

Supplement for:

A Novel Approach Combining Mechanism and Machine Learning to Unravel Oxidant-Driven Formation Regimes of α -Pinene-Derived Oxygenated Organic Molecules (OOMs)

Rui Tan^{1,2}, Xinshi Ni¹, Zichao Wan¹, Xudong Tian³, Jiasi Shen³, Bingye Xu³, Da Xu³, Sihua Lu¹, Xin Li¹, Song Guo^{1,2*}

Correspondence: Song Guo (songguo@pku.edu.cn)

S1 Precursors, oxidants, and OOMs abundance

The general meteorological conditions, trace gas concentrations, and total OOMs mass concentrations during the campaign (20 April to 28 May 2024) are summarized in Table S1 and illustrated in Figure S1. During the observation period, continental air masses from the north and marine air masses from the east alternately influenced the measurement site. The continental air masses were generally more polluted, whereas the marine air masses were relatively clean. This meteorological variability drove dynamic fluctuations in the chemical regime, with average O₃ and NO_x concentrations of 50.10 ± 17.15 ppb and 8.67 ± 3.79 ppb, respectively. Meanwhile, monoterpene mixing ratios exhibited substantial diurnal and episodic variability (ranging from 8.62 to 794.40 ppt, averaging 102.41 ± 99.49 ppt), primarily driven by the interplay between nighttime accumulation in a shallow boundary layer and rapid daytime oxidative depletion.

Table S1 Mean values of meteorological variables, trace gases, and monoterpene

Parameters	Mean $\pm$ SD	Unit
Meteorological		
Temperature	19.72 $\pm$ 2.86	°C
Relative Humidity (RH)	86.69 $\pm$ 15.30	%
Pressure	1011.77 $\pm$ 4.05	hPa
Trace Gases & Aerosols		
NO	1.32 $\pm$ 0.54	ppb
NO ₂	7.35 $\pm$ 3.56	ppb
NO _x	8.67 $\pm$ 3.79	ppb
O ₃	50.10 $\pm$ 17.15	ppb
PM _{2.5}	16.23 $\pm$ 10.44	$\mu\text{g m}^{-3}$
Volatile Organic Compounds		
monoterpene	102.41 $\pm$ 99.49	ppt

Despite the continuous availability of precursors and oxidants, a visual inspection of the time series (Figure S1) shows a notable feature: the total OOMs abundance (ranging from 0.05 to 0.46 $\mu\text{g m}^{-3}$) did not simply co-vary with the linear abundance of monoterpene or O₃ (Pearson $r = 0.297$ and 0.369 , respectively; $p < 0.001$). This moderate correlation suggests that while precursor and oxidant concentrations exert some influence, OOMs production cannot be explained by a simple linear dependency on any single variable. Instead, it is governed by complex, multi-factorial interactions (Hallquist et al., 2009).

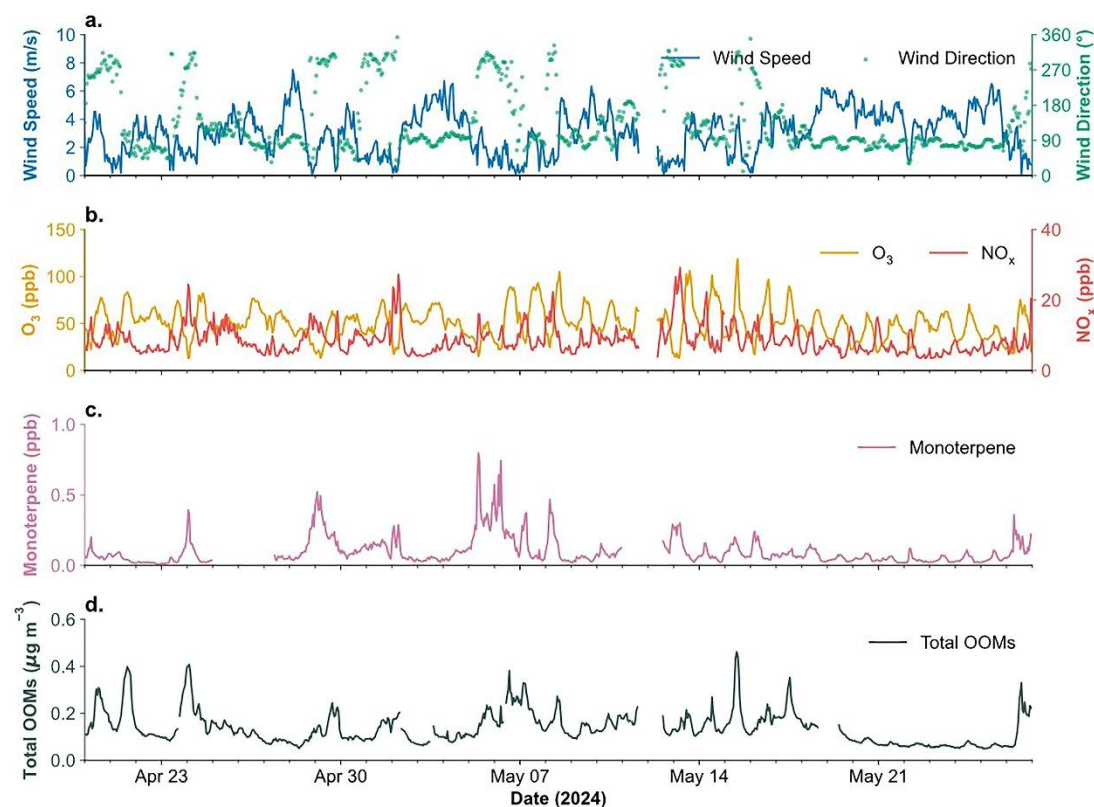


Figure S1. Time series of (a) wind speed (m/s) and wind direction (°), (b) O₃ and NO_x concentrations (both in ppb), (c) monoterpene concentration (ppb), and (d) total OOMs concentration (μg m⁻³) during the campaign.

All four OOMs classes followed a broadly similar diurnal pattern, with concentrations increasing from morning to a clear maximum around noon (12:00–13:00 LST) and decreasing afterwards. However, the amplitude of the diurnal cycle and the absolute concentrations differed markedly among classes. H14, the most unsaturated class (C₁₀H₁₄), showed the highest daytime concentration (~0.065 μg m⁻³) and the strongest relative increase from nighttime minimum (0.045 μg m⁻³) to peak (~44% increase). H16, which retains the parent monoterpene hydrogen count (C₁₀H₁₆), also exhibited a

pronounced midday peak ($0.045 \mu\text{g m}^{-3}$), with a larger relative increase ($\sim 60\%$ from its nighttime minimum of $0.028 \mu\text{g m}^{-3}$). The more saturated classes, H18 and H20, showed weaker diurnal amplitudes. H18 increased from a nighttime low of $\sim 0.030 \mu\text{g m}^{-3}$ to a midday peak of $\sim 0.042 \mu\text{g m}^{-3}$ ($\sim 40\%$ increase), while H20, the most saturated class ($\text{C}_{10}\text{H}_{20}$), displayed the smallest diurnal variation, rising from $\sim 0.018 \mu\text{g m}^{-3}$ at night to $\sim 0.025 \mu\text{g m}^{-3}$ at noon ($\sim 39\%$ increase). Notably, all classes exhibited their maxima at the same time of day (12:00–13:00 LST), coinciding with peak O_3 and maximum photochemical activity.

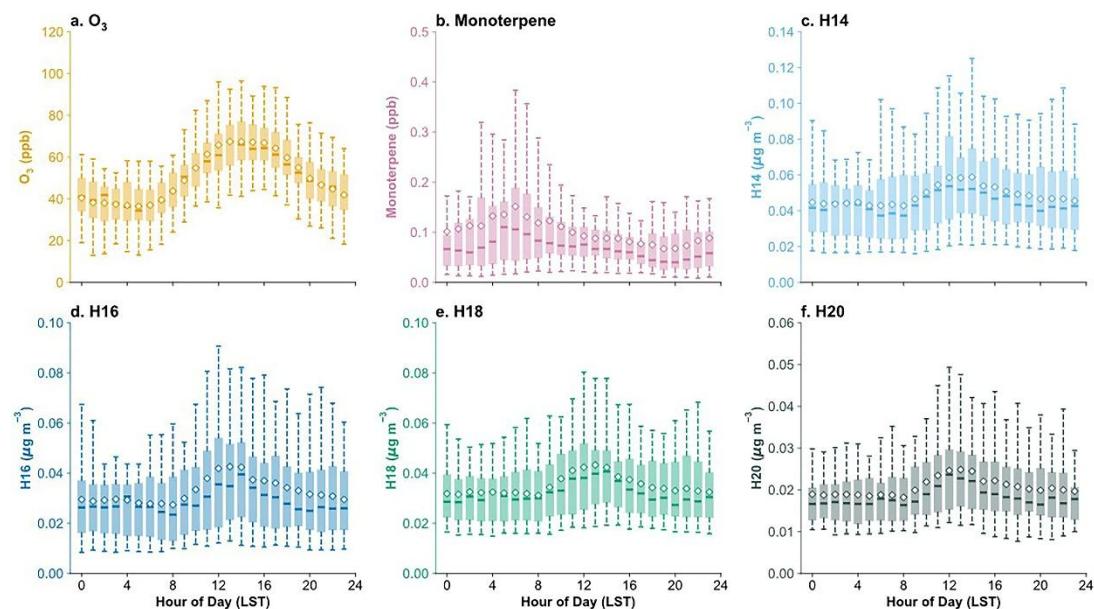


Figure S2. Diurnal variations of (a) O_3 (ppb), (b) monoterpene (ppb), and OOMs classes (c) H14, (d) H16, (e) H18, and (f) H20 (all in $\mu\text{g m}^{-3}$) during the campaign. The x-axis represents the Hour of Day (Local Standard Time, LST). The boxes indicate the 25th and 75th percentiles, the solid lines inside the boxes represent the medians, and the whiskers represent the data range.

Figure S3 presents a representative high-resolution mass spectrum of the monoterpene-

derived OOMs measured by the FIGAERO-I-CIMS. The spectrum displays the iodide adduct ions $[M+I]^-$, with the top 10 most abundant species labeled based on their relative signal intensities. The base peak at $m/z \sim 342$ corresponds to the $C_{10}H_{14}O_5$ species, indicating that this initial, highly unsaturated oxidation product is the most abundant OOM detected during the campaign. As shown in the spectrum, the detected OOMs are primarily characterized by molecular formulas ranging from $C_{10}H_{14}O_5$ to $C_{10}H_{20}O_{10}$. A general trend is observable: the relative abundance tends to decrease as the number of oxygen atoms increases (e.g., from O_5 to O_{10}), and as the degree of hydrogen saturation increases (from H_{14} to H_{20}). This suggests that the primary or lightly oxidized products dominate the measured signal, while highly oxygenated and multi-generation aging products are present at lower relative intensities, which aligns with the expected oxidation progression in the ambient atmosphere (Ehn et al., 2014).

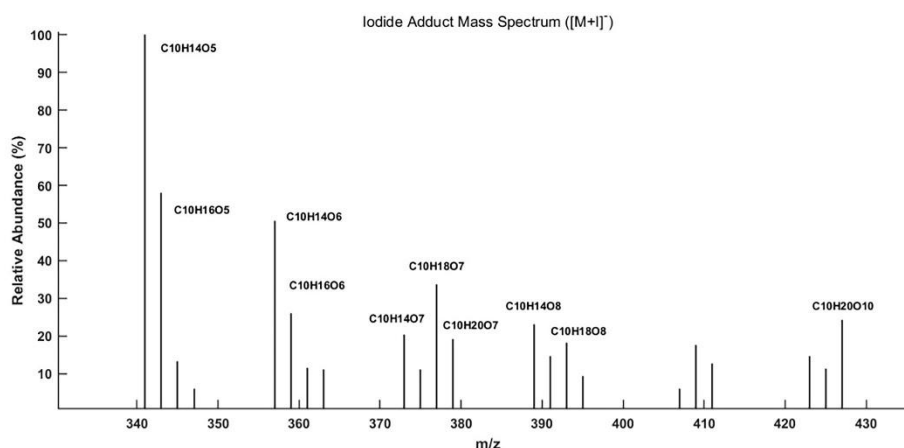


Figure S3. High-resolution mass spectrum of the monoterpene-derived OOMs measured by the FIGAERO-I-CIMS.

Figure S4 illustrates the average concentration proportion of the four main OOM

homologous series (H14, H16, H18, and H20) derived from the sum of gas- and particle-phase measurements over the entire observation period. The result reveals a distinct descending trend in abundance with increasing hydrogen saturation: the most unsaturated class, H14, exhibits the largest contribution at 37.8%, followed by H16 at 25.4%, H18 at 20.6%, and the most saturated class, H20, accounting for 16.2%. This compositional distribution is chemically consistent with the nature of monoterpene oxidation chemistry: initial O_3 or OH attack typically yields unsaturated products that retain a carbon skeleton similar to the parent $C_{10}H_{16}$, whereas the more saturated classes (H18 and H20) require multi-generation oxidation, reduction, or rearrangement processes (Atkinson and Arey, 2003). The relative predominance of H14 and H16 over H18 and H20 provides a useful baseline chemical context for interpreting the class-specific oxidant sensitivities and non-linear activation regimes revealed by our machine learning analysis.

Average Concentration Proportion (H14/H16/H18/H20)

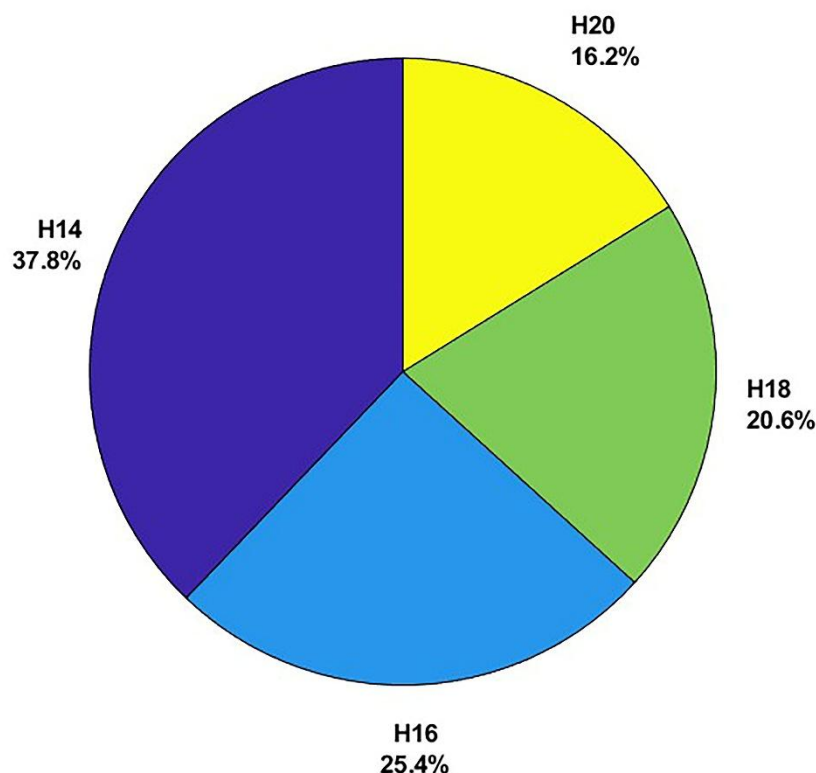


Figure S4. The average concentration proportion of the four main OOM homologous series (H14, H16, H18, and H20) during the entire campaign.

Based on the diurnal variation results, the diurnal variation of OOMs is relatively consistent with that of O_3 . The necessity of dissecting these formation pathways is further amplified when analyzing the day–night relationships between OOMs classes and O_3 (Figure S5). During daylight hours, all four classes show moderate correlations with O_3 ($R = 0.32$ – 0.38). However, because daytime O_3 strongly covaries with solar radiation and OH production, this correlation reflects O_3 acting simultaneously as a direct oxidant and a broader proxy for photochemistry.

A clearer signal comes from the nighttime data: the correlations between nighttime O_3 and all OOMs classes flatline to near zero ($R \leq 0.01$). The near-absence of nighttime correlation indicates that, under dark conditions, O_3 concentration alone does not account for the OOM variance; which factor does is not resolved by these data. The pronounced nighttime accumulation of monoterpene is consistent with ozonolysis being constrained by other chemical bottlenecks rather than by O_3 availability itself.

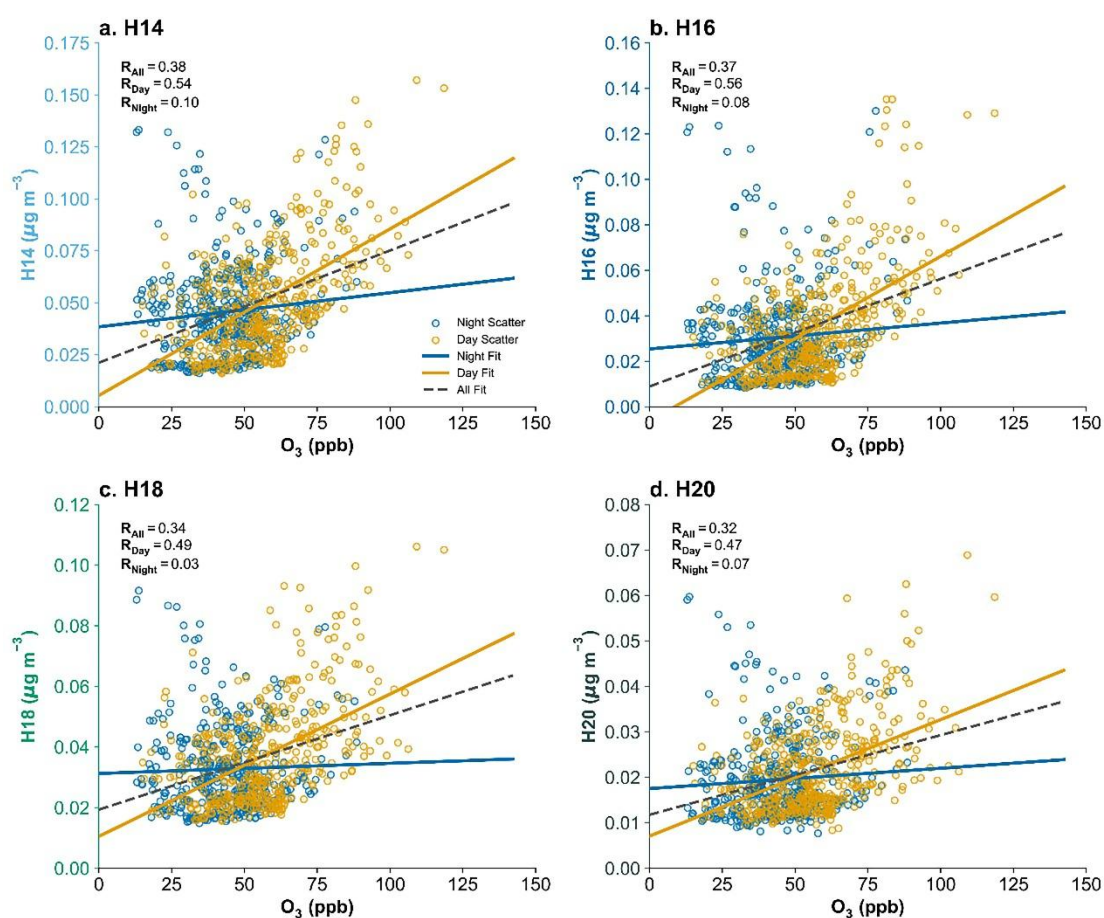


Figure S5. Day-night relationships between OOMs classes and O_3 . Data points are color-coded by day (yellow) and night (blue) during the campaign.

S2 Separate ML analysis for gas-phase and particle-phase OOMs

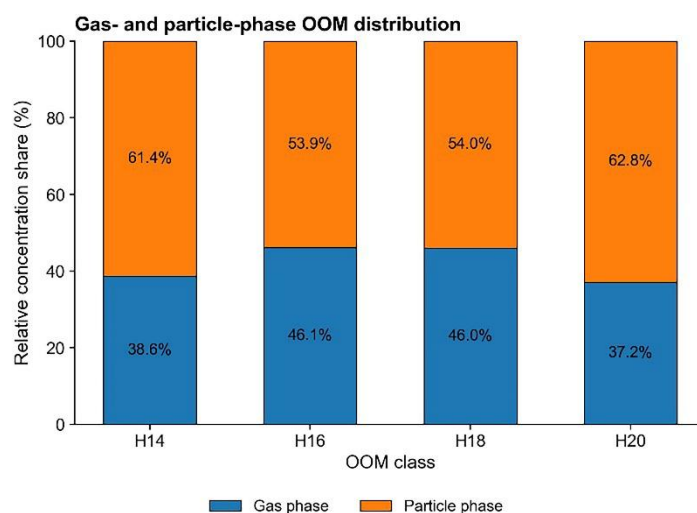


Figure S6. Gas- and particle-phase relative concentration distribution of the four OOM classes (H14, H16, H18, and H20) during the campaign.

The results show that the particle-phase fractions for H14, H16, H18, and H20 are 61.4%, 53.9%, 54.0%, and 62.8%, respectively, indicating that the particle-phase contribution is non-negligible (Shrivastava et al., 2017).

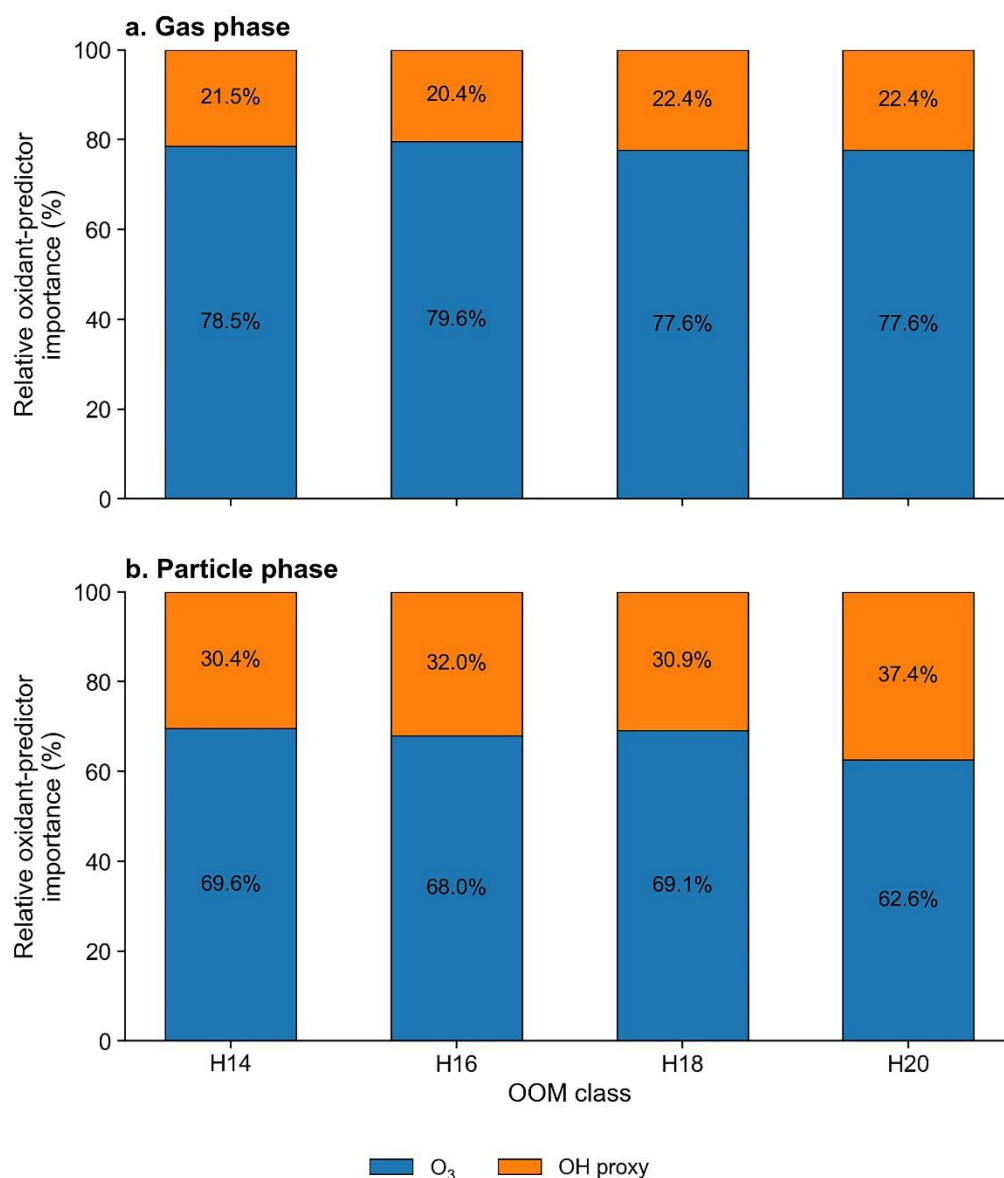


Figure S7. Relative predictive importance of O₃ vs. OH proxy for (a) gas-phase and (b) particle-phase OOMs, derived from separate ensemble ML models coupled with SHAP analysis.

The results show that the overall conclusions are consistent across both phases, but notable phase-dependent differences emerge. In the gas phase, the relative SHAP importance of O₃ is approximately 77.6–79.6%; in the particle phase, O₃ remains dominant but decreases to 62.6–69.6%. Correspondingly, the importance of the OH proxy increases from 20.4–22.4% in the gas phase to 30.4–37.4% in the particle phase. Thus, within the models, O₃ remains more important than the OH proxy in both phases.

S3 Theoretical upper bound of O₃ contribution to OOM formation: qualitative consistency check

A qualitative consistency check based on independent kinetic calculations is reported in Appendix B of the main text; the calculation itself is set out here in full. We reiterate that these values represent the statistical contributions of the input features to the model's predictive output, rather than direct chemical formation fractions. To bridge this gap, we have independently calculated the theoretical maximum contribution of ozonolysis using ambient kinetic branching ratios and pathway-specific molar yields. The qualitative agreement between the statistical findings and these independent kinetic calculations supports the chemical plausibility of our ML-derived drivers, rather than claiming causality solely from the SHAP values.

To explicitly bridge the gap between our statistical attributions and fundamental physical chemistry, we calculated the ambient pseudo-first-order kinetic branching ratio of α -pinene (as a representative of monoterpenes) oxidation ($Ratio_{O_3}$) using the high-resolution measurements of OH and O₃ during the campaign

$$Ratio_{O_3} = \frac{k_{O_3}[O_3]}{k_{O_3}[O_3] + k_{OH}[OH]}$$

where the bimolecular rate coefficients for α -pinene oxidation are $k_{O_3} \approx 8.7 \times 10^{-17} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$ and $k_{OH} \approx 5.3 \times 10^{-11} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$.

Based on our real-world observations, the average kinetic fraction of α -pinene oxidized by O₃ was 53.2%. While O₃ accounts for roughly half of the total precursor loss, its modeled explanatory weight for OOMs abundance is appreciably higher (~75%). This

upward shift from kinetic loss to product contribution is physically consistent with the aforementioned disparate molar OOMs yields. The theoretical relative contribution of O_3 to total OOMs formation ($Contr_{O_3}$) can be estimated by weighting the kinetic branching ratio with the respective pathway yields ($Y_{O_3} \approx 3.4\%$, $Y_{OH} \approx 0.44\%$):

$$Contr_{O_3} = \frac{Ratio_{O_3} \cdot Y_{O_3}}{Ratio_{O_3} \cdot Y_{O_3} + (1 - Ratio_{O_3}) \cdot Y_{OH}}$$

Substituting our observed $Ratio_{O_3}$ (53.2%) into this equation yields a theoretical O_3 contribution of approximately 89.7%. This theoretical value (89.7%) represents an upper bound, as it excludes other oxidants (e.g., NO_3 radicals) and the effects of multi-component mixtures, and assumes ideal conditions without radical sinks (Yang et al., 2025). Our SHAP-derived statistical importance (70.8%–79.4%) falls below this theoretical maximum, which is plausible given complex ambient loss processes (e.g., $RO_2 + NO$ termination) that reduce the effective contribution of the ozonolysis pathway under real-world conditions. This calculation is not intended to numerically validate the exact percentages from the machine learning model, given the vast differences in experimental timescales and instrumentation. Rather, the qualitative convergence between the statistical dominance of O_3 in the ML model and the intrinsic higher yield of ozonolysis suggests that our data-driven findings are consistent with established physical chemistry rather than spurious correlations. The two estimates therefore agree in direction and in order of magnitude, which is what a qualitative consistency check of this kind can establish; the main text uses the same comparison to place a bound on the SHAP-derived weights rather than to replace them.

S4 Comparison of measured OOM abundances across the two O₃ regimes

To test whether the model-derived statistical activation regime (~63 ppb) is visible in the measurements, we directly compared the measured total OOM concentrations during periods with O₃ below 62 ppb and periods with O₃ above 64 ppb, as shown in Figure S8.

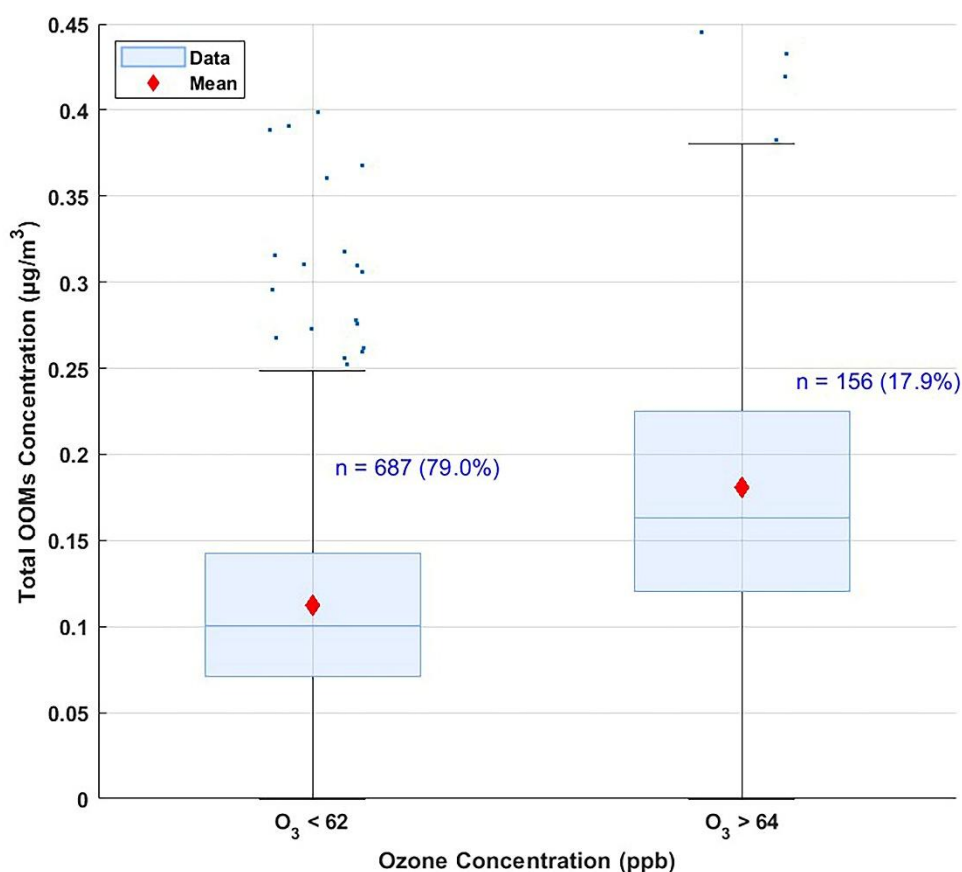


Figure S8. OOMs abundances across the two O₃ regimes

The box plot reveals a distinct separation in OOM abundances across the two O₃ regimes. During the low-O₃ period (< 62 ppb), which accounts for the majority of the observational samples (n = 687, 79.0%), the average total OOM concentration (indicated by the red diamond) is ~0.11 µg m⁻³. In contrast, during the high-O₃ period (> 64 ppb), despite the smaller sample size (n = 156, 17.9%), the average total OOM concentration increases substantially to ~0.18 µg m⁻³.

This comparison of measured abundances is consistent with the SHAP-derived

activation behaviour: the mean OOM abundance is higher when ambient O₃ exceeds the identified ~63 ppb transition point. Because it contrasts two regime means without a formal test, it does not by itself establish the transition as a chemical threshold. The wide range of the box plots (especially the upper whiskers and outliers) in both regimes also highlights the inherent chemical complexity of the ambient environment, where OOM formation is influenced not only by O₃, but by the synergistic effects of other factors such as NO_x and temperature (Pye et al., 2019).

References

- Atkinson, R. and Arey, J.: Atmospheric Degradation of Volatile Organic Compounds, *Chemical Reviews*, 103, 4605-4638, 10.1021/cr0206420, 2003.
- Ehn, M., Thornton, J. A., Kleist, E., Sipilä, M., Junninen, H., Pullinen, I., Springer, M., Rubach, F., Tillmann, R., Lee, B., Lopez-Hilfiker, F., Andres, S., Acir, I.-H., Rissanen, M., Jokinen, T., Schobesberger, S., Kangasluoma, J., Kontkanen, J., Nieminen, T., Kurtén, T., Nielsen, L. B., Jørgensen, S., Kjaergaard, H. G., Canagaratna, M., Maso, M. D., Berndt, T., Petäjä, T., Wahner, A., Kerminen, V.-M., Kulmala, M., Worsnop, D. R., Wildt, J., and Mentel, T. F.: A large source of low-volatility secondary organic aerosol, *Nature*, 506, 476-479, 10.1038/nature13032, 2014.
- Hallquist, M., Wenger, J. C., Baltensperger, U., Rudich, Y., Simpson, D., Claeys, M., Dommen, J., Donahue, N. M., George, C., Goldstein, A. H., Hamilton, J. F., Herrmann, H., Hoffmann, T., Iinuma, Y., Jang, M., Jenkin, M. E., Jimenez, J. L., Kiendler-Scharr, A., Maenhaut, W., McFiggans, G., Mentel, T. F., Monod, A., Prévôt, A. S. H., Seinfeld, J. H., Surratt, J. D., Szmigielski, R., and Wildt, J.: The formation, properties and impact of secondary organic aerosol: current and emerging issues, *Atmos. Chem. Phys.*, 9, 5155-5236, 10.5194/acp-9-5155-2009, 2009.
- Pye, H. O. T., D'Ambro, E. L., Lee, B. H., Schobesberger, S., Takeuchi, M., Zhao, Y., Lopez-Hilfiker, F., Liu, J., Shilling, J. E., Xing, J., Mathur, R., Middlebrook, A. M., Liao, J., Welti, A., Graus, M., Warneke, C., de Gouw, J. A., Holloway, J. S., Ryerson, T. B., Pollack, I. B., and Thornton, J. A.: Anthropogenic enhancements to production of highly oxygenated molecules from autoxidation, *Proceedings of the National Academy of Sciences*, 116, 6641-6646, doi:10.1073/pnas.1810774116, 2019.
- Shrivastava, M., Cappa, C. D., Fan, J., Goldstein, A. H., Guenther, A. B., Jimenez, J. L., Kuang, C., Laskin, A., Martin, S. T., Ng, N. L., Petaja, T., Pierce, J. R., Rasch, P. J., Roldin, P., Seinfeld, J. H., Shilling, J., Smith, J. N., Thornton, J. A., Volkamer, R., Wang, J., Worsnop, D. R., Zaveri, R. A., Zelenyuk, A., and Zhang, Q.: Recent advances in understanding secondary organic aerosol: Implications for global climate forcing, *Reviews of Geophysics*, 55, 509-559, <https://doi.org/10.1002/2016RG000540>, 2017.
- Yang, L., Nie, W., Yan, C., Ehn, M., Liu, Y., Roldin, P., Qi, X., Kerminen, V.-M., Donahue, N. M., Worsnop, D., Kulmala, M., and Ding, A.: A mechanistic understanding of the varying

yields of highly oxygenated organic molecules, *Nature Communications*, 17, 302, 10.1038/s41467-025-67007-w, 2025.