S1. Measurement results of gas species, VOCs, and meteorology.

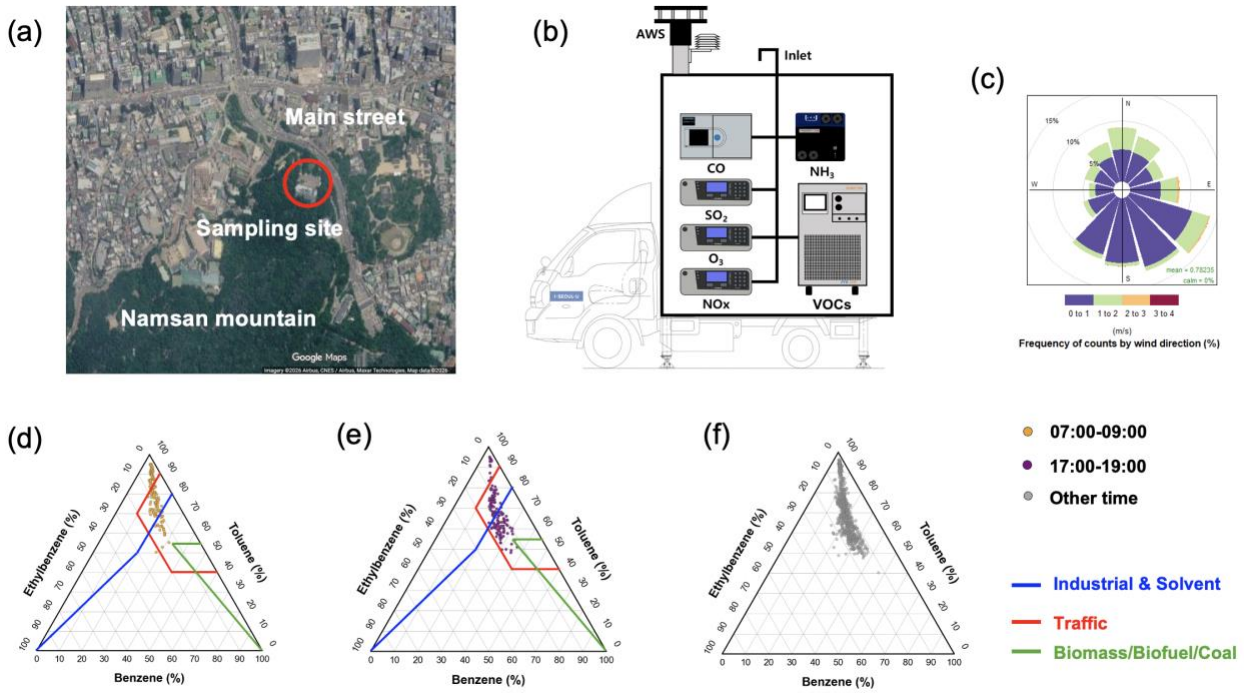

Figure S1: Overview of the sampling campaign setup. (a) Satellite map showing the location of the sampling site (red circle) in Seoul, situated between a main street and Namsan Mountain. (b) Schematic diagram of the mobile laboratory equipped with instruments for measuring trace gases (CO, SO_2, O_3, NO_x, NH_3) and VOCs, along with an automatic weather station (AWS). (c) Wind rose diagram depicting the frequency of wind speed and direction during the observation period. (d-f) Ternary diagram for morning rush hour, afternoon rush hour and other periods. Satellite imagery and map data in panel (a): Google Maps; Imagery ©2026 Airbus, CNES / Airbus, Maxar Technologies; Map data ©2026 Google.

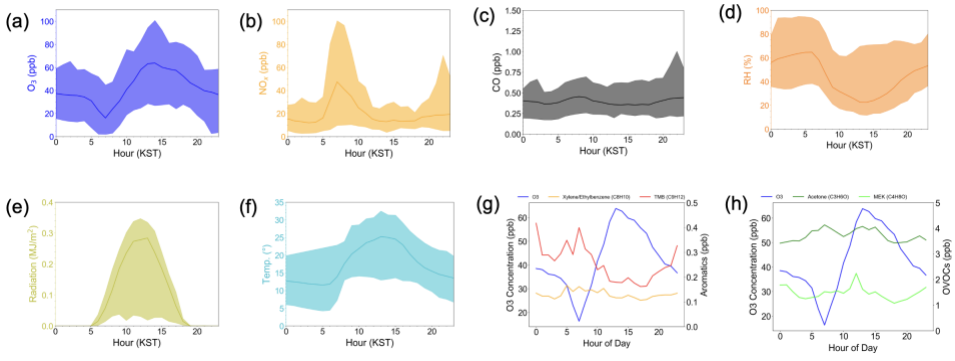

Figure S2: Diurnal variations of (a) ozone, (b) NO_x, (c) CO, (d) RH, (e) solar radiation, and (f) temperature. Panels (g) and (h) compare the diurnal profiles of ozone (left axis) with representative VOC species (right axis): (g) aromatics (Xylenes/Ethylbenzene and TMB) and (h) OVOCs (Acetone and MEK). Solid lines represent hourly averages, while shaded areas indicate the range (minimum to maximum values).

Table S1: Average concentrations of the top 20 VOCs measured by PTR-MS. Species were selected based on their overall average mixing ratios and are grouped by chemical category. Daytime (06:00-19:00) and nighttime (19:00-06:00) periods are distinguished based on the presence of solar radiation.

Formula (PTR-MS)	VOCs	Concentration (ppb)		
		Overall average	Daytime average	Nighttime average
OVOCs				
CH_2O	Formaldehyde	1.07 ± 0.53	1.07 ± 0.56	1.08 ± 0.50
CH_4O	Methanol	4.66 ± 2.11	4.42 ± 1.90	4.94 ± 2.29
CH_2O_2	Formic acid	0.91 ± 0.30	0.96 ± 0.32	0.85 ± 0.27
C_2H_4O	Acetaldehyde	1.84 ± 0.73	1.93 ± 0.75	1.75 ± 0.70
$C_2H_4O_2$	Acetic acid	2.37 ± 1.10	2.46 ± 1.10	2.28 ± 1.08
C_3H_4O	Acrolein	1.16 ± 0.96	1.17 ± 1.05	1.14 ± 0.86
C_3H_6O	Acetone	3.71 ± 1.44	3.85 ± 1.40	3.55 ± 1.47
$C_3H_6O_2$	Methyl acetate/Propionic acid	0.52 ± 0.29	0.54 ± 0.32	0.49 ± 0.26
C_4H_8O	Methyl Ethyl Ketone	1.46 ± 1.11	1.49 ± 1.07	1.43 ± 1.14
Aromatics				
C_6H_6	Benzene	0.67 ± 0.36	0.69 ± 0.34	0.66 ± 0.38
C_7H_8	Toluene	4.51 ± 6.02	4.38 ± 4.88	4.66 ± 7.08
C_8H_8	Styrene	0.13 ± 0.08	0.13 ± 0.07	0.13 ± 0.09
C_8H_{10}	Xylene / Ethylbenzene	1.27 ± 1.03	1.29 ± 0.87	1.26 ± 1.18
$C_{10}H_8$	Naphthalene	0.08 ± 0.03	0.08 ± 0.04	0.08 ± 0.03
C_9H_{12}	Trimethylbenzene	0.26 ± 0.25	0.24 ± 0.20	0.29 ± 0.30
Biogenic				
C_5H_8	Isoprene	0.19 ± 0.09	0.20 ± 0.09	0.18 ± 0.08
$C_{10}H_{16}$	Limonene / α-Pinene	0.08 ± 0.04	0.08 ± 0.04	0.07 ± 0.03
$C_{15}H_{24}$	Sesquiterpenes	1.51 ± 0.17	1.55 ± 0.21	1.46 ± 0.09

S2. Details of photochemical clock and MCM-OBM model.

Text S1. OFP and OH reactivity

To quantify the theoretical influence of individual VOCs on ozone production, the OFP and OHR were calculated. OFP is defined as the product of the species concentration and its MIR factor (Carter, 2009), providing a static estimate of the ultimate ozone-forming capacity under optimum conditions. Complementing this, OHR was calculated by multiplying the VOC concentration by its reaction rate constant with the hydroxyl radical (Atkinson and Arey, 2003). Unlike OFP, OHR represents the instantaneous kinetic driver of the oxidation cycles that fuel ozone production (Sinha et al., 2012).

Text S2. Observation-Based Model (OBM)

To quantify in situ ozone production dynamics, we simulated the chemical system using an Observation-Based Model (OBM) utilizing the Master Chemical Mechanism (MCM) via the AtChem2 framework (Sommariva et al., 2020). The model was constrained by hourly measurements of NO, NO_2, O_3, CO, meteorology, and a full suite of measured VOCs (Zhu et al., 2024). To address instrumental limitations in isomer separation, un-speciated C_8H_{10} masses were allocated to xylenes and ethylbenzene using representative ratios derived from the KORUS-AQ campaign in Seoul (Simpson et al., 2020). Physical loss processes (dilution and deposition) were parameterized using a first-order loss rate equivalent to a 12-hour mixing lifetime (1/43200 s^{-1}) (Zhu et al., 2024), and unmeasured HONO was constrained to 2% of NO_x (Wang et al., 2025).

The net ozone production rate ($P(O_3)$) was calculated as the difference between the gross production rate and the loss rate (Gu et al., 2022). To quantify the chemical sensitivity of the system, the Relative Incremental Reactivity (RIR) was calculated, representing the percentage change in ozone production resulting from a 1% change in a specific precursor or precursor group, serving as a direct metric of effective control potential (Wang et al., 2025).

Text S3. OH exposure estimation

To distinguish the extent of chemical processing from physical dilution, we estimated the photochemical age of the air mass. The time-integrated exposure to OH radicals (OH exposure or [OH]Δt) was quantified using the ratio of benzene to trimethylbenzene (TMB), as these compounds share common sources but differ significantly in their reactivity (Borbon et al., 2013; Hayes et al., 2013). The OH exposure was solved using the following equation:

$$[OH]\Delta t = \frac{1}{(k_{TMB+OH} - k_{benz+OH})} \times \left(ln\frac{[benzene]}{[TMB]}\bigg|_{t=0} - ln\frac{[benzene]}{[TMB]} \right) \quad (1)$$

where $[OH]\Delta t$ represents the time integrated OH exposure (molecules cm^{-3} s); k_{TMB+OH} and $k_{benz+OH}$ are the rate constants for the reaction of TMB and benzene with OH radicals, respectively; $[benzene]/[TMB]_{t=0}$ is the initial emission ratio determined at nighttime when photochemistry is negligible (derived from the regression slope in Fig. S3); and [benzene]/[TMB] is the measured concentration ratio at sampling time t.

The rate constant for TMB ($k = 36.0 \times 10^{-12}$ cm^3 s^{-1}) was calculated as the average of its isomers (1,2,3-; 1,2,4-; and 1,3,5-TMB), considering their measured ratio in KORUS-AQ (Simpson et al., 2020), as the PTR-ToF-MS technique used in this study could not resolve specific isomers.

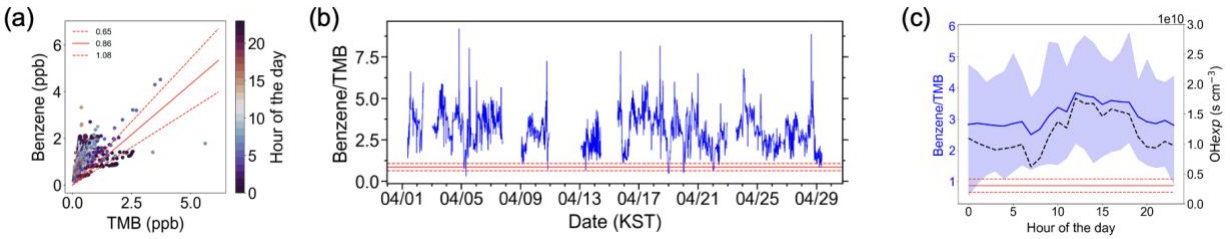

Figure S3: Photochemical clock analysis using [benzene]/[TMB]. Panel (a) displays scatter plots of benzene vs TMB, colored by hour of the day. Panel (b) shows the time series of estimated OH exposure (OH_{exp}). Panel (c) presents the diurnal variations of VOC ratios (left axis, blue) and corresponding OH_{exp} (right axis, black). Red solid lines indicate the estimated emission ratios, and dashed red lines represent their uncertainty bounds.

Text S4. Verification of Chemical Oxidation versus Physical Dilution

To verify that the ΔVOC term represents chemical oxidation rather than physical dilution, we analyzed the timing of VOC depletion relative to the photochemical age of the air mass (Fig. S4). Following the morning rush hour, observations show a sharp decline in the concentrations of both toluene and TMB. However, the process-informed ML model distinguishes the drivers of these losses. Toluene, despite its rapid concentration decay, is assigned a low SHAP slope (Fig. 3c), indicating that its depletion is primarily driven by physical dilution (boundary layer expansion) rather than ozone-producing oxidation. In contrast, the model assigns high sensitivity to TMB and xylenes. These species exhibit depletion rates that exceed physical dilution, coincident with the rise in calculated OH exposure. This selective attribution confirms that the model successfully filters out the 'background' signal of physical dispersion.

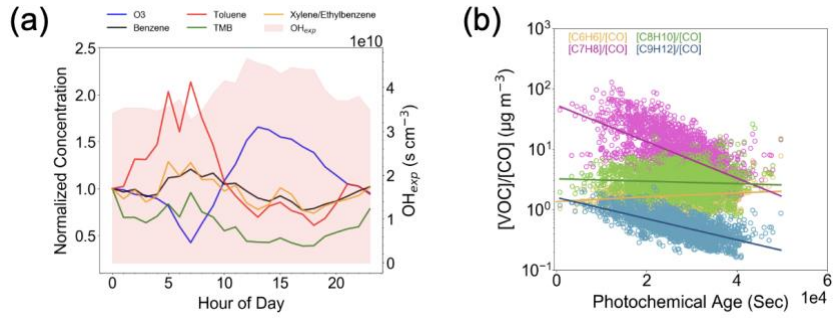

Figure S4: Relationship between VOC oxidation and photochemical age. (a) Diurnal profiles of normalized O_3, benzene, toluene, xylene/ethylbenzene, and trimethylbenzene (TMB), shown together with the estimated OH exposure (shaded, right axis). (b) VOC-to-CO ratios as a function of photochemical age, illustrating progressive depletion of reactive aromatics with increasing oxidative processing.

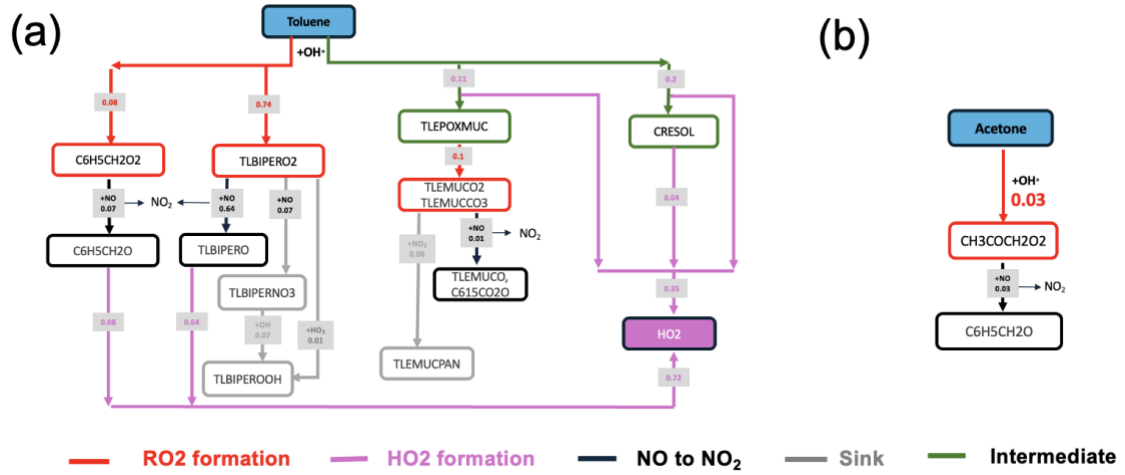

Figure S5: Radical reaction pathways simulated by the MCM-OBM for (a) toluene and (b) acetone oxidation. Arrows indicate formation of RO_2 (red) and HO_2 (purple), conversion of NO to NO_2 (black), intermediate formation (green), and termination/sink reactions (gray). Values denote the reaction rates (ppt/h) of each pathway, illustrating substantially stronger peroxy-radical production from toluene compared to acetone under the observed conditions.

S3. Validation, selection of input parameters, and results of the ML model.

Text S5. Evaluation of Feature Selection and Grouping Strategies

Selecting a chemically interpretable subset of input variables from high-dimensional PTR-MS data is critical for accurate ozone attribution. We found purely statistical feature selection approaches unsuitable for this task, as they tended to prioritize oxidation coproducts that correlate strongly with atmospheric aging rather than causal precursors (Guyon et al., 2002). To resolve this, we implemented a knowledge-guided feature curation process. First, the comprehensive VOC list was filtered against Maximum Incremental Reactivity (MIR) scale, retaining only recognized ozone precursors (Carter, 2009). This curation retained key primary aromatics, biogenics, and major oxygenated species including formaldehyde (HCHO). Following this curation, we tested whether reducing dimensionality through feature engineering could improve robustness. We aggregated the selected VOCs into six physicochemical groups based on their OH reactivity (kOH, separated by the median into high and low reactivity subsets) and source origin (primary/biogenic and secondary/mixed). However, comparative evaluation demonstrated that the model utilizing individual, curated VOCs significantly outperformed the aggregated-group models in predictive accuracy (R^2 value). This confirms that preserving specific chemical identity is essential for capturing the non-linear sensitivities of ozone formation. Therefore, the full list of chemically curated individual species was adopted to maximize both model performance and the granularity of the subsequent SHAP analysis.

Text S6. Details of the ML Training and Validation Process

The predictive performance of the knowledge-informed model was optimized through systematic hyperparameter tuning and robust cross-validation. Based on the XGBoost framework, the model configuration was finalized with 500 estimators, a learning rate of 0.01, and a maximum tree depth of 6.

To ensure the model's generalizability and minimize overfitting, a comprehensive 10-fold cross-validation (CV) strategy was implemented. The dataset was partitioned into ten subsets; in each iteration, nine subsets (90%) were used for training, while the remaining subset (10%) served as an independent validation set. This iterative process was repeated ten times, ensuring that the entire dataset was effectively utilized for both training and validation phases. The optimized model achieved a mean R^2 of 0.79 across the validation folds, indicating good predictive performance. Statistical performance metrics and scatter plots are detailed in Fig. S7.

Text S7. Description of SHAP explainer

SHAP has substantially improved researchers' ability to interrogate the internal structure of ML models, which are often treated as black boxes because their decision-making is not readily interpretable. SHAP assigns an importance value to each predictor for a given prediction, thereby quantifying individual contributions to the output and clarifying the model's reasoning. This enables insights that are difficult to obtain with traditional analytical approaches. The global and local explanations provided by SHAP support two main objectives: (1) quantifying the overall influence of environmental drivers on ozone formation, and (2) interpreting how specific predictor values shape predictions, including linear and nonlinear effects of key factors (e.g., RH, temperature, and temporal variability). Computations were performed using the TreeExplainer in the SHAP package (v0.42) developed by Lundberg et al (Lundberg et al., 2020). The SHAP output can be expressed in a linear additive form, as shown in Equation (1).

$$f(x) = g(x) = \Phi_0 + \sum_{i=1}^{M} \Phi_i \quad (2)$$

where $f(x)$ is the XGBoost predictive model; $g(x)$ is the output from the explanation model from the SHAP method; and Φ_0 is the expected or base value over the dataset, while Φ_i is the SHAP value of the *i*th factor, both of which are generated by the SHAP explanation model.

In the context of our models, a positive value ($\Phi_i > 0$) of variable i signifies a favorable association that elevated the prediction above the base value (Φ_0), whereas a negative value ($\Phi_i < 0$) implied a negative influence that reduced the prediction below the base value. In the context of atmospheric pollution, the variables were associated with their effect on the pollutant concentration of the chemical components.

Table S2: VOC categories and corresponding kOH values (in units of 10^{-12} cm^3 $molecule^{-1}$ s^{-1}).

Potential Source	High reactivity (>median)	Low reactivity (<median)
Primary/Biogenic	C_9H_{12} – TMB (kOH = 32.5-56.7) $C_{10}H_{16}$ – Monoterpene (kOH = 52.3-74.3) C_8H_8 – Styrene (kOH = 58.0) C_5H_8 – Isoprene (kOH = 100.0) C_8H_{10} – Ethylbenzene (kOH = 7.0)/xylene (kOH=13.6-23.1)	CH_4O – Methanol (kOH = 0.94) C_4H_8O – MEK (kOH = 1.22) C_7H_8 – Toluene (kOH = 5.63)
Secondary/Mixed	C_3H_4O – Acrolein (kOH = 29.0)	C_3H_6O – Acetone (kOH = 0.17) C_2H_4O – Acetaldehyde (kOH = 15.0) CH_2O – Formaldehyde (kOH = 9.37)

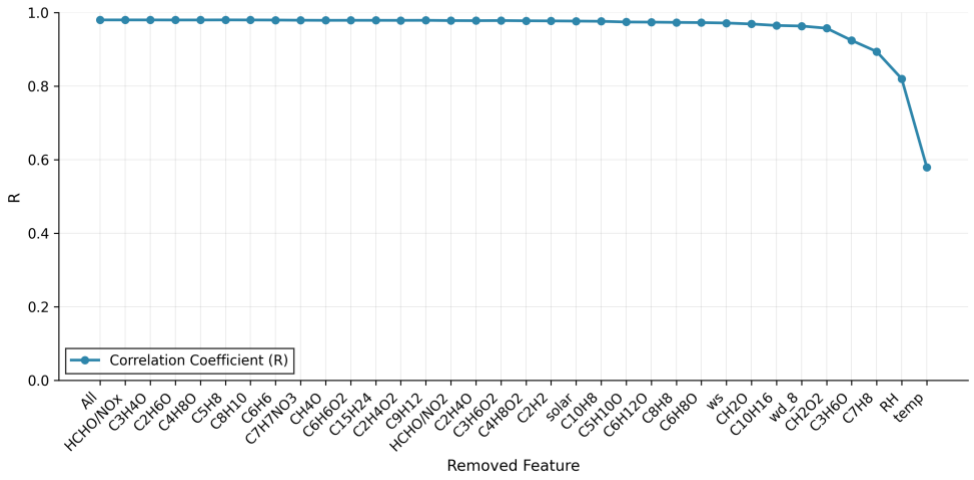

Figure S6: Recursive Feature Elimination (RFE) results showing the change in model performance (R) as input features are individually removed according to feature importance. The x-axis lists the features in the order of their removal (from least important on the left to most important on the right).

Table S3: Model input and output variables for the conventional ML model

Data source	Input parameters	Output
KMA	temperature (°C), WS (m/s), WD (°), RH (%)	O_3 (ppb)
PTR-MS	(VOCs unit in ppb) Aromatic: C_7H_8, C_8H_8, C_8H_{10}, C_9H_{12} Biogenic: C_5H_8, $C_{10}H_{16}$ OVOCs: CH_2O, CH_4O, C_2H_4O, C_3H_4O, C_3H_6O, C_4H_8O	
Auxiliary	HoD (h)	

Table S4: Model input and output variables for the process-informed ML model

Data source	Input parameters	Output
KMA	temperature (Δ°C/Δt), WS (m/s), WD (°), RH (Δ%/Δt)	O_3 (Δppb/Δt)
PTR-MS	(VOCs unit in Δppb/Δt) Aromatic: C_7H_8, C_8H_8, C_8H_{10}, C_9H_{12} Biogenic: C_5H_8, $C_{10}H_{16}$ OVOCs: CH_2O, CH_4O, C_2H_4O, C_3H_4O, C_3H_6O, C_4H_8O	
Auxiliary	HoD (h)	

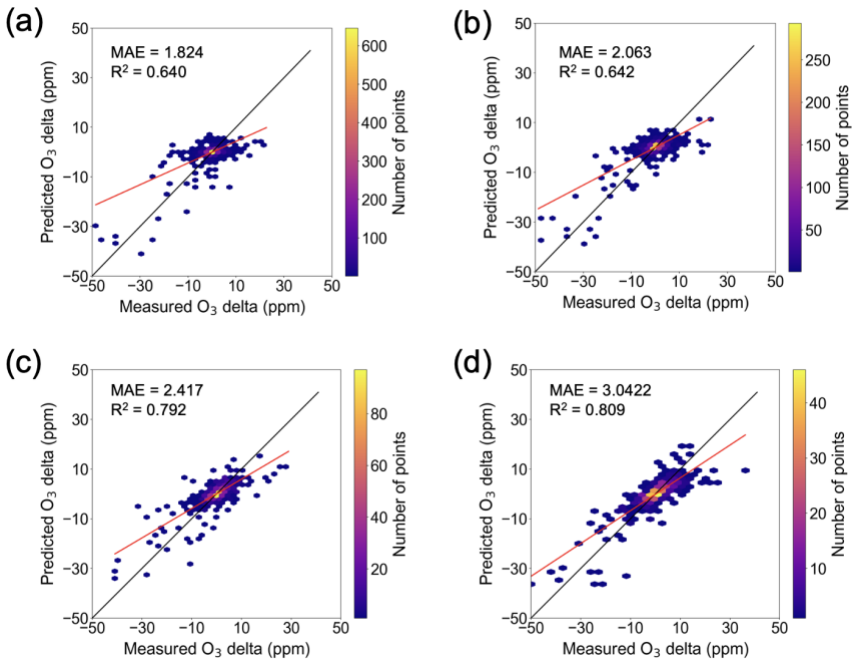

Figure S7: Cross-validation performance of the model on held-out validation folds under different temporal integration windows (Δt): (a) 3 min; (b) 9 min; (c) 18 min; (d) 30 min.

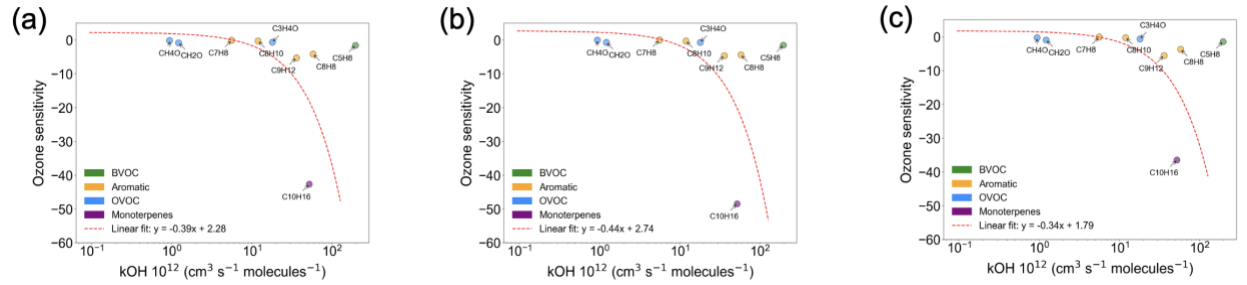

Figure S8: Linear regression analysis of SHAP-derived response slope versus OH reaction rate constant kOH for Aromatics and Monoterpenes. Red dashed curves represent the linear fit applied to the combined subset of Aromatics and Monoterpenes ($C_{10}H_{16}$). Panels (a–c) display the regression equations for overall, high ozone (ozone > mean) and low ozone periods (ozone < mean). Isoprene was excluded from the linear fit because its daytime emissions mask its net chemical consumption (ΔVOC), introducing variance that complicates the extraction of a simple linear response slope.

References

Atkinson, R. and Arey, J.: Atmospheric Degradation of Volatile Organic Compounds, Chem. Rev., 103, 4605–4638, https://doi.org/10.1021/cr0206420, 2003.

Borbon, A., Gilman, J. B., Kuster, W. C., Grand, N., Chevaillier, S., Colomb, A., Dolgorouky, C., Gros, V., Lopez, M., Sarda-Esteve, R., Holloway, J., Stutz, J., Petetin, H., McKeen, S., Beekmann, M., Warneke, C., Parrish, D. D., and de Gouw, J. A.: Emission ratios of anthropogenic volatile organic compounds in northern mid-latitude megacities: Observations versus emission inventories in Los Angeles and Paris, J. Geophys. Res. Atmospheres, 118, 2041–2057, https://doi.org/10.1002/jgrd.50059, 2013.

Carter, W. P. L.: Updated maximum incremental reactivity scale and hydrocarbon bin reactivities for regulatory applications, Calif. Air Resour. Board Contract, 339, 2009.

Gu, C., Wang, S., Zhu, J., Wu, S., Duan, Y., Gao, S., and Zhou, B.: Investigation on the urban ambient isoprene and its oxidation processes, Atmos. Environ., 270, 118870, https://doi.org/10.1016/j.atmosenv.2021.118870, 2022.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V.: Gene Selection for Cancer Classification using Support Vector Machines, Mach. Learn., 46, 389–422, https://doi.org/10.1023/A:1012487302797, 2002.

Hayes, P. L., Ortega, A. M., Cubison, M. J., Froyd, K. D., Zhao, Y., Cliff, S. S., Hu, W. W., Toohey, D. W., Flynn, J. H., Lefer, B. L., Grossberg, N., Alvarez, S., Rappenglück, B., Taylor, J. W., Allan, J. D., Holloway, J. S., Gilman, J. B., Kuster, W. C., de Gouw, J. A., Massoli, P., Zhang, X., Liu, J., Weber, R. J., Corrigan, A. L., Russell, L. M., Isaacman, G., Worton, D. R., Kreisberg, N. M., Goldstein, A. H., Thalman, R., Waxman, E. M., Volkamer, R., Lin, Y. H., Surratt, J. D., Kleindienst, T. E., Offenberg, J. H., Dusanter, S., Griffith, S., Stevens, P. S., Brioude, J., Angevine, W. M., and Jimenez, J. L.: Organic aerosol composition and sources in Pasadena, California, during the 2010 CalNex campaign, J. Geophys. Res. Atmospheres, 118, 9233–9257, https://doi.org/10.1002/jgrd.50530, 2013.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I.: From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell., 2, 56–67, https://doi.org/10.1038/s42256-019-0138-9, 2020.

Simpson, I. J., Blake, D. R., Blake, N. J., Meinardi, S., Barletta, B., Hughes, S. C., Fleming, L. T., Crawford, J. H., Diskin, G. S., Emmons, L. K., Fried, A., Guo, H., Peterson, D. A., Wisthaler, A., Woo, J.-H., Barré, J., Gaubert, B., Kim, J., Kim, M. J., Kim, Y., Knote, C., Mikoviny, T., Pusede, S. E., Schroeder, J. R., Wang, Y., Wennberg, P. O., and Zeng, L.: Characterization, sources and reactivity of volatile organic compounds (VOCs) in Seoul and surrounding regions during KORUS-AQ, Elem. Sci. Anthr., 8, 37, https://doi.org/10.1525/elementa.434, 2020.

Sinha, V., Williams, J., Diesch, J. M., Drewnick, F., Martinez, M., Harder, H., Regelin, E., Kubistin, D., Bozem, H., Hosaynali-Beygi, Z., Fischer, H., Andrés-Hernández, M. D., Kartal, D., Adame, J. A., and Lelieveld, J.: Constraints on instantaneous ozone production rates and regimes during DOMINO derived using in-situ OH reactivity measurements, Atmospheric Chem. Phys., 12, 7269–7283, https://doi.org/10.5194/acp-12-7269-2012, 2012.

Sommariva, R., Cox, S., Martin, C., Borońska, K., Young, J., Jimack, P. K., Pilling, M. J., Matthaios, V. N., Nelson, B. S., Newland, M. J., Panagi, M., Bloss, W. J., Monks, P. S., and Rickard, A. R.: AtChem (version 1), an open-source box model for the Master Chemical Mechanism, Geosci. Model Dev., 13, 169–183, https://doi.org/10.5194/gmd-13-169-2020, 2020.

Wang, M., Lu, Y., Ji, W., Chai, W., Wang, H., Ding, F., Zhao, X., Qin, M., Zhang, Y., Ge, X., and Shao, M.: Maximum incremental reactivity for volatile organic compounds in three city clusters of China: quantification, variability, and implications for ozone control, Atmos. Environ., 361, 121459, https://doi.org/10.1016/j.atmosenv.2025.121459, 2025.

Zhu, J., Wang, S., Gu, C., Jiang, Z., Zhang, S., Xue, R., Yan, Y., and Zhou, B.: Why did ozone concentrations remain high during Shanghai's static management? A statistical and radical-chemistry perspective, Atmospheric Chem. Phys., 24, 8383–8395, https://doi.org/10.5194/acp-24-8383-2024, 2024.

231