of higher k_w (i.e. 0.4 s^{-1}) and lower k_w (i.e. 0.04 s^{-1}) values on the gas- and particle-phase concentrations are simulated and shown in Figure 3a-c for generic oxidation products of differing saturation vapor concentrations ranging from 10^{-2} to 10^6 μg m^{-3}. The OFR wall is also assumed to be a perfect sink for organic vapors, i.e. no back-partitioning of organic vapor from the wall to the gas-phase is considered.

The remaining gas-phase concentration, G_{remain} and the condensed particle-phase concentration, P_{cond} during seed injection are expressed in relative terms with respect to the steady gas-phase concentration prior to the seed injection, G_{ss} (e.g. G_{remain}/G_{ss} and P_{cond}/G_{ss}). So that they are not dependent on the absolute value of G_{ss}, and vice versa on the actual production rate, provided that the production rate is not affected by the seed injection.

The modeled gas-particle partitioning is shown below in Figure S4. A sensitivity analysis was performed by varying the organic aerosol concentration (OA), the condensation sink (CS), or the wall loss rate (k_w) from the base condition (20 μg m^{-3} OA, 1 s^{-1} CS, and 0.04 s^{-1} k_w) in Figure S4a-c. The observed OA and CS values were used to simulate the partitioning behaviors as shown in Figure S4d-i. For each VOC system, the observed OA concentration and CS roughly followed a linear correlation. Figure S4d shows the P_{cond} normalized to the maximum value as a function of CS, and suggests that it may be possible to infer the saturation vapor concentration, C^* of semi-volatile compounds based on the uptake trend without the knowledge of near-molecular particle-phase sensitivity or gas-phase concentration (as long as G_{SS} remains constant in this case). However, compounds of different C^* may exhibit similar

trends, i.e. high inter-correlations, which cannot be numerically resolved due to noise. Visually, this is obvious for compounds with log(C^*) > 2 or < -1 as shown in Figure S4d.

To determine the range of log(C^*) that could be in theory numerically resolved from the P_{cond} behaviors alone, we modeled the normalized P_{cond} for compounds with log(C^*) ranging from -2 to 6 using OA and CS values observed for each system. The lower C^* threshold is set at the point beyond which all compounds with lower C^* would exhibit normalized P_{cond} trends with intercorrelation (R^2 value from linear regression between the normalized P_{cond} values corresponding to any pair of C^* values, i.e. any two "vertical slices" from Figure S4e and S4f) above 0.99. The decision to set the cutoff at R^2 = 0.99 is arbitrary. The upper C^* threshold is similarly defined in Figure S4g-i. The experimentally constrainable log(C^*) ranges based on the uptake behavior alone are narrow: 1.25 to 2.02 for the cresol system, 1.18 to 2.09 for the TMB system, and 0.57 to 1.85 for the limonene system. The span of the constrainable C^* range is wider for the limonene system due to the higher maximum CS range explored experimentally (>2 s^{-1} as compared to <1 s^{-1} for the anthropogenic systems). The upper constrainable C^* range for limonene system (i.e. log(C^*) =1.85) is lower compared to that for either cresol (i.e. log(C^*) =2.02) or TMB system (i.e. log(C^*) =2.09) due to the lower maximum OA uptake as a function of CS for the limonene system as compared to the anthropogenic systems. All else being equal, the constrainable range of log(C^*) increases with the experimental CS range, which is limited by the maximum particle concentrations the instruments could accommodate before clogging or signal depletion becomes too severe.



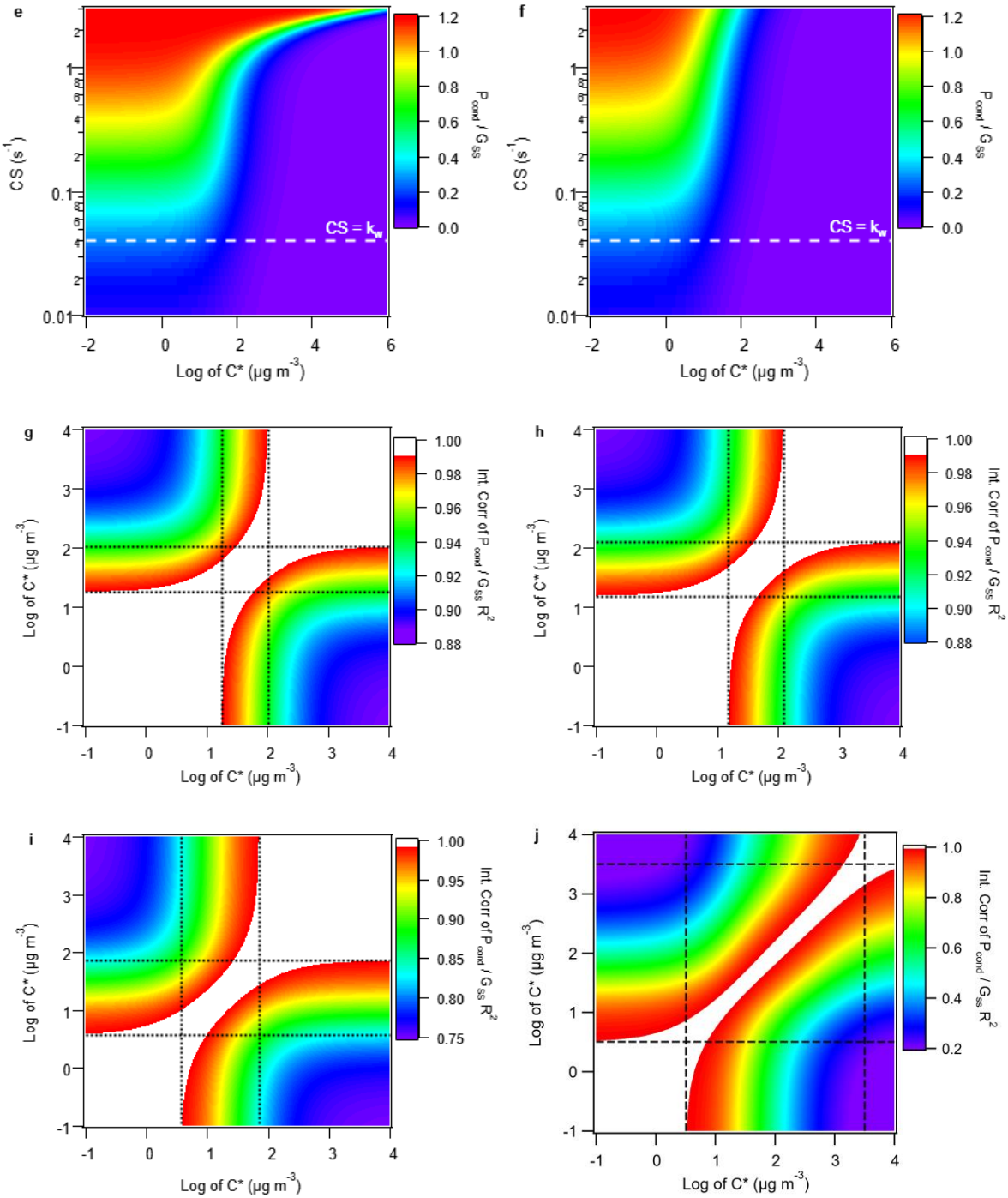


Figure S4. Modeled partitioning

(a-c) Expected distribution of organic oxidation products of differing volatilities between the gas- and particle-phase during the seed injection period for a hypothetical base case scenario of 20 µg m^{-3} organic aerosol concentration (*OA*), 1 s^{-1} condensation sink (*CS*), and 0.04 s^{-1} vapor wall loss rate (k_w). Alternative scenarios assume higher or lower *OA*, *CS*, and k_w. (d-i) Modeled ratio of P_{cond} to G_{ss} for compounds of varying $log(C^*)$ under observed OA and CS conditions. (a) Ratio of condensed organic material during seed injection, P_{cond} to the steady-state gas-phase concentration prior to seed injection, G_{SS}. The ratio can exceed 1 under high *CS* conditions. (b) Ratio of P_{cond} to the sum of P_{cond} with the gas-phase concentration during the seed injection period, G_{remain}. Partitioning between P_{cond} and G_{remain} is invariant with respect

to k_w. (c) Ratio of P_{cond} to the sum of P_{cond} and G_{SS}. (d) Normalized P_{cond} relative to the maximum expected value, $P_{cond,max}$ as a function of *CS* for compounds of different volatility. Note again that observed *CS* and *OA* values from the anthropogenic experiments are used to simulate the uptake behavior shown in (d), whereas hypothetical *CS*, *OA*, and k_w conditions are used to simulate the behaviors shown in (a-c). (e) Ratio of P_{cond} to G_{SS} for compounds of varying log(C^*) at different *CS* for the cresol and TMB systems, which exhibited similar intercorrelations between observed *OA* and *CS*. (f) Ratio of P_{cond} to G_{SS} for compounds of varying log(C^*) at different *CS* for the limonene system. (g) Inter-correlation of the expected normalized P_{cond}, similar to those shown in (d), for compounds of varying log(C^*) under the uptake conditions in the cresol system. (h) Inter-correlation of the expected normalized P_{cond}, for compounds of varying log(C^*) under the uptake conditions in the TMB system. (i) Inter-correlation of the expected normalized P_{cond}, for compounds of varying log(C^*) under the uptake conditions in the limonene system. (j) Similar to g, but with the maximum CS range extrapolated to 2 s^{-1} from ~0.8 s^{-1} to examine its effect on constrainable log(C^*) range. Regions with R^2 values exceeding 0.99 are shown in white in (g-j), where the log(C^*) empirically determined from the normalized P_{cond} is considered as highly uncertain due to experimental noise and high intercorrelations of the normalized P_{cond} behavior with compounds of different log(C^*). Behaviors of compounds with log(C^*) below -1 or above 4 are not shown, as they are indistinguishable per our definition based on the intercorrelation value R^2.

Section S6. EESI-TOF vs Vocus-PTR Composition

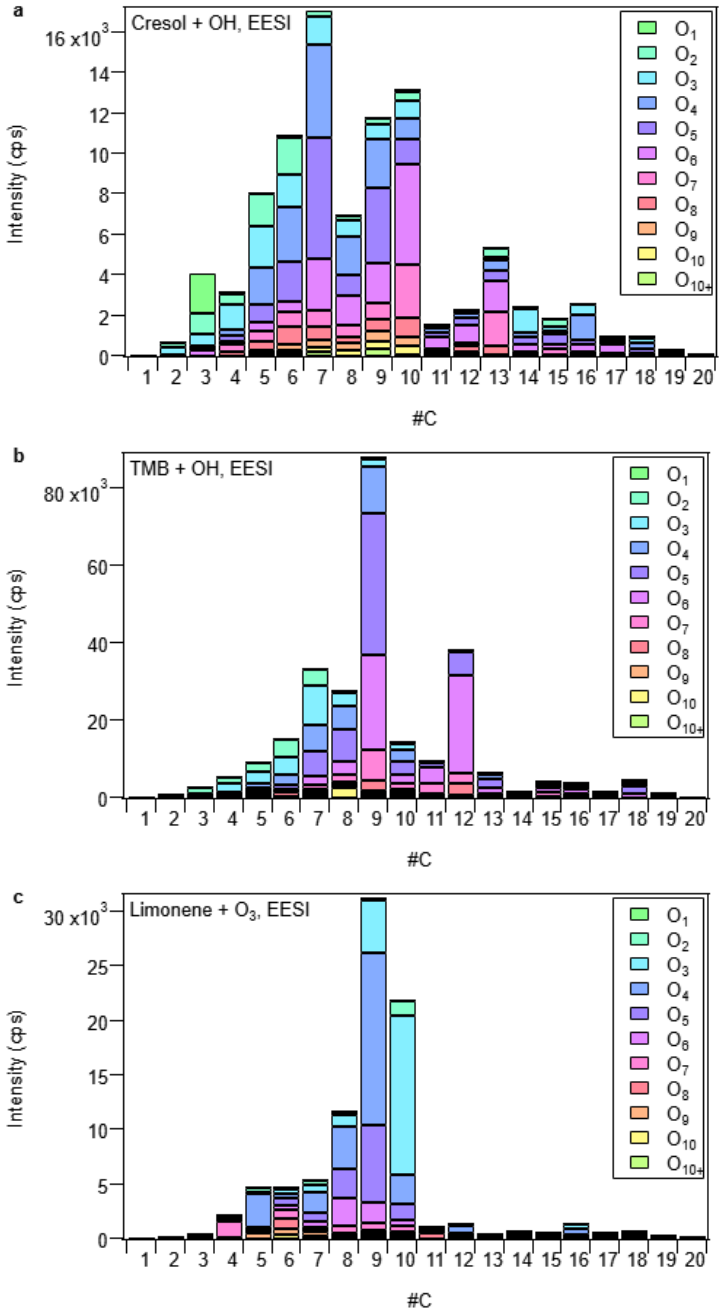


Figure S5. Average particle-phase composition

Ion intensity of $[M+Na]^+$ adducts observed during (a) OH-oxidation of cresol, (b) OH-oxidation of TMB, and (c) ozonolysis of limonene. For each VOC and oxidant system, the average composition over all seed injection / organic aerosol uptake events is shown. Ion intensities are grouped by their carbon number (*#C*) and further distinguished by the oxygen number as shown in the legend.

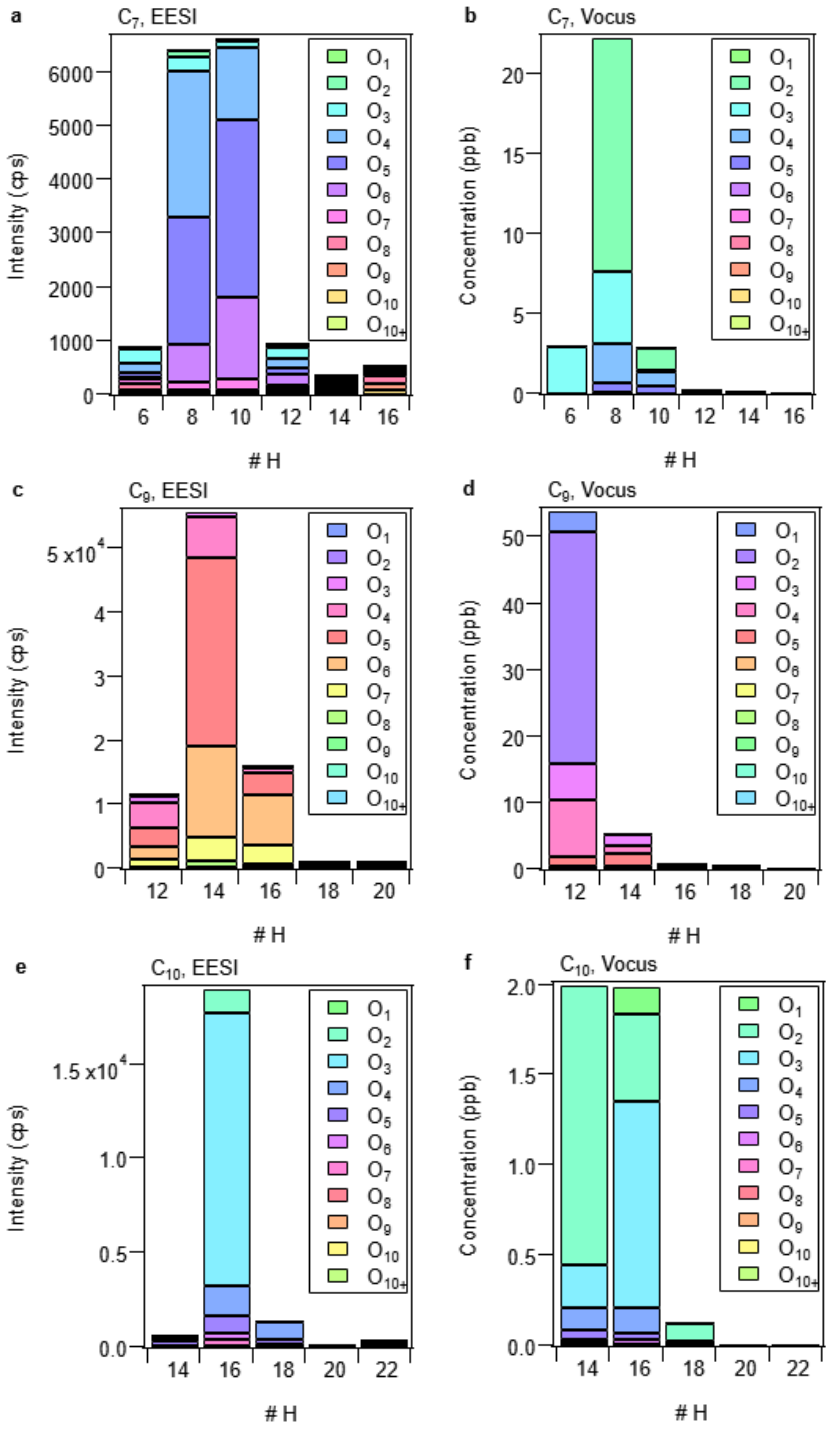


Figure S6. Comparison of major particle- and gas-phase oxidation products
Intensities of selected $[M+Na]^+$ adducts observed by the EESI-TOF for the particle-phase are shown for (a) C_7 OH + cresol oxidation products, (c) C_9 OH + TMB oxidation products, and (e) C_{10} limonene + O_3 oxidation products. Intensities of selected $[M+H]^+$ ions observed by the Vocus-PTR in the gas-phase are shown for (b) C_7 OH + cresol oxidation products, (d) C_9 OH + TMB oxidation products, and (f) C_{10} limonene + O_3 oxidation products. Average particle-phase signals over all uptake events are shown in (a), (c), and (e). Average steady-state gas-phase concentrations prior to each uptake event are shown in (b), (d), and (f). Note that the color scales are only consistent within each of the (a-b), (c-d), and (e-f) pairs. Ion intensities

are grouped by the number of hydrogens (*#H*) and further distinguished by the number of oxygen as indicated in the legends. The *o*-cresol is not included in (a) and (b) because it is the VOC precursor and not an oxidation product.

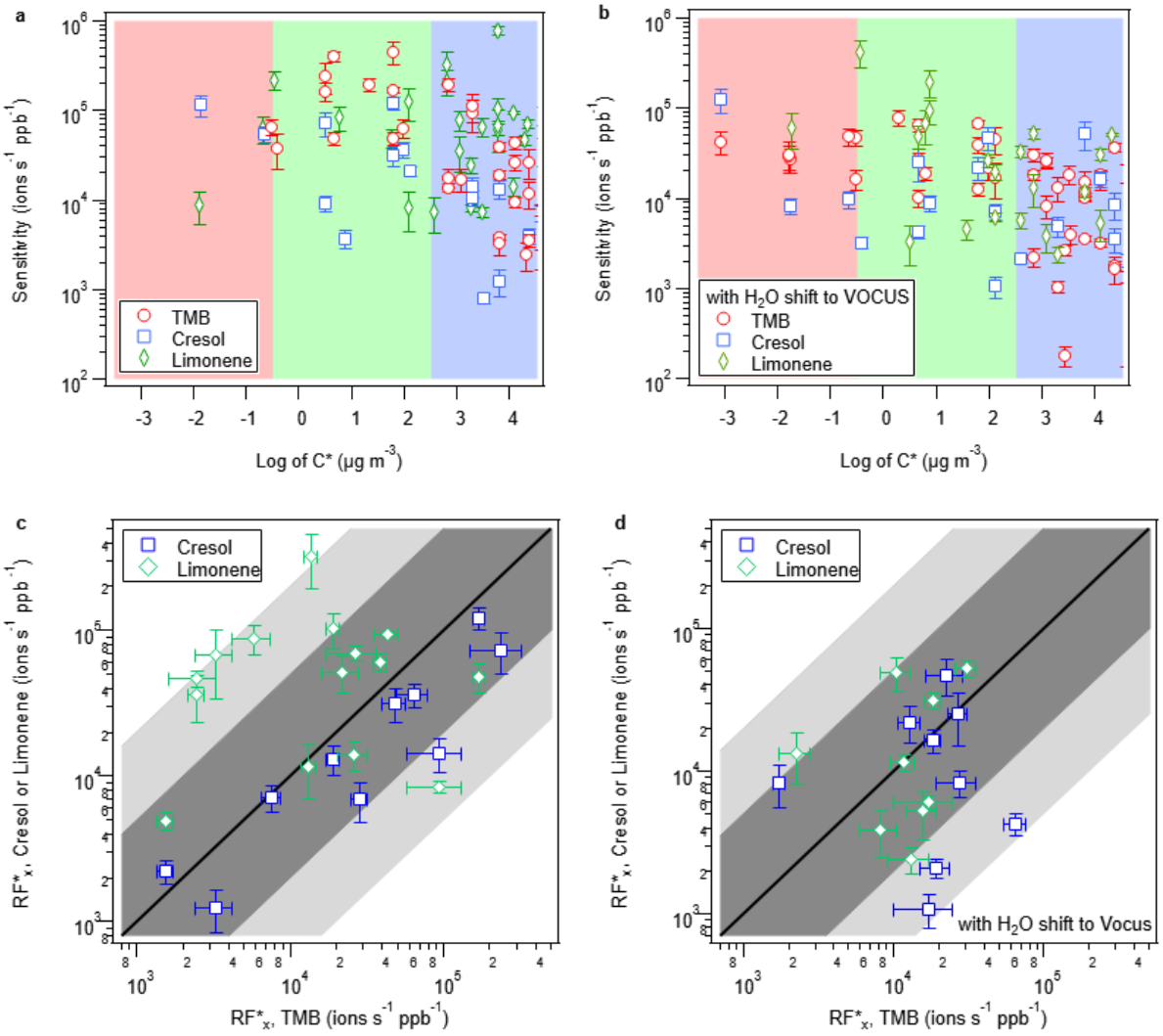


Figure S7. Sensitivity estimation with and without assuming –H_2O loss

The EESI-TOF response factor, RF_x^* estimated by performing a linear regression of the observed increase in particle-phase EESI-TOF ion intensity vs. the decrease in the gas-phase concentration as measured by the Vocus-PTR under two different assumptions: (a) and (c), where no ion-fragmentation occurs in the Vocus-PTR or (b) and (d), where all ions undergo H_2O loss inside the Vocus-PTR. In (a) and (b), RF_x^* values are plotted against the log(C^*) estimated based on the molecular formula. In (c) and (d), the RF_x^* of isomers observed from different oxidation systems are compared. Only oxidation products with $R^2 \geqslant 0.5$ for the linear regression of EESI-TOF signal increase and Vocus-PTR mixing ratio decreases are shown. The red, green, and blue shaded regions in (a) and (b) indicate the volatility ranges corresponding to LVOC, SVOC, and IVOC, respectively. Overall, the inverse correlation between EESI RF^*_x and log(C^*) is retained under different H_2O loss assumptions. The lighter and darker shaded regions in (c) and (d) indicate a factor of 5 or 20 deviation from the 1-to-1 line, respectively. Overall, the isomeric sensitivities agree within a factor of 20, where the RF^*_x deviation between

limonene and TMB isomers is less pronounced under the H_2O loss assumption as compared to the no fragmentation assumption.

Section S7. Parameterization and Model Validation

Based on the elemental formulae measured by the EESI-TOF and the Vocus-PTR, several additional features could be derived from the number of carbon (n_C), hydrogen (n_H), and oxygen (n_O), including the exact molecular mass (*MW*), the mass defect (Δm), the hydrogen-to-carbon ratio (*H:C*), the oxygen-to-carbon ratio (*O:C*), the double bond equivalent (*DBE*), and the double bond equivalent per carbon (*DBEpC*)

$$DBE = 1 + \frac{1}{2}(2C - H + N + P) \quad \text{Eq. (S16)}$$

The aromaticity index (*AI*) can be calculated as

$$AI = \frac{DBE_{AI}}{C_{AI}} = \frac{1+C-O-S-0.5H}{C-O-S-N-P} \quad \text{Eq. (S17)}$$

Which has been reported to underestimate the aromaticity compared to the aromaticity equivalent (X_c) proposed by Yassine et al. (Yassine et al., 2014)

$$X_C = \frac{C-(H-C)}{DBE} + 1 \quad \text{Eq. (S18)}$$

where, if $DBE \leq 0$, X_c is set to 0. Note that for CHO compounds, Eq. S18 simplifies to

$$X_C = 3 - \frac{2}{DBE} \quad \text{Eq. (S19)}$$

In addition, the carbon-oxygen non-ideality (NI_{CO}) from Eq. (7) itself is an interaction term between the product of the number of carbon and oxygen atoms (P_{CO}) and the inverse of the sum of carbon and oxygen atoms (I_{CO}),

$$NI_{CO} = \frac{n_C n_O}{n_C + n_O} = P_{CO} \times I_{CO} \quad \text{Eq. (S20)}$$

In addition to the aforementioned features, the log of effective saturation vapor concentration, $\log(C^*)$ is included as a feature.

Preliminary ordinary least square (OLS) regressions of the near-molecular EESI-TOF response factor, RF^*_x with n_C, n_O, *MW*, P_{CO}, I_{CO}, and NI_{CO} are shown in Figure S8a-f for each of the three VOC systems studied. The RF^*_x values estimated for cresol and TMB oxidation products appear to increase as the molecules increase in size (i.e. positive correlation with *MW* and n_C) and/or become more functionalized (i.e. positive correlation with n_O). The correlations also appear to be steeper for the TMB system than for the cresol system. In contrast, the RF^*_x values estimated for limonene oxidation products do not appear to be well correlated with n_C, n_O, *MW*, P_{CO}, I_{CO}, or NI_{CO}. The discrepancies observed between the aromatic systems and the biogenic system are likely due to differences in the structure of the oxidation products as discussed in the main text.

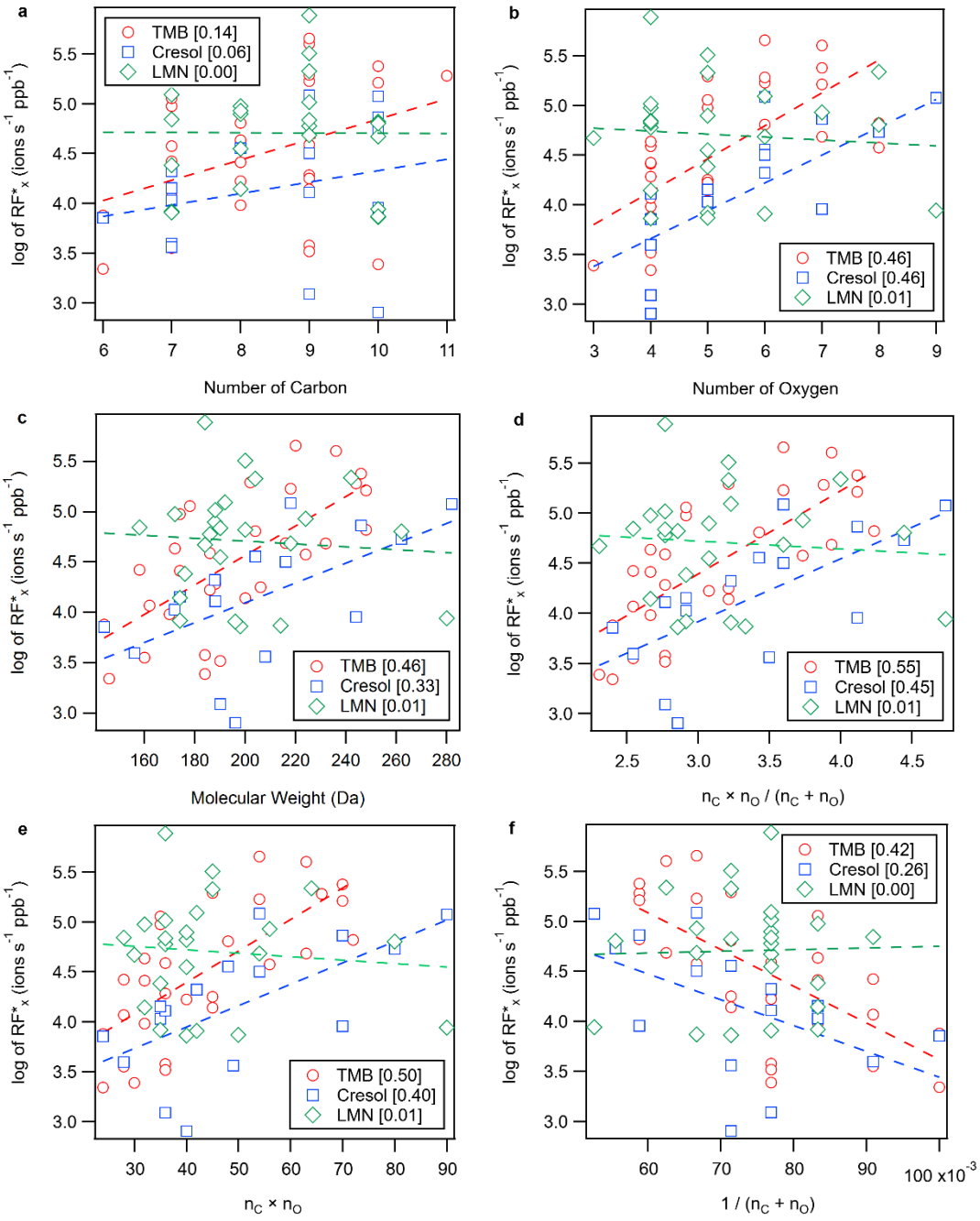


Figure S8. Preliminary regression analysis

OLS regression analysis of the log of RF^*_x with respect to (a) the number of carbon, n_C, (b) the number of oxygen, n_O, (c) the molecular weight, MW, (d) the carbon-oxygen non-ideality, NI_{CO}, (e) the product of n_C and n_O, P_{CO}, and (*f*) the inverse of the sum of n_C and n_O, I_{CO}. The red, blue, and green dashed lines correspond to the linear fitting lines for the log(RF^*_x) values of TMB, cresol, and LMN oxidation products, respectively. The coefficient of determination, R^2 of ordinary linear regression for the log(RF^*_x) as a function of the feature is shown in brackets after the corresponding VOC label.

The full regression analysis was performed on two types of datasets: The log of measured EESI sensitivity in ions s^{-1} ppb^{-1}, log(RF^*_x) from (1) the TMB system alone, or (2) all three VOC systems. Two approaches were taken for the combined dataset: (2a) the precursor

VOC identity was not included as a feature or (2b) the VOC identity was digitized and included as a feature. Based on the trends observed for the relative isomer sensitivities, which were highest for the oxidation products of limonene followed by those of TMB and of cresol, the VOC identity feature value for limonene, TMB, and cresol data were digitized as 1, 0, and -1.

First, an exhaustive search over the feature space was performed to determine the optimal set of features for each regressor using their respective default hyperparameter values. Leave-one-out (LOO) cross-validation was used to evaluate the model performance in terms of the coefficient of determination, R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \text{Eq. (S21)}$$

where y_i and $\hat{y}_i$ are the true and the predicted value for the i-th sample among a total of n samples, and $\bar{y}$ is the mean of the n samples. If the model always predicts $\bar{y}$, the R^2 will be 0, e.g. a naive model where all values are predicted to equal that of the sample mean regardless of input. The R^2 can be negative if it performs worse than this naive model, i.e. assuming the mean value regardless of model input produces on average better results. For a dataset of size n, LOO involves setting aside each data point (y_i) in turn as the test sample while the remaining (n-1) data points are used to train the model and make a prediction, $\hat{y}_i$. y_i and $\hat{y}_i$ are then used to estimate the R^2 using Eq. (S21). LOO can be considered as performing a K-fold cross-validation where the number of K is equal to the number of data points. Compared to the K-fold cross-validation method, LOO is more computationally intensive to perform, but is nonetheless appropriate given the small size of the dataset used here (n_{sample} = 30 for case 1 and 70 for 2a and 2b). During cross-validation, a portion of the dataset is used to train the model (i.e. "train" set), while the remaining dataset is withheld to validate against the model predictions (i.e. "test" set). For each train-test set, the training feature values ($n = n_{sample} - 1$) were standardized, which involves subtracting by their mean and dividing by their standard deviation. The same transformation was then applied to the feature values from the test set (n = 1), which was not included in deriving the transformation required for the standardization to prevent information leak between the training and test sets.

The results of the feature optimization are shown in Figure S8 in terms of the best R^2 vs. the number of features used. The optimal feature sets are shown in Table S2. In addition to OLS, linear ridge regression ("Ridge") and Bayesian ridge regression (BayRR) are included. Both Ridge and BayRR implement L_2 regularization, making them more resilient against overfitting and feature co-linearity. Support vector regression (SVR) with linear kernel is also included as a linear regression model for comparison. Exploratory analysis using SVR with radial basis functions (rbf) yielded better R^2, but the relative feature importance was not easily interpretable when rbf was used, hence the choice of linear kernel. Lastly, nonparametric regressions such as random forest regressor (RFR) and gradient boosting regressor (GBR) were included, as the RF^*_x is likely not a linear function of features already included. While it is possible that RF^*_x could be well-described by a linear combination of engineered features, it is not feasible to explore all nonlinear (e.g. n_C^2) or interaction ($n_C n_H$) feature terms, hence the necessity of nonparametric regressors.

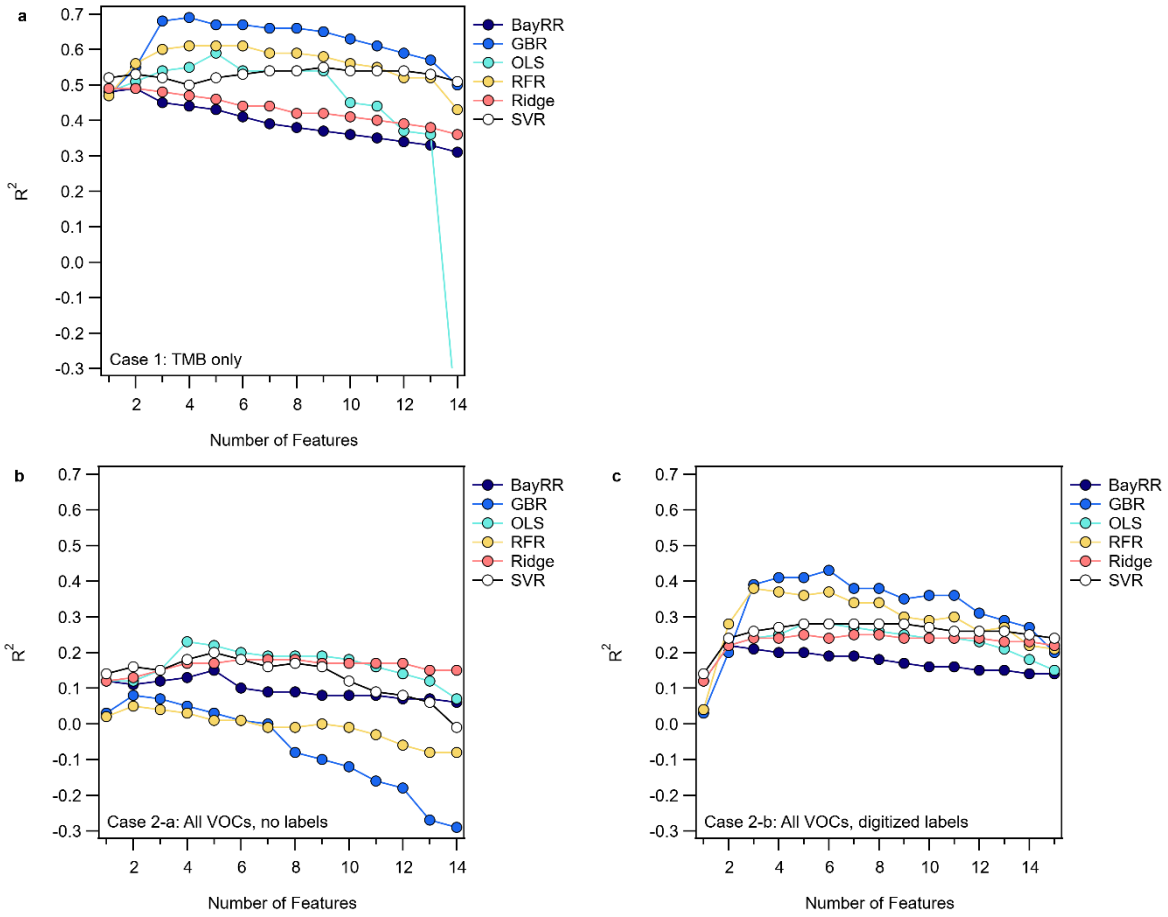


Figure S9. Feature selection

The best R^2 from LOO cross-validation test for each regressor using different permutations of features as a function of the number of features included for (a) Case 1, where only the TMB dataset is used, (b) Case 2a, where data from all the VOC systems were used without providing the digitized VOC identity as one of the input features, or (c) Case 2b, where data from all the VOC systems were used with the digitized VOC identity provided as one of the input features, hence the one extra feature over cases 1 and 2a.

Table S2. Best R^2 scores and their corresponding feature combination obtained using leave-one-out cross-validation with default model hyperparameters

Case #	OLS	Ridge	BRR	SVR	RFR	GBR
1	0.59 NI_{CO}, P_{CO}, $\log(C^*)$, X_C, DBEpC	0.49 NI_{CO}	0.49 NI_{CO}, $\log(C^*)$	0.53 NI_{CO}, $\log(C^*)$	0.61 NI_{CO}, n_O, n_H, H:C	0.69 NI_{CO}, P_{CO}, n_H, H:C
2a	0.23 NI_{CO}, P_{CO}, n_H, X_C	0.17 NI_{CO}, P_{CO}, X_C, DBEpC	0.15 NI_{CO}, P_{CO}, n_H, X_C, mW	0.20 NI_{CO}, P_{CO}, $\log(C^*)$, X_C, DBEpC	0.05 n_C, X_C	0.08 I_{CO}, X_C
2b	0.25 NI_{CO}, P_{CO}, O:C, VOC	0.24 NI_{CO}, P_{CO}, $\log(C^*)$, VOC	0.22 NI_{CO}, VOC	0.27 NI_{CO}, H:C, X_C, VOC	0.38 n_O, Δm, VOC	0.43 n_C, n_H, n_O, DBEpC, Δm, VOC

Note that in some cases (e.g. SVR in case 1), the optimal feature set selected does not correspond to the set with the highest R^2, but rather one with slightly lower R^2 score but also (sometimes substantially) lower total number of features used. The feature abbreviations used are as followed: Carbon-oxygen non-ideality (NI_{CO}), product of the number of oxygen and carbon numbers (P_{CO}), the inverse of the sum of the number of oxygen and carbon numbers (I_{CO}), logarithm of saturation vapor concentration ($\log(C^*)$), aromaticity (X_C), double bond equivalent per carbon (*DBEpC*), number of oxygen atoms (n_O), number of hydrogen atoms (n_H), molecular weight (*MW*), mass defect (Δm), hydrogen-to-carbon ratio (*H:C*), oxygen to carbon ratio (*O:C*), digitized precursor VOC label (VOC).

For Case 1, NI_{CO} is unanimously identified as an essential feature in predicting the (log of) EESI-TOF response factor. For Case 2a, the decision-tree model performs negligibly better at predicting the EESI-TOF sensitivity than simply assuming the dataset mean. The feature selection results for linear regression models suggest that NI_{CO}, P_{CO}, and X_C are essential features to include. In general, the regression model performances are poor for Case 2a, and it would be simpler to assume instead a uniform RF^*_x, for example the geometric mean of $10^{4.5}$ ions s^{-1} ppb^{-1}. For Case 2b, inclusion of the VOC label as the 2nd feature results in substantial increase in R^2 for all regressors, as shown in Figure S9 and Table S3 below. As the number of features increase beyond 3, regressor performances do not show any substantial improvement and may even deteriorate. Comparing the regressor performances for Case 2a and 2b, we see that nonparametric models (i.e. RFR and GBR) benefit much more from the inclusion of the VOC precursor as an input feature than do the linear regression models. This is partially due to the specific digitization applied to the VOC precursor (-1 for cresol, 0 for TMB, 1 for limonene), which come with *a priori* ranking and weighting information that would affect the linear models more so than the decision-tree type models.

Table S3. Best R^2 scores for different feature combinations obtained using leave-one-out cross-validation with default model hyperparameters for Case 2b

Feature #	OLS	Ridge	BRR	SVR	RFR	GBR
1	0.12 I_{CO}	0.12 I_{CO}	0.12 I_{CO}	0.14 mW	0.04 n_C	0.03 n_C
2	0.22 NI_{CO}, VOC	0.22 NI_{CO}, VOC	0.22 NI_{CO}, VOC	0.24 NI_{CO}, VOC	0.28 mW, VOC	0.20 Δm, VOC
3	0.24 NI_{CO}, $\log(C^*)$, VOC	0.24 NI_{CO}, P_{CO}, $\log(C^*)$, VOC	0.21 NI_{CO}, $\log(C^*)$, VOC	0.26 NI_{CO}, $\log(C^*)$, VOC	0.38 n_O, Δm, VOC	0.39 n_O, Δm, VOC

Having identified the optimal feature sets, we then performed grid search to find the optimal model hyperparameters using R^2 from *LOO* as the metric. The hyperparameter spaces explored for each regressor are listed in Table S4a-c below, along with the *LOO* R^2 obtained using the default vs. the optimal model hyperparameters.

Table S4a. Regressor hyperparameter grid search results for Case 1

Regressor	Hyperparameter Space	Optimal	R^2 (Optimal)	R^2 (Default)
RFR	n_estimator: [10, 20, 30, 40, 50,100]	10	0.60	0.61
	min_samples_split: [2, 3, 4, 5]	4		
	max_features: [“auto”, “sqrt”, “log2”]	“auto”		
SVR	C: [0.1, 0.2, 0.5, 1, 2,10]	1	0.53	0.53
	epsilon: [0.1, 0.2, 0.5, 1, 10, 100]	0.1		
GBR	n_estimator: [5, 10, 20, 30, 40, 50,100]	100	0.76	0.69
	loss: [“ls”, “lad”, “huber"]	ls		
	learning_rate: [0.05, 0.1, 0.2, 0.5]	0.1		
	subsample: [0.3, 0.5, 0.7, 1]	0.3		
	max_features: [“auto”, “sqrt”, “log2”]	“auto”		
	min_samples_split: [2, 3, 4, 5]	4		
BRR	n_iter: [100, 200, 500, 1000]	100	0.49	0.49
	alpha_1: [10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}]	10^{-4}		
	alpha_2: [10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}]	10^{-8}		
	lambda_1: [10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}]	10^{-8}		
	lambda_2: [10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}]	10^{-4}		
Ridge	alpha: [0.1, 0.2, 0.5, 1, 2, 10, 100]	1	0.49	0.49

Note: Optimal hyperparameter values that are identical to the default values are underlined. For decision-tree type models such as GBR and RFR, the model can vary from run to run, and the “optimal” hyperparameter values may produce worse scores than the default case by chance.

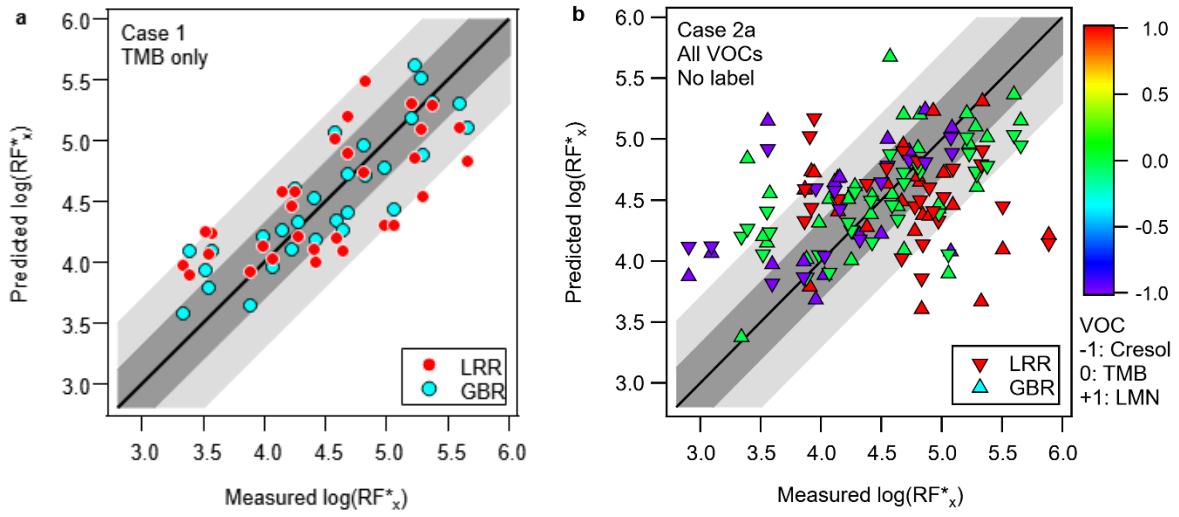
Table S4b. Regressor hyperparameter grid search results for Case 2a

Regressor	Hyperparameter Space	Optimal	R^2 (Optimal)	R^2 (Default)
RFR	n_estimator: [10, 20, 30, 40, 50,100]	20	0.10	0.05
	min_samples_split: [2, 3, 4, 5]	4		
	max_features: [“auto”, “sqrt”, “log2”]	“auto”		
SVR	C: [0.1, 0.2, 0.5, 1, 2,10]	2	0.20	0.20
	epsilon: [0.1, 0.2, 0.5, 1, 10, 100]	0.1		
GBR	n_estimator: [5, 10, 20, 30, 40, 50,100]	40	0.00	0.08
	loss: [“ls”, “lad”, “huber"]	ls		
	learning_rate: [0.05, 0.1, 0.2, 0.5]	0.5		
	subsample: [0.3, 0.5, 0.7, 1]	0.7		
	max_features: [“auto”, “sqrt”, “log2”]	“auto”		
	min_samples_split: [2, 3, 4, 5]	4		
BRR	n_iter: [100, 200, 500, 1000]	100	0.15	0.15
	alpha_1: [10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}]	10^{-4}		
	alpha_2: [10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}]	10^{-8}		
	lambda_1: [10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}]	10^{-8}		
	lambda_2: [10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}]	10^{-4}		
Ridge	alpha: [0.1, 0.2, 0.5, 1, 2, 10, 100]	0.2	0.18	0.17

Table S4c. Regressor hyperparameter grid search results for Case 2b

Regressor	Hyperparameter Space	Optimal	R^2 (Optimal)	R^2 (Default)
RFR	n_estimator: [10, 20, 30, 40, 50,100]	100	0.38	0.38
	min_samples_split: [2, 3, 4, 5]	2		
	max_features: [“auto”, “sqrt”, “log2”]	“auto”		
SVR	C: [0.1, 0.2, 0.5, 1, 2,10]	0.5	0.28	0.27
	epsilon: [0.1, 0.2, 0.5, 1, 10, 100]	0.2		
GBR	n_estimator: [5, 10, 20, 30, 40, 50,100]	100	0.48	0.43
	loss: [“ls”, “lad”, “huber"]	ls		
	learning_rate: [0.05, 0.1, 0.2, 0.5]	0.7		
	subsample: [0.3, 0.5, 0.7, 1]	1		
	max_features: [“auto”, “sqrt”, “log2”]	“auto”		
	min_samples_split: [2, 3, 4, 5]	3		
BRR	n_iter: [100, 200, 500, 1000]	100	0.22	0.22
	alpha_1: [10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}]	10^{-4}		
	alpha_2: [10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}]	10^{-8}		
	lambda_1: [10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}]	10^{-8}		
	lambda_2: [10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}]	10^{-4}		
Ridge	alpha: [0.1, 0.2, 0.5, 1, 2, 10, 100]	0.2	0.24	0.24

The log(RF^*_x) predicted from the LOO cross-validation test (see discussion around Eq. S21) by the linear ridge regressor (LRR) and the gradient boosting regressor (GBR) using their respective optimal features sets and hyperparameters for Cases 1, 2a, and 2b are shown in Figure S10 and compared to the measured log(RF^*_x). For a single VOC system (Case 1), the predicted and measured RF^*_x values mostly agree within a factor of 5 using LRR or a factor of 2 using GBR. When dealing with compounds from multiple VOC systems, where the VOC precursor identities are unknown (i.e. Case 2a), the predictions do not fare better than simply assuming a uniform response factor equal to that of the sample mean, as shown in Figure S10b. If the VOC precursor identity is used as one of the features, GBR can produce reasonable predictions that agree with the measured values within a factor of 2-5, as shown in Figure S10c, where most of the scatter outside this range was related to the limonene dataset, which did not appear to have a clear predictor for log(RF_x^*), as we have also shown during our preliminary regression analysis in Figure S8.



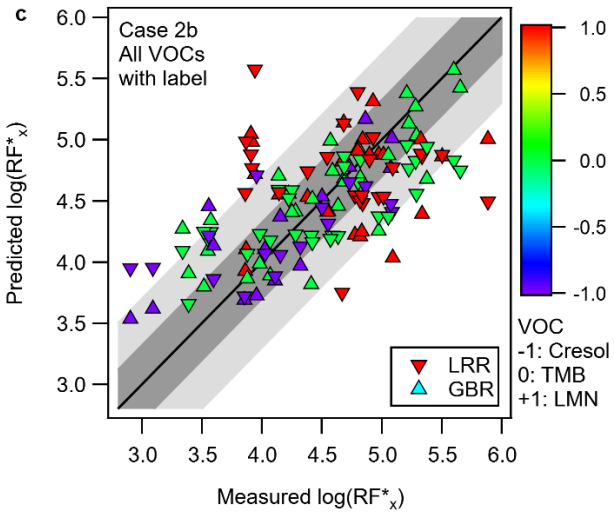


Figure S10. Comparison of model performance for different cases

Comparison of the log of the measured response factor, log(RF^*_x) with those predicted using the leave-one-out cross-validation method by the linear ridge regressor (LRR) and the gradient boosting regressor (GBR) using their optimal feature sets and hyperparameters for (a) Case 1, (b) Case 2a, and (c) Case 2b. The digitized VOC labels are shown in color scale in (b) and (c) for combined VOC scenarios, but it was made available to the regression models to use as a potential feature for Case 2b in (c). The 1-to-1 line is shown in solid black. The darker shaded region represents a factor of 2 deviation from the 1-to-1 line. The lighter shaded region represents a factor of 5 deviation from the 1-to-1 line.

Table S5. R^2 for each regressor using their optimal features and model hyperparameters, and the weights/importance of fitted features.

Case #	OLS	Ridge	BRR	SVR	RFR	GBR
1	0.59 NI_{CO}: 2.73 P_{CO}: -1.77 log(C^*): 0.60 X_C: 0.50 DBEpC: -0.51	0.49 NI_{CO}: 0.47	0.49 NI_{CO}: 0.70 log(C^*): 0.23	0.53 NI_{CO}: 0.73 log(C^*): 0.26	0.61 NI_{CO}: 0.42 n_O: 0.28 n_H: 0.16 H:C: 0.15	0.69 NI_{CO}: 0.32 P_{CO}: 0.24 n_H: 0.20 H:C: 0.24
2a	0.23 NI_{CO}: 2.38 P_{CO}: -2.35 n_H: 0.53 X_C: 0.39	0.17 NI_{CO}: 1.05 P_{CO}: -0.88 X_C: 0.47 DBEpC: -0.49	0.15 NI_{CO}: 1.76 P_{CO}: -1.02 n_H: 0.58 X_C: 0.36 mW: -0.74	0.20 NI_{CO}: 0.54 P_{CO}: -0.28 log(C^*): 0.07 X_C: 0.36 DBEpC: -0.43	0.10 n_C: 0.52 X_C: 0.47	0.00 I_{CO}: 0.64 X_C: 0.36
2b	0.25 NI_{CO}: 2.11 P_{CO}: -1.64 O:C: -0.31 VOC: 0.24	0.24 NI_{CO}: 1.25 P_{CO}: -0.46 log(C^*): 0.51 VOC: 0.24	0.22 NI_{CO}: 0.27 VOC: 0.22	0.27 NI_{CO}: 0.30 H:C: 0.19 X_C: 0.24 VOC: 0.23	0.38 n_O: 0.28 Δm: 0.43 VOC: 0.30	0.49 n_C: 0.10 n_H: 0.03 n_O: 0.23 DBEpC: 0.17 Δm: 0.20 VOC: 0.26

The R^2 determined from the leave-one-out (LOO) cross-validation test is shown. For ordinary least square (OLS) regression, linear ridge regression (LRR), Bayesian ridge regression (BRR), and support vector regression (SVR), the weight for each feature is shown. For random forest (RFF) and gradient boosting regression (GBR), the importance is shown, which is a measure of the usefulness of a feature in constructing the decision tree.

Note that if we were to use the entire dataset to train and validate the model, the resulting R^2 would be overly optimistic, as shown in Figure S11 especially for those obtained using the nonparametric models.

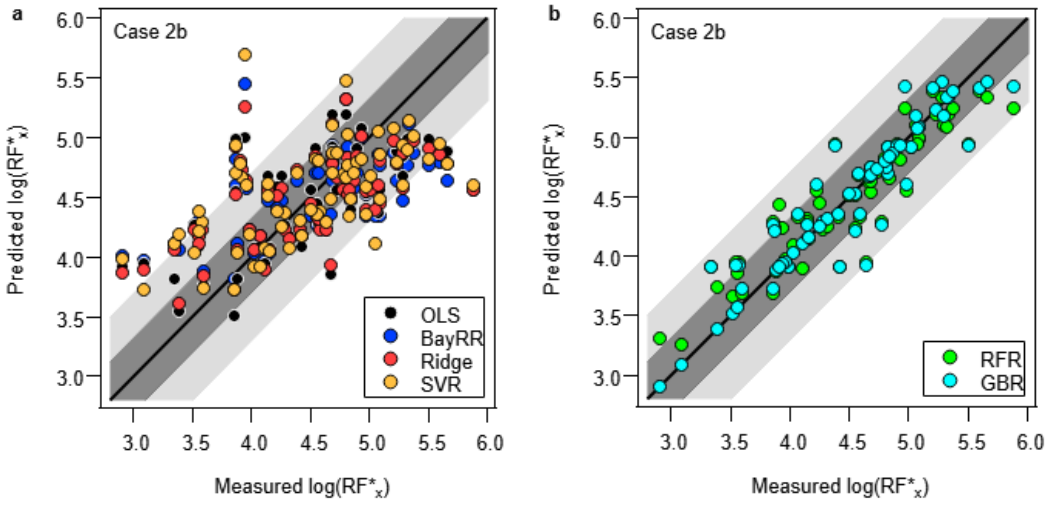
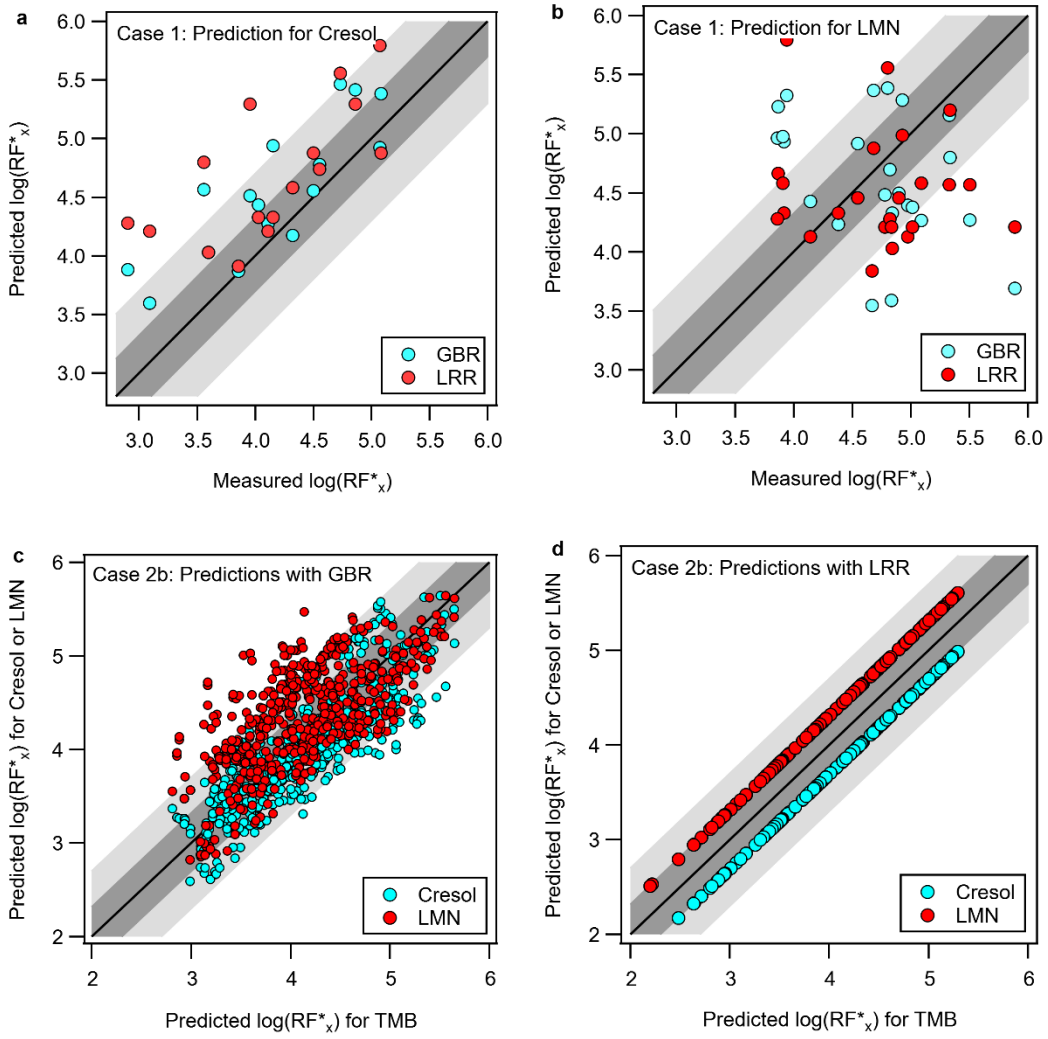


Figure S11. Regression using the entire dataset
Comparison of the predicted log(RF^*_x) using the entire dataset with VOC label included as one of the features using (a) linear regression models and (b) nonparametric regression models. The optimal feature sets and hyperparameters used for each model are identical to those used for Figure S10 and Table S5, except that now each model was trained with the entire dataset to predict the entire dataset, instead of following the LOO procedure. The 1-to-1 line is shown in solid black. The darker shaded region represents a factor of 2 deviation from the 1-to-1 line. The lighter shaded region represents a factor of 5 deviation from the 1-to-1 line.

For typical ambient measurements or chamber experiments with complex precursor mixtures, the VOC precursor identity is often not known without additional constraints (e.g. ion mobility or gas chromatography measurements supported with chemical reaction box models). The prediction capability of the regression model for an unknown VOC is examined in Figures S12a and S12b, using the TMB dataset as the "known" VOC system to predict the log(RF^*_x) for the "unknown" cresol and limonene (LMN) systems. As shown in Figure S12a, while the regression models trained with TMB dataset tend to overestimate the log(RF^*_x) for the cresol system, the predictions and observations are qualitatively consistent in terms of the relative log(RF^*_x), likely due to the structural similarity of cresol and TMB, which would be reflected to varying degrees in their respective oxidation products. In contrast, regression models trained with the TMB dataset are unfit to predict the log(RF^*_x) for the limonene oxidation products, as shown in Figure S12b.

The effect of the VOC precursor on the predicted log(RF^*_x) values, using the model trained in Case 2b (all data with digitized VOC label), for all CHO molecular formulae used for EESI-TOF spectral fitting is shown in Figures S12c and S12d. In general, the predicted log(RF^*_x) trend in the same direction for all VOCs. The predicted effect of VOC precursor is distinct when a linear regressor is used, as shown in Figure S12d, where log(RF^*_x) is treated as a linear combination of features, one of which is the digitized VOC precursor identity. When a decision-tree type regressor is used, the VOC precursor identity effect is not as obvious, as

shown in Figure S12c. Lastly, the combination of dataset from multiple VOC systems also affects the predicted log(RF^*_x), as shown in Figure S12e and S12f for the TMB system. Models trained with the combined dataset (i.e. Case 2b) appear to underestimate the log(RF^*_x) as compared to the models trained with a single VOC dataset (i.e. Case 1). Furthermore, regressors that performed reasonably well (e.g. LRR for Case 1) for the training dataset with a limited number of features (e.g. NI_{CO}) may be ill-equipped when predicting for a more diverse set of compounds, whose variabilities are only reflected in other features (e.g. optimal features for LRR in Case 2b, see Table S5).



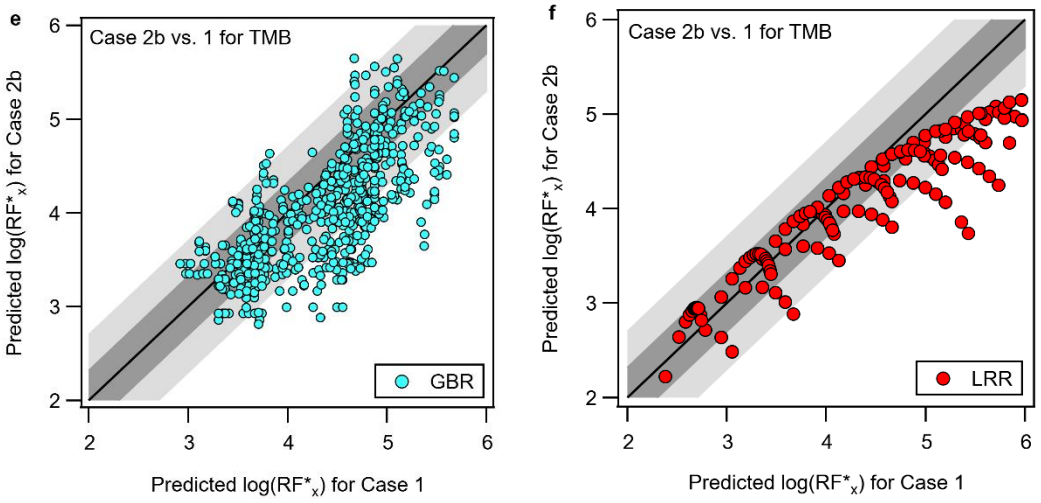


Figure S12. Prediction of all RF^*_x

(a) Comparison of the observed log(RF^*_x) for cresol oxidation products with that predicted using gradient boosting regression (GBR) and the linear ridge regression (LRR) models trained with the TMB dataset (b) Same as (a) but for the limonene (LMN) system. (c) Comparison of the log(RF^*_x) for all molecular formulae used for EESI-TOF MS fitting predicted using the GBR model from Case 2b for different VOC systems, i.e. all feature values used during prediction were identical expect for that of the digitized VOC precursor identity. (d) Same as (c), but with the LRR model from Case 2b. (e) Comparison of the log(RF^*_x) for all molecular formulae used for EESI-TOF MS fitting predicted using the GBR model from Case 1 and Case 2b for TMB system only. (f) Same as (e), but with the LRR model from Case 1 and Case 2b. The optimal feature sets and hyperparameters used for each model are listed in Table S5. The 1-to-1 line is shown in solid black. The darker shaded region represents a factor of 2 deviation from the 1-to-1 line. The lighter shaded region represents a factor of 5 deviation from the 1-to-1 line. The case number indicated on the axis legend and in annotations indicate the how the model was trained as described throughout Tables S2-5.

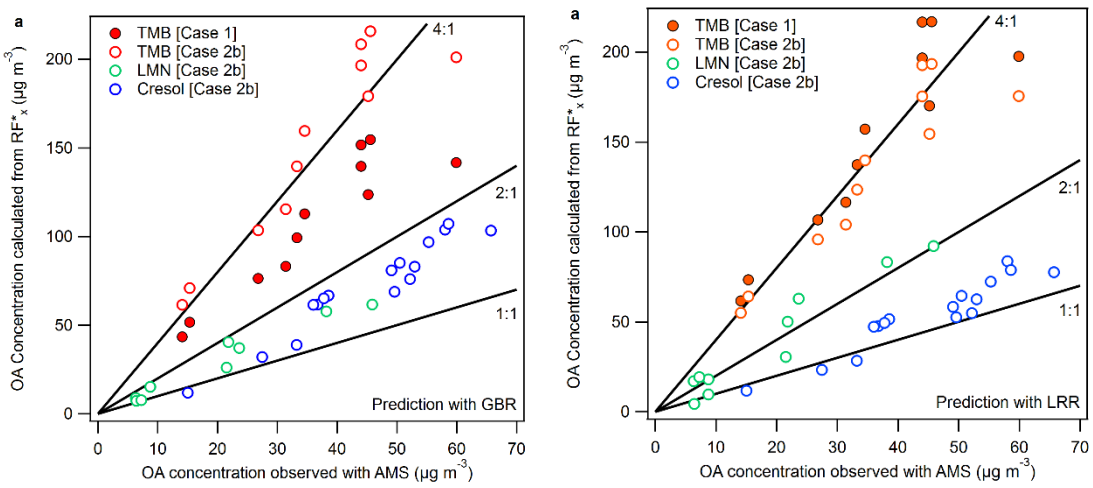


Figure S13. Comparison of estimated and observed OA concentration

(a) Comparison of the observed organic aerosol (OA) as measured by the AMS with the OA concentration estimated using EESI-TOF measurements converted from ions s^{-1} to μg m^{-3} using the RF^*_x (ions s^{-1} ppb^{-1}) predicted using the gradient boosting regression (GBR) model. (b) Same as (a), but using the RF_x^* predicted by the linear ridge regression model (LRR).

Conversion of ppb to molecules cm^{-3} is performed under standard conditions, i.e. $2.46 \cdot 10^{10}$ molecules cm^{-3} per ppb. The 1-to-1, 2-to-1, and 4-to-1 lines are shown in solid black. Two versions of the regression models are used to predicted the RF^*_x for TMB, one trained with single VOC dataset (Case 1) and one trained with combined VOC datasets where the VOC precursor identity is used as a training feature (Case 2b).

References

Brune, W. H.: The Chamber Wall Index for Gas-Wall Interactions in Atmospheric Environmental Enclosures, Environ. Sci. Technol., 53(7), 3645–3652, doi:10.1021/acs.est.8b06260, 2019.

Dal Maso, M., Kulmala, M., Lehtinen, K. E. J., Ḿkelä, J. M., Aalto, P. and O'Dowd, C. D.: Condensation and coagulation sinks and formation of nucleation mode particles in coastal and boreal forest boundary layers, J. Geophys. Res. Atmos., 107(19), doi:10.1029/2001JD001053, 2002.

Fuller, E. N., Schettler, P. D. and Giddings, J. C.: A new method for prediction of binary gas-phase diffusion coefficients, Ind. Eng. Chem., 58(5), 18–27, doi:10.1021/ie50677a007, 1966.

George, I. J., Vlasenko, A., Slowik, J. G., Broekhuizen, K. and Abbatt, J. P. D.: Heterogeneous oxidation of saturated organic aerosols by hydroxyl radicals: Uptake kinetics, condensed-phase products, and particle size change, Atmos. Chem. Phys., 7(16), 4187–4201, doi:10.5194/acp-7-4187-2007, 2007.

Holzinger, R., Joe Acton, W. F., Bloss, W. W., Breitenlechner, M., Crilley, L. L., Dusanter, S., Gonin, M., Gros, V., Keutsch, F. F., Kiendler-Scharr, A., Kramer, L. L., Krechmer, J. J., Languille, B., Locoge, N., Lopez-Hilfiker, F., Materi, D., Moreno, S., Nemitz, E., Quéléver, L. L., Sarda Esteve, R., Sauvage, S., Schallhart, S., Sommariva, R., Tillmann, R., Wedel, S., Worton, D. D., Xu, K. and Zaytsev, A.: Validity and limitations of simple reaction kinetics to calculate concentrations of organic compounds from ion counts in PTR-MS, Atmos. Meas. Tech., 12(11), 6193–6208, doi:10.5194/amt-12-6193-2019, 2019.

Jennings, S. G.: The mean free path in air, J. Aerosol Sci., 19(2), 159–166, doi:10.1016/0021-8502(88)90219-4, 1988.

Krechmer, J. E., Pagonis, D., Ziemann, P. J. and Jimenez, J. L.: Quantification of Gas-Wall Partitioning in Teflon Environmental Chambers Using Rapid Bursts of Low-Volatility Oxidized Species Generated in Situ, Environ. Sci. Technol., 50(11), 5757–5765, doi:10.1021/acs.est.6b00606, 2016.

Krechmer, J. E., Day, D. A., Ziemann, P. J. and Jimenez, J. L.: Direct Measurements of Gas/Particle Partitioning and Mass Accommodation Coefficients in Environmental Chambers, Environ. Sci. Technol., 51(20), 11867–11875, doi:10.1021/acs.est.7b02144, 2017.

Kulmala, M. and Wagner, P. E.: Mass accommodation and uptake coefficients - A quantitative comparison, J. Aerosol Sci., 32(7), 833–841, doi:10.1016/S0021-8502(00)00116-6, 2001.

Lehtinen, K. E. J., Korhonen, H., Dal Maso, M. and Kulmala, M.: On the concept of condensation sink diameter, Boreal Environ. Res., 8(4), 405–411 [online] Available from: http://www.borenv.net/BER/pdfs/ber8/ber8-405.pdf (Accessed 22 May 2014), 2003.

Liu, X., Day, D. A., Krechmer, J. E., Brown, W., Peng, Z., Ziemann, P. J. and Jimenez, J. L.: Direct measurements of semi-volatile organic compound dynamics show near-unity mass accommodation coefficients for diverse aerosols, Commun. Chem., 2(1), 98, doi:10.1038/s42004-019-0200-x, 2019.

Palm, B. B., Campuzano-Jost, P., Ortega, A. M., Day, D. A., Kaser, L., Jud, W., Karl, T., Hansel, A., Hunter, J. F., Cross, E. S., Kroll, J. H., Peng, Z., Brune, W. H. and Jimenez, J. L.:

In situ secondary organic aerosol formation from ambient pine forest air using an oxidation flow reactor, Atmos. Chem. Phys., 16(5), 2943–2970, doi:10.5194/acp-16-2943-2016, 2016.

Pieber, S. M., El Haddad, I., Slowik, J. G., Canagaratna, M. R., Jayne, J. T., Platt, S. M., Bozzetti, C., Daellenbach, K. R., Fröhlich, R., Vlachou, A., Klein, F., Dommen, J., Miljevic, B., Jiménez, J. L., Worsnop, D. R., Baltensperger, U. and Prévôt, A. S. H.: Inorganic Salt Interference on CO2+ in Aerodyne AMS and ACSM Organic Aerosol Composition Studies, Environ. Sci. Technol., 50(19), 10494–10503, doi:10.1021/acs.est.6b01035, 2016.

Tang, M. J., Shiraiwa, M., Pöschl, U., Cox, R. A. and Kalberer, M.: Compilation and evaluation of gas phase diffusion coefficients of reactive trace gases in the atmosphere: Volume 2. Diffusivities of organic compounds, pressure-normalised mean free paths, and average Knudsen numbers for gas uptake calculations, Atmos. Chem. Phys., 15(10), 5585–5598, doi:10.5194/acp-15-5585-2015, 2015.

Yassine, M. M., Harir, M., Dabek-Zlotorzynska, E. and Schmitt-Kopplin, P.: Structural characterization of organic aerosol using Fourier transform ion cyclotron resonance mass spectrometry: aromaticity equivalent approach, Rapid Commun. Mass Spectrom., 28(22), 2445–2454, doi:10.1002/rcm.7038, 2014.