https://doi.org/10.5194/egusphere-2026-2647
Preprint. Discussion started: 18 June 2026

Reactivity Over Abundance: Unveiling the True Kinetic Drivers of Urban Ozone Using Process-Informed Machine Learning

Qihua Hu [1], Hwajin Kim [1, 2], Joost de Gouw [3], Sujin Kwon [1], Yoojin Park [4], Sojin Lee [5]

[1] Department of Environmental Health Sciences, Graduate School of Public Health, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

[2] Institute of Health and Environment, Seoul National University, 1 Gwanak, Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

[3] Department of Chemistry and Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, Colorado 80309, United States

[4] Hanwha Solutions Corporation, Jincheon plant, 27816, Republic of Korea

[5] The Seoul Institute, Seoul 06756, Republic of Korea

Correspondence to: Hwajin Kim (khj0116@snu.ac.kr)

Abstract Ground-level ozone remains a persistent challenge in East Asian urban centers, where concentrations continue rising despite significant reductions in precursor emissions. Designing effective mitigation is complicated by the non-linear relationship between precursor abundance and reactivity. Standard data-driven approaches often suffer from "survivor bias," systematically undervaluing highly reactive precursors that are rapidly depleted. We introduce a process-informed machine learning framework that uses net chemical consumption (ΔVOC) rather than ambient concentrations to resolve this attribution failure. Applied to measurements from a NO_x-saturated roadside site in Seoul, South Korea, the approach reveals a robust quantitative relationship between intrinsic hydroxyl radical reactivity (kOH) and ozone formation sensitivity.

Across a comprehensive suite of precursors—including oxygenated VOCs (e.g., formaldehyde, acetone) and aromatics—the framework identifies reactive aromatics (trimethylbenzenes, xylenes) and biogenics (isoprene, monoterpenes) as the dominant kinetic drivers. Whereas static metrics such as OFP rank precursors by stoichiometric capacity under idealized accumulated conditions, the process-informed framework shows that realized ozone production in fresh urban plumes is governed by kinetic turnover rather than abundance, with this kinetic selectivity further amplified during high-ozone episodes. These results indicate that mass-based VOC inventories and OFP-style rankings, when applied without kinetic context, can systematically misallocate control priorities in NO_x-saturated urban regimes. Because the framework requires only sub-hourly co-located VOC and ozone observations and no prescribed mechanism, it offers a complementary empirical pathway in settings where explicit mechanism-based modeling is constrained by incomplete VOC speciation or unmeasured radical precursors.

1 Introduction

Ground-level ozone (O_3) remains a persistent environmental and health challenge in East Asian urban centers, continuing to climb despite significant reductions in precursor emissions (Li et al., 2020; Wang et al., 2019). For instance, between 2015 and 2021, South Korea reduced national emissions of nitrogen oxides (NO_x) by 26% and VOCs by approximately 25%, yet urban O_3 levels increased over the same period (Colombi et al., 2023). This divergence is characteristic of NO_x-saturated urban regimes, where non-linear chemistry renders strategies centered on bulk emission mass reductions ineffective (Chang et al., 2020). Addressing this trend requires shifting regulatory focus from aggregate reductions to prioritizing the control of specific, high-impact precursor classes.

However, the extreme chemical diversity of atmospheric precursors complicates this targeted approach. Ozone formation efficiency varies dramatically between VOC classes—driven by the high reactivity of alkenes and aromatics compared to less reactive alkanes—and is dependent on the specific VOC mixture in each environment (Atkinson and Arey, 2003). Consequently, the central challenge for modern air quality management is accurately identifying which specific VOCs drive local ozone production to prioritize control efforts effectively (Shi et al., 2024).

Current analytical frameworks struggle to bridge atmospheric complexity with policy applicability. Reactivity scales like Maximum Incremental Reactivity (MIR) are useful for ranking but rely on static, idealized conditions that are not inherently designed to capture the dynamics of urban atmospheres (Carter, 2009; Huang et al., 2025). Conversely, explicit Observation-Based Models (OBMs) utilizing explicit mechanisms like the Master Chemical Mechanism (MCM) provide detailed oxidation pathways; however, they are highly sensitive to unmeasured radical precursors and intermediate reaction rates (e.g., RO_2 radical propagation) (Berndt et al., 2018; Huang et al., 2025). Recently, machine learning (ML) has been applied to model the nonlinear ozone dynamics with high predictive accuracy, but its utility for chemical interpretation remains controversial (Zhang et al., 2024). An inherent limitation in conventional ML approaches trained on atmospheric data is their reliance on ambient concentrations as primary features. In atmospheric chemistry, a critical "survivor bias" exists: the most reactive—and thus most important—precursors are rapidly depleted, leaving low ambient concentrations that models statistically undervalue (Huang et al., 2025; Wang et al., 2022). This leads to skewed interpretations where inert accumulators are misidentified as primary drivers while reactive species are ignored, potentially misguiding control policies (Huang et al., 2025).

In this study, we introduce, to our knowledge for the first time, a process-informed machine learning framework that uses net VOC consumption (ΔVOC) rather than ambient concentrations as the predictive feature. By embedding the physical constraints of mass conservation and reaction kinetics directly into the learning process, the framework forces the model to relate the extent of VOC processing—rather than standing abundance—to ozone production. As a concrete illustration of the survivor-bias problem, we show that conventional concentration-based ML applied to our Seoul roadside dataset assigns acetone a higher ozone-driver rank than toluene, despite an OBM-MCM radical budget analysis indicating that toluene oxidation contributes an order of magnitude more peroxy radical flux. The ΔVOC framework, in contrast, autonomously recovers from observations alone the kinetic hierarchy expected from fundamental rate theory, including the mechanistic penalties of radical termination (organic nitrate formation, suppressed OH recycling) and the boost from non-photolytic radical generation by monoterpene ozonolysis. Beyond

https://doi.org/10.5194/egusphere-2026-2647
Preprint. Discussion started: 18 June 2026

this methodological validation, we report that during high-ozone episodes the kinetic selectivity of the system intensifies rather than the chemical regime shifting—providing a process-level explanation for why peak-ozone reduction strategies based on bulk emission inventories may underperform in fresh-emission urban regimes. Because the framework requires only sub-hourly co-located VOC and ozone observations and no prescribed mechanism, it offers a complementary empirical pathway in environments where explicit mechanism-based modeling is constrained by incomplete VOC speciation, unmeasured radical precursors (e.g., HONO, $ClNO_2$), or computational cost.

2 Methods

2.1 Measurement information

The sampling campaign was conducted in Seoul, South Korea, at the Seoul Youth Hostel (37.56°N, 126.99°E). Located in a dense urban center adjacent to a major roadside (Fig. S1), this site was selected to capture the complex mixture of VOCs typical of dense metropolitan environments, ranging from vehicular to commercial sources. Sampling occurred from April 1–30, 2022, a period selected to minimize the interference of biomass burning and maximize the representativeness of urban anthropogenic sources.

VOC concentrations were quantified using a proton transfer reaction time-of-flight mass spectrometer (PTR-ToF-MS 1000, Ionicon Analytik, Austria), enabling real-time detection of a broad spectrum of oxygenated VOCs (OVOCs), aromatics, and other high-proton-affinity compounds. Inorganic trace gases (NO_x, SO_2, O_3, CO) were monitored using standard gas analyzers (Models 42i, 43i, 49i, and 48i, Thermo Fisher Scientific, USA). Meteorological parameters (temperature, relative humidity, wind speed, and direction) were measured by a collocated automatic weather station (AWS).

2.2 Machine Learning Model

2.2.1. Curation of input VOC species

To ensure chemical interpretability, input variables were selected from the high-dimensional PTR-MS data through a knowledge-guided curation process. The input feature set comprised recognized ozone precursors based on the MIR scale, including key oxygenated VOCs (OVOCs) such as formaldehyde (HCHO) and acetone, alongside primary aromatics and biogenic species (Carter, 2009). We utilized these individual, curated species as inputs to maximize granular insight. Alternative selection strategies, including statistical recursive feature elimination (RFE) (Guyon et al., 2002) and physicochemical grouping based on reactivity or source, were evaluated and subsequently rejected due to issues with chemical circularity and reduced predictive accuracy, respectively (details presented in Text S5). Finally, it must be noted that certain key precursors, particularly light alkanes (e.g., ethane, propane), were not accurately quantified by the specific PTR-MS configuration used and are absent from model inputs.

2.2.2 ML Model Configuration and Temporal Integration

We employed the XGBoost gradient boosting framework, selected for its handling of non-linear interactions and compatibility with Shapley Additive Explanations (SHAP) (Chen and Guestrin, 2016; Lundberg et al., 2020). Two

ML model configurations were developed: a conventional model using ambient concentrations, and a process-informed model using the net change in VOC concentration (ΔVOC). Input features included NO_x (representing the titration sink), meteorological variables, and temporal markers to capture diurnal emission cycles (Tables S3 and S4). To prevent overfitting and ensure robust model performance, we employed a 10-fold cross-validation approach. The dataset was partitioned into ten distinct subsets, with each subset serving iteratively as an independent validation set while the remaining data were used for training (detailed in Text S6).

To capture the relevant chemical timescale, data were resampled into time steps (Δt) of 18 minutes. The process-informed variables (ΔX) were calculated as:

$$\Delta \mathrm{X} = \frac{X_t - X_{t-1}}{\Delta \mathrm{t}} \tag{1}$$

where X represents the input (VOC, NO_x, and meteorological factors) or output (ozone) species. The 18-minute integration window was rigorously determined via optimization tests (Fig. S7) to balance signal fidelity against physical transport. This interval is sufficiently long to average out high-frequency turbulent noise (dominant at < 5 min) yet short enough to precede significant shifts in planetary boundary layer dynamics (dominant at > 30 min, effectively isolating the local chemical signal) (Mauder et al., 2013; Metzger, 2018).

Time-shift analysis (introducing a lag between model inputs and the output target)—testing for potential delays between VOC oxidation and ozone response—confirmed optimal model performance at zero lag. This immediate response is consistent with the rapid kinetics of a NO_x-saturated roadside regime, where titration and alkene oxidation drive ozone variability on timescales shorter than the integration window (Clapp and Jenkin, 2001). While ΔVOC mathematically integrates emissions, transport, and chemistry, the study site offers a simplifying constraint: the strong regularity of traffic emissions is captured by the 'hour of day' variable. Consequently, the remaining variance in ΔVOC serves as a robust empirical proxy for the net chemical turnover driving local ozone production. It is important to distinguish this from metrics like OHR and OFP, which are computed from current concentrations at the sampling site and therefore represent an instantaneous propensity (a box model snapshot) rather than a realized response that depends on prior airmass evolution. By design, ΔVOC summarizes the net change over Δt, so it is sensitive to the recent history of the sampled air mass within the selected window.

Effectively, the evaluation of the Δ derivative over this optimized 18-minute window acts as a dynamic kinetic high-pass filter. It mathematically strips away the inert, accumulated background of long-lived species (where ΔVOC =~ 0) and exclusively captures the instantaneous oxidative turnover of fresh local emissions.

2.3 Chemical Indicators and Observation-Based Validation

To benchmark the ML results against established chemical frameworks, we employed three complementary metrics. First, theoretical ozone formation contributions were assessed using Ozone Formation Potential (OFP) based on MIR factors (Carter, 2009) and OH reactivity (OHR), which quantifies the instantaneous kinetic driving force (Text S1) (Sinha et al., 2012). Second, in situ ozone production dynamics were simulated using an Observation-Based Model (OBM) driven by the Master Chemical Mechanism (MCM v3.3.1) via the AtChem2 framework (Text S2) (Sommariva

https://doi.org/10.5194/egusphere-2026-2647
Preprint. Discussion started: 18 June 2026

et al., 2020). The model yielded the Relative Incremental Reactivity (RIR) for each precursor, providing a direct metric of effective control potential (Wang et al., 2025). Finally, to distinguish chemical processing from physical dilution, photochemical age (time-integrated OH exposure) was calculated using a hydrocarbon clock based on the ratio of benzene to trimethylbenzenes (TMB) (Text S3) (Borbon et al., 2013). Detailed model constraints, isomer allocations derived from KORUS-AQ data, and the specific rate constants utilized for photochemical aging are provided in the Supporting Information.

3 Results and Discussion

3.1 Overview of measurements

The sampling site exhibited a chemically active urban regime driven by the coexistence of fresh anthropogenic emissions and regional background oxidation. While the bulk VOC composition was numerically dominated by oxygenated species (OVOCs) such as methanol and acetone (Table S1), the substantial abundance of primary aromatics—specifically toluene (4.51 ± 6.02 ppb) and C_8 aromatics (1.40 ± 1.10 ppb)—confirms the immediate influence of local, continuous emissions rather than aged air masses. This precursor abundance fueled intense in situ photochemistry: ground-level O_3 exhibited a sharp diurnal amplitude, rising from a morning minimum of 16.2 ppb to an afternoon peak of 63.8 ppb (Fig. S2). Crucially, 83% of hourly measurements exceeded the WHO peak-season guideline of 31 ppb (World Health Organization, 2021), highlighting the persistence of hazardous ozone levels in this NO_x-saturated environment.

To verify the dominance of local primary sources, we employed a ternary diagram using benzene, toluene, and ethylbenzene (Fig. 1) (Zhao et al., 2024). Because PTR-ToF-MS cannot chemically resolve C_8H_{10} isomers, ethylbenzene was differentiated from xylenes using a fixed ratio (0.366) derived from speciated measurements during the local KORUS-AQ campaign (Simpson et al., 2020). The resulting distribution clusters predominantly within the traffic-specific wireframe, establishing vehicular exhaust as the persistent baseline source. However, a noticeable skew toward the solvent/industrial sector indicates distinct episodic enhancements driven by toluene-rich commercial activities.

This dual-source influence is further clarified by the diurnal stability of the toluene-to-benzene (T/B) ratio (Fig. 1b). A distinct temporal divergence emerges between the two rush-hour periods. The morning rush (07:00–09:00) exhibits a significantly elevated toluene-to-benzene (T/B) ratio. This enhancement reflects the combined impact of cold-start traffic emissions (dominated by unburned fuel components rich in toluene due to inefficient catalytic conversion at low temperatures) and the onset of daytime solvent usage (Zhang et al., 2022; Simpson et al., 2020). In contrast, the afternoon rush (17:00–19:00) returns to a suppressed ratio indistinguishable from background levels. This suggests that while traffic provides the fundamental source signature, high-intensity solvent emissions are temporally confined to the morning operation, creating specific periods of heightened reactivity.

3.2 Main drivers of ozone formation

This section identifies the principal drivers of ozone formation by comparing the interpretability of a conventional machine learning model against a process-informed framework. We systematically evaluated the sensitivity of ozone production to inorganic precursors, meteorological parameters, and specific VOC classes. To verify the physical consistency of the machine learning attribution, model outputs were cross-validated using observation-based chemical mechanisms (MCM-OBM) and kinetic reactivity metrics (OFP, OHR). The following analysis aims to distinguish between the environmental conditions that facilitate ozone accumulation and the specific chemical precursors that drive rapid local production.

3.2.1. Resolving instantaneous drivers with the process-informed framework

By resolving the temporal dynamics of the system, the process-informed (Δ) framework decouples the accumulated atmospheric state from its active drivers, revealing a regime governed by rapid titration and chemical limitation.

The analysis identifies active titration as the primary high-frequency control governing local ozone variability within the 18-minute integration window. While the conventional model (Fig. 2a) broadly characterizes the environment as a static NO_x-saturated state, the process-informed model (Fig. 2a; blue scatter points) captures the instantaneous kinetic reality: fresh NO emissions are associated with a clear negative response in ozone production ($\Delta NO_x > 0 \rightarrow \Delta O_3 < 0$). This sharp negative response is not merely a statistical correlation but a mechanistic signature of the rapid titration reaction ($NO + O_3 \rightarrow NO_2 + O_2$). Since this reaction operates on timescales of roughly one minute in urban plumes (Liu et al., 2008; Sillman, 1999; Zhang et al., 2022), its full effect is completely captured within the 18-minute integration window. By isolating this rapid suppression, the framework confirms that titration functions as a dynamic modulator of ozone on sub-hourly scales rather than a static background condition, distinct from the slower evolution of the underlying oxidation regime.

The dynamic model further corrects a critical misinterpretation regarding the stability of the oxidation regime. The conventional model (Fig. 2e; red scatter points) displays a precipitous "titration cliff" across the range of $HCHO/NO_x$ ratios. While this plot accurately captures the diurnal association—where low morning ratios correlate with low ozone and high afternoon ratios with high ozone—the static framework conflates this long-term evolution with instantaneous sensitivity. This pattern could be misread as an instantaneous hypersensitivity, as if a small, immediate perturbation in the ratio would trigger a disproportionate surge in ozone production. The process-informed model (Fig. 2e) resolves this ambiguity by analyzing the rate of change (Δ). By showing negligible dependence on short-term ratio fluctuations ($\Delta Ratio \rightarrow \Delta O_3$), it reveals that the oxidation regime functions as a stable thermodynamic boundary condition. The $HCHO/NO_x$ ratio evolves gradually over diurnal timescales—driven by changing solar angles and cumulative emissions—rather than shifting rapidly enough to drive high-frequency ozone variance within the 18-minute window. Consequently, the delta model confirms that the "cliff" observed in static analysis is a statistical artifact of time-aggregated data, whereas the dynamic framework recovers the physical reality of a buffered, slowly evolving radical cycle (Pusede and Cohen, 2012).

Finally, the framework clarifies that peak ozone production is chemically limited, rather than meteorologically driven. Conventional models often conflate high ozone with its environmental prerequisites (high temperature, low humidity). However, the process-informed model (Fig. 2; blue scatter points) shows negligible sensitivity to short-term fluctuations in these parameters. This decoupling confirms that once the meteorological "stage" is set, the rate of ozone production is controlled strictly by the supply and turnover of reactive precursors, aligning with box-model studies demonstrating that in stagnation events, chemical throughput, not micrometeorological variance, dictates the production rate (Tan et al., 2019).

3.2.2 Uncovering VOC Drivers via Net Chemical Consumption

The stability of the underlying oxidation regime (established in Section 3.2.1) implies that high-frequency ozone fluctuations are driven by the specific reactive turnover of precursors processed within that regime. However, correctly identifying these precursors requires filtering out the "survivor bias" that obscures the true chemical drivers in concentration-based analyses.

Conventional ML analysis yields a chemically counterintuitive attribution: abundant, low-reactivity OVOCs (e.g., acetone) are identified as primary drivers, while highly reactive aromatics (toluene, TMB) are assigned negative or negligible ozone formation potentials (Fig. 3; red scatter points). The OBM-MCM radical budget analysis (Fig. S5) refutes this, demonstrating that toluene oxidation is a net source of peroxy radicals (RO_2), contributing an order of magnitude more to the propagation cycle than acetone. Consequently, the conventional model's attribution is confirmed as a statistical artifact arising from diurnal anti-correlation.

The process-informed framework resolves this causality gap by utilizing net chemical consumption (ΔVOC) as the predictor. The resulting SHAP dependence plots (Fig. 3; blue scatter points) reveal a chemically coherent mechanism: ozone production is strictly coupled to the reactive loss of primary precursors.

Specifically, the model recovers a strong inverse relationship for reactive aromatics ($\Delta\text{VOC} < 0 \rightarrow$ High SHAP Value), indicating that the rate of ozone formation accelerates in direct proportion to the consumption rate of toluene, xylenes, and TMB. Importantly, this relationship is absent for accumulated low reactivity OVOCs. This distinction provides a critical atmospheric insight: despite the numerical dominance of oxygenated species in the ambient mix, the instantaneous production of ozone in this NO_x-saturated environment is kinetically limited by the supply of fresh, high-reactivity aromatics. In this sense, realized ozone production at the site appears tied more to the fast turnover of reactive aromatics than to the standing abundance of oxygenated species.

3.3 Kinetic Control of Ozone Production and Radical Termination Pathways

The application of the process-informed machine learning framework revealed that local ozone production in the Seoul Metropolitan Area is governed by a strict kinetic hierarchy, where the intrinsic reactivity of a precursor outweighs its ambient abundance. We quantified this relationship by examining the net consumption of each VOC and its corresponding SHAP response slope. This was achieved by calculating a "SHAP slope" for each species, defined as the slope of the linear regression between the input ΔVOC and its corresponding SHAP value (contribution to ΔO_3).

In this framework, a steeper negative slope indicates that the net depletion of a precursor (negative ΔVOC) is associated with a higher contribution to ozone production (positive ΔO_3).

3.3.1 The Primacy of Kinetic Reactivity (kOH)

As shown in Fig. 4a, there is a distinct non-linear relationship between this data-driven SHAP response slope and the hydroxyl radical reaction rate constant (kOH). VOCs with low kinetic reactivity, such as toluene (kOH=5.7×10^{-12} cm^3 $molecule^{-1}$ s^{-1}) and benzene (kOH=1.2×10^{-12} cm^3 $molecule^{-1}$ s^{-1}), exhibit SHAP slopes near zero. Despite their high dominance in the ambient mixing ratio (Table S1), their slow oxidation rates prevent them from driving the rapid radical cycling required to sustain elevated ozone levels in this NO_x-saturated regime.

Conversely, the sensitivity curve steepens dramatically for highly reactive compounds, including isoprene (kOH=100×10^{-12} cm^3 $molecule^{-1}$ s^{-1}) and trimethylbenzenes (kOH=36.0×10^{-12} cm^3 $molecule^{-1}$ s^{-1}; average of three isomers). This trend confirms that the rate of oxidative turnover is the primary predictor of ozone variability.

Isoprene presents a unique challenge in this analysis. Typically, its strong daytime emissions, originating from both adjacent urban vegetation (e.g., Namsan Mountain) and vehicular traffic, mask its chemical consumption, often resulting in positive ΔVOC values even during active oxidation. However, the model assigns isoprene a high sensitivity index. This suggests that the model successfully detects deviations from the expected accumulation rate—identifying moments where rapid chemical loss competes with strong emissions. This high sensitivity aligns with the premise that in a NO_x-saturated regime, the system is strictly limited by the supply of peroxy radicals, which high-kOH species like isoprene generate most efficiently. These results extend established chemical frameworks by demonstrating that rapid depletion dynamics are the strongest statistical predictor of ozone formation potential (de Gouw et al., 2017; Pusede and Cohen, 2012). Crucially, this result demonstrates a major methodological breakthrough: without being explicitly programmed with physical or chemical mechanisms, the ΔVOC-informed ML framework autonomously recovered the fundamental kinetic rate laws of atmospheric chemistry purely from observational data.

Finally, to verify that the ΔVOC term represents chemical oxidation rather than physical dilution, we performed an independent observational verification detailed in Text S4. By analyzing VOC depletion relative to photochemical age (Fig. S4), we confirmed that dilution-dominated species (e.g., toluene) follow CO decay profiles, while oxidation-driven drivers (e.g., TMB) exhibit rapid depletion relative to CO. This confirmed that the derived sensitivities isolated the chemical consumption specifically responsible for local ozone production.

3.3.2 Mechanistic Penalties: Radical Termination and Outliers

While the process-informed model broadly aligns SHAP response slope with kOH, specific outliers reveal the model's capacity to distinguish between simple consumption speed and the ultimate efficiency of radical recycling (Fig. 4a).

Specifically, heavy aromatics like styrene and TMB exhibit lower sensitivities than their reactivity would suggest (falling below the trendline). For TMB, this suppressed sensitivity is consistent with the competition between chain propagation and termination via organic nitrate formation. Upon oxidation, large alkyl-substituted aromatics exhibit significantly higher organic nitrate yields ($RONO_2$) compared to lighter species (Perring et al., 2013). This branching pathway effectively sequesters both NO_x and radicals, reducing the net ozone yield per oxidized TMB molecule relative to less-substituted aromatics. The model correctly detects that a substantial fraction of the oxidized TMB mass is diverted into these termination reservoirs rather than fueling the catalytic cycle. Similarly, styrene shows a suppressed sensitivity consistent with inefficient radical regeneration. Unlike light alkenes, the ozonolysis of styrene is characterized by a negligible yield of recycled OH radicals (Tuazon et al., 1993). Consequently, styrene acts partially as a radical sink, consuming oxidants without regenerating the oxidative capacity required to sustain the photochemical chain.

In contrast, monoterpenes (measured as $\Sigma C_{10}H_{16}$) exhibit higher-than-expected sensitivities, placing them above the general reactivity trend. Theoretically, monoterpenes also possess high organic nitrate yields, which should penalize their production. However, our results suggest this penalty is outweighed by efficient non-photolytic radical generation. The ozonolysis of monoterpenes—particularly α-pinene and limonene—is a potent source of non-photolytic OH radicals with a yield of approximately 0.85 (Chew and Atkinson, 1996). This strong radical boosting effect could compensate for the loss of NO_x to nitrates, allowing these species to sustain rapid oxidation chains even under NO_x-saturated conditions.

It is important to acknowledge that the PTR-MS technique measures the sum of monoterpene isomers, preventing the distinct separation of species with varying yields (e.g., α-pinene vs. β-pinene). Since the efficiency of OH recycling varies among isomers, the elevated sensitivity observed here likely reflects a composite signal dominated by high-yield species in the local biogenic mix. Nevertheless, the model's differentiation of these species from aromatics underscores its capacity to capture the nuance of radical stoichiometry beyond simple reaction kinetics.

3.3.3 Drivers of Peak Pollution Episodes

The practical implication of this kinetic control is clearest when diagnosing high-ozone episodes. Fig. 5 compares the model sensitivity during baseline conditions versus high-ozone events. A critical evolution is observed in the NO_x response: while the negative sensitivity (titration) persists during high-ozone events, the magnitude attenuates (slope shifts from -0.45 to -0.28). This indicates that while the system remains strictly within the NO_x-saturated regime (where increasing NO_x suppresses ozone), the efficiency of this suppression diminishes as the air mass evolves toward peak concentrations.

Crucially, as the titration effect attenuates, the system exhibits an increase in response to reactive radical precursors. During high-ozone episodes, the reactivity-response slope significantly steepens (linear fit coefficient shifts from -0.34 to -0.44; Fig. S8). This steepening indicates that the ozone influence for high-kOH species is amplified: the system becomes disproportionately sensitive to monoterpenes and reactive aromatics, while the response to lower-kOH species like toluene remains negligible. This confirms that the transition to hazardous pollution levels is not driven by a fundamental shift in the chemical regime (e.g., from VOC-limited to NO_x-limited), but by the intensified processing of specific rate-limiting precursors.

These results show the distinct roles of different chemical metrics in diagnosing local pollution episodes. Conventional metrics like OHR and OFP (Fig. 4b) measure the current ozone production potential at sampling time, which effectively represents the theoretical yield if the air mass remained isolated in a static 0-D box model. In the real atmosphere, ozone depends on air mass history, and the ΔVOC metric captures some of this dynamic evolution by accounting for rapid chemical depletion. While OFP identifies the ultimate stoichiometric capacity over longer scales, the process-informed framework reveals that in this fresh-emission environment, the rapid reaction kinetics (kOH) dictate the realized ozone formation.

In this NO_x-saturated regime, ozone production is rate-limited by the instantaneous supply of peroxy radicals (RO_2) required to convert NO to NO_2. High-kOH species provide this radical flux immediately upon emission, rapidly initiating the oxidation cycle. By prioritizing this instantaneous turnover over integrated potential (OFP and OHR), the process-informed framework effectively closes the gap between data-driven modeling and kinetic theory. The model recovers the kinetic dominance of high-kOH species—specifically aromatics and isoprene—consistent with mechanistic box-modeling studies (Lu et al., 2019). This validates that by resolving short-term chemical variances (e.g., the 18-minute window), data-driven approaches can successfully align with the fundamental kinetic principles established in experimental atmospheric chemistry.

4 Conclusions

This study demonstrates that, for the Seoul roadside observations analyzed here, concentration-based machine-learning interpretation can systematically misidentify the kinetic drivers of urban ozone—a 'survivor bias' to which standard data-driven analyses are particularly susceptible in NO_x-saturated, fresh-emission regimes. Within the process-informed (ΔVOC) framework, reactive aromatics (xylenes, TMB) and biogenic species (isoprene, monoterpenes) emerge as the compounds most closely tied to short-timescale ozone production, whereas the more abundant OVOCs primarily reflect accumulated background conditions rather than immediate chemical control. Realized ozone production in fresh urban plumes is therefore governed by precursor turnover and radical-production efficiency rather than by bulk abundance alone. The persistent negative response to NO_x is also consistent with a continuing titration influence during the observation period, suggesting that, in this regime, NO_x reductions are unlikely to translate into proportional peak-ozone reductions without accompanying VOC controls. Importantly, during high-ozone episodes the system′ s response to reactive precursors steepens while the titration efficiency attenuates, indicating that hazardous events are driven not by a regime change but by intensified turnover of specific rate-limiting species—a finding with direct implications for the marginal-benefit calculations underlying VOC control strategies. More broadly, these results indicate that mass-based emission inventories and OFP-style rankings, when applied without kinetic context, can misallocate control priorities in NO_x-saturated urban environments; reactivity-informed and process-constrained metrics should therefore complement—and in such regimes potentially supersede—conventional inventories when informing precursor control. Because the analysis is based on a single roadside site, a one-month spring observation period, an 18-minute integration window, and a PTR-ToF-MS suite that does not fully capture all precursor classes (notably light alkanes), further applications across seasons, source environments, and chemical regimes will be needed to test generality. The framework itself, however, requires only sub-hourly co-located

VOC and ozone observations and is therefore broadly applicable wherever such data are available—including settings where explicit mechanism-based modeling is constrained by incomplete VOC speciation or unmeasured radical precursors.

Data availability.

The observational data and ML code used in this study are available from the corresponding author upon reasonable request.

Author contributions

Conceptualization, Methodology, Writing—review & editing, H.K.; Methodology, Investigation, Modeling, Writing—original draft, Q.H.; Discussion—J.G.; Measurement—S.K., Y.P., S.L.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplemental Information

S1. Measurement results of gas species, VOCs, and meteorology; S2. Details of the photochemical clock and MCM-OBM model; S3. Validation, selection of input parameters, and results of the ML model.

Acknowledgments

This work was supported by the project "development of SMaRT based aerosol measurement and analysis systems for the evaluation of climate change and health risk assessment" operated by Seoul National University (900-20240101).

References

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

Berndt, T., Mentler, B., Scholz, W., Fischer, L., Herrmann, H., Kulmala, M., and Hansel, A.: Accretion Product Formation from Ozonolysis and OH Radical Reaction of α-Pinene: Mechanistic Insight and the Influence of Isoprene and Ethylene, Environ. Sci. Technol., 52, 11069–11077, https://doi.org/10.1021/acs.est.8b02210, 2018.

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.

Chang, L.-S., Choi, J.-Y., Son, J., Lee, S., Lee, D., Jo, Y.-J., and Kim, C.-H.: Interpretation of decadal-scale ozone production efficiency in the Seoul Metropolitan Area: Implication for ozone abatement, Atmos. Environ., 243, 117846, https://doi.org/10.1016/j.atmosenv.2020.117846, 2020.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Chew, A. A. and Atkinson, R.: OH radical formation yields from the gas-phase reactions of O3 with alkenes and monoterpenes, J. Geophys. Res. Atmospheres, 101, 28649–28653, https://doi.org/10.1029/96JD02722, 1996.

Clapp, L. J. and Jenkin, M. E.: Analysis of the relationship between ambient levels of O3, NO2 and NO as a function of NOx in the UK, Atmos. Environ., 35, 6391–6405, https://doi.org/10.1016/S1352-2310(01)00378-8, 2001.

Colombi, N. K., Jacob, D. J., Yang, L. H., Zhai, S., Shah, V., Grange, S. K., Yantosca, R. M., Kim, S., and Liao, H.: Why is ozone in South Korea and the Seoul metropolitan area so high and increasing?, Atmospheric Chem. Phys., 23, 4031–4044, https://doi.org/10.5194/acp-23-4031-2023, 2023.

de Gouw, J. A., Gilman, J. B., Kim, S.-W., Lerner, B. M., Isaacman-VanWertz, G., McDonald, B. C., Warneke, C., Kuster, W. C., Lefer, B. L., Griffith, S. M., Dusanter, S., Stevens, P. S., and Stutz, J.: Chemistry of Volatile Organic Compounds in the Los Angeles basin: Nighttime Removal of Alkenes and Determination of Emission Ratios, J. Geophys. Res. Atmospheres, 122, 11,843-11,861, https://doi.org/10.1002/2017JD027459, 2017.

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.

Huang, X., Zheng, W., Li, Y., Li, C., Li, Y., Zhang, N., Pan, C., Lei, Y., Li, H., Zhang, Y., Qin, Y., Zheng, Y., Fu, T.-M., Yang, X., Ge, X., Zhang, H., and Ye, J.: Ozone Formation in a Representative Urban Environment: Model Discrepancies and Critical Roles of Oxygenated Volatile Organic Compounds, Environ. Sci. Technol. Lett., 12, 297–304, https://doi.org/10.1021/acs.estlett.4c01026, 2025.

Kenagy, H. S., Sparks, T. L., Wooldridge, P. J., Weinheimer, A. J., Ryerson, T. B., Blake, D. R., Hornbrook, R. S., Apel, E. C., and Cohen, R. C.: Evidence of Nighttime Production of Organic Nitrates During SEAC4RS, FRAPPÉ, and KORUS-AQ, Geophys. Res. Lett., 47, e2020GL087860, https://doi.org/10.1029/2020GL087860, 2020.

Li, K., Jacob, D. J., Shen, L., Lu, X., De Smedt, I., and Liao, H.: Increases in surface ozone pollution in China from 2013 to 2019: anthropogenic and meteorological influences, Atmospheric Chem. Phys., 20, 11423–11433, https://doi.org/10.5194/acp-20-11423-2020, 2020.

Liu, Y., Shao, M., Fu, L., Lu, S., Zeng, L., and Tang, D.: Source profiles of volatile organic compounds (VOCs) measured in China: Part I, Atmos. Environ., 42, 6247–6260, https://doi.org/10.1016/j.atmosenv.2008.01.070, 2008.

Lu, K., Guo, S., Tan, Z., Wang, H., Shang, D., Liu, Y., Li, X., Wu, Z., Hu, M., and Zhang, Y.: Exploring atmospheric free-radical chemistry in China: the self-cleansing capacity and the formation of secondary air pollution, Natl. Sci. Rev., 6, 579–594, https://doi.org/10.1093/nsr/nwy073, 2019.

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.

Mauder, M., Cuntz, M., Drüe, C., Graf, A., Rebmann, C., Schmid, H. P., Schmidt, M., and Steinbrecher, R.: A strategy for quality and uncertainty assessment of long-term eddy-covariance measurements, Agric. For. Meteorol., 169, 122–135, https://doi.org/10.1016/j.agrformet.2012.09.006, 2013.

Metzger, S.: Surface-atmosphere exchange in a box: Making the control volume a suitable representation for in-situ observations, Agric. For. Meteorol., 255, 68–80, https://doi.org/10.1016/j.agrformet.2017.08.037, 2018.

Perring, A. E., Pusede, S. E., and Cohen, R. C.: An Observational Perspective on the Atmospheric Impacts of Alkyl and Multifunctional Nitrates on Ozone and Secondary Organic Aerosol, Chem. Rev., 113, 5848–5870, https://doi.org/10.1021/cr300520x, 2013.

Pusede, S. E. and Cohen, R. C.: On the observed response of ozone to NO_x and VOC reactivity reductions in San Joaquin Valley California 1995–present, Atmospheric Chem. Phys., 12, 8323–8339, https://doi.org/10.5194/acp-12-8323-2012, 2012.

Shi, G., Yu, H., Li, J., Zhang, Z., Wang, F., Huang, J., Li, Y., Yan, Y., Feng, Y., and Russell, A. G.: Unexpected Trade-Offs of Fossil Fuel Reduction on PM2.5 and O3 Pollution Regulation Can Be Offset by Synergistic Control of VOCs Source, ACS EST Air, 1, 660–669, https://doi.org/10.1021/acsestair.4c00014, 2024.

Sillman, S.: The relation between ozone, NOx and hydrocarbons in urban and polluted rural environments, Atmos. Environ., 33, 1821–1845, https://doi.org/10.1016/S1352-2310(98)00345-8, 1999.

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.

Tan, Z., Lu, K., Jiang, M., Su, R., Wang, H., Lou, S., Fu, Q., Zhai, C., Tan, Q., Yue, D., Chen, D., Wang, Z., Xie, S., Zeng, L., and Zhang, Y.: Daytime atmospheric oxidation capacity in four Chinese megacities during the photochemically polluted season: a case study based on box model simulation, Atmospheric Chem. Phys., 19, 3493–3513, https://doi.org/10.5194/acp-19-3493-2019, 2019.

Tuazon, E. C., Arey, J., Atkinson, R., and Aschmann, S. M.: Gas-phase reactions of 2-vinylpyridine and styrene with hydroxyl and NO3 radicals and ozone, Environ. Sci. Technol., 27, 1832–1841, https://doi.org/10.1021/es00046a011, 1993.

Wang, F., Zhang, Z., Wang, G., Wang, Z., Li, M., Liang, W., Gao, J., Wang, W., Chen, D., Feng, Y., and Shi, G.: Machine learning and theoretical analysis release the non-linear relationship among ozone, secondary organic aerosol and volatile organic compounds, J. Environ. Sci., 114, 75–84, https://doi.org/10.1016/j.jes.2021.07.026, 2022.

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.

Wang, T., Dai, J., Lam, K. S., Nan Poon, C., and Brasseur, G. P.: Twenty-Five Years of Lower Tropospheric Ozone Observations in Tropical East Asia: The Influence of Emissions and Weather Patterns, Geophys. Res. Lett., 46, 11463–11470, https://doi.org/10.1029/2019GL084459, 2019.

World Health Organization: WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, World Health Organization, 300 pp., 2021.

Zhang, L., Wang, L., Ji, D., Xia, Z., Nan, P., Zhang, J., Li, K., Qi, B., Du, R., Sun, Y., Wang, Y., and Hu, B.: Explainable ensemble machine learning revealing the effect of meteorology and sources on ozone formation in megacity Hangzhou, China, Sci. Total Environ., 922, 171295, https://doi.org/10.1016/j.scitotenv.2024.171295, 2024.

Zhang, Z., Man, H., Zhao, J., Jiang, Y., Zeng, M., Cai, Z., Huang, C., Huang, W., Zhao, H., Jing, S., Shi, X., He, K., and Liu, H.: Primary organic gas emissions in vehicle cold start events: Rates, compositions and temperature effects, J. Hazard. Mater., 435, 128979, https://doi.org/10.1016/j.jhazmat.2022.128979, 2022.

Zhao, D., Wang, Q., Hui, Y., Liu, Y., Wang, F., and Chu, B.: Characteristics, sources, and health risks of volatile organic compounds in different functional regions of Shenyang, Sci. Total Environ., 933, 173148, https://doi.org/10.1016/j.scitotenv.2024.173148, 2024.

Figure Captions

- Figure 1: (a) Ternary diagram of normalized benzene, toluene, and ethylbenzene mixing ratios. Data points are stratified by sampling time. The wireframes indicate source profiles for industrial and solvent emissions, traffic, and biomass/biofuel/coal burning. (b) Box plots of Toluene/Benzene ratios categorized by non-rush, morning rush, and afternoon rush hours. The horizontal line within each box represents the median value.
- Figure 2: SHAP dependence plots comparing the concentration-based model (red scatter points, bottom/left axes) and the delta model (blue scatter points, top/right axes) for (a) NO_x, (b) temperature, (c) RH, (d) solar radiation, and (e) HCHO/NO_x ratio. The color of the markers represents the hour of day, as indicated by the respective blue and red color bars.
- Figure 3: SHAP dependence plots comparing the concentration-based model (red markers, bottom/left axes) and the delta model (blue markers, top/right axes) for: (a) $C_{10}H_{16}$, (b) C_9H_{12}, (c) C_7H_8, (d) C_3H_4O, and (e) C_3H_6O.
- Figure 4: Relationship between SHAP-derived response slope (slope of SHAP dependence for ΔVOC) and chemical reactivity metrics: (a) OH reaction rate constant (kOH), and (b) OH reactivity (OHR, left axis) and ozone formation potential (OFP, right axis) using the same x-axis VOC ordering as in (a). Analysis includes only VOCs with negative SHAP slopes, where consumption correlates with ozone formation. kOH for isomeric groups are weighted averages from KORUS-AQ: C_8H_{10} (1.39×10^{-11}) for ethylbenzene and xylenes; C_9H_{12} (3.60×10^{-11}) for trimethylbenzenes; and $C_{10}H_{16}$ (6.63×10^{-11}) for a 1:1 α- to β-pinene ratio (Kenagy et al., 2020; Simpson et al., 2020).
- Figure 5: SHAP dependence plots for (a) NO_x, (b) $C_{10}H_{16}$, (c) C_9H_{12}, and (d) C_7H_8. Red and green markers represent high ($O_3 > 45$ ppb) and low ($O_3 < 45$ ppb) ozone periods, respectively. Dashed lines denote linear regressions, where the fitted slope indicates the ozone response slope to the precursor.

https://doi.org/10.5194/egusphere-2026-2647
Preprint. Discussion started: 18 June 2026

Figures list

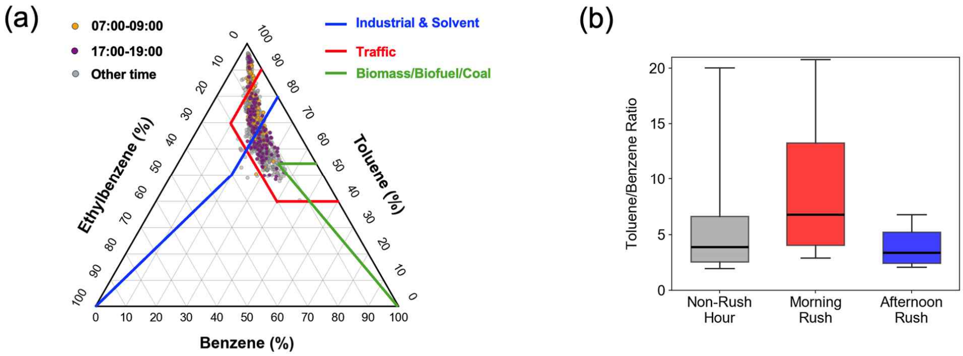

Figure 1: (a) Ternary diagram of normalized benzene, toluene, and ethylbenzene mixing ratios. Data points are stratified by sampling time. The wireframes indicate source profiles for industrial and solvent emissions, traffic, and biomass/biofuel/coal burning. (b) Box plots of Toluene/Benzene ratios categorized by non-rush, morning rush, and afternoon rush hours. The horizontal line within each box represents the median value.

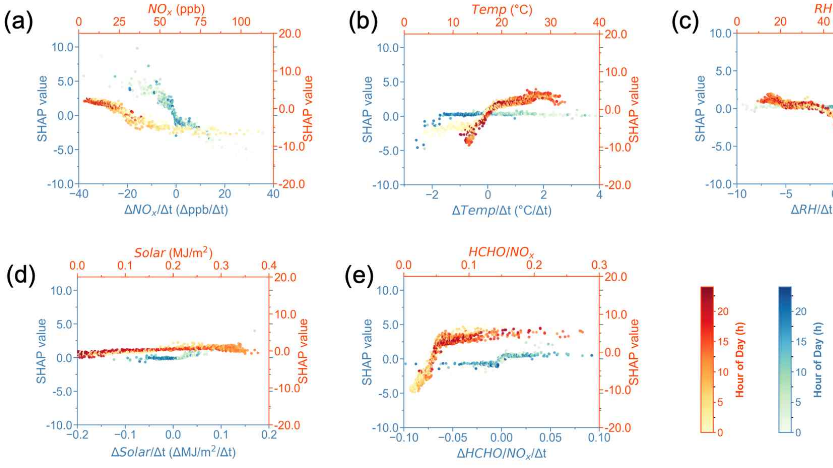

Figure 2: SHAP dependence plots comparing the concentration-based model (red scatter points, bottom/left axes) and the delta model (blue scatter points, top/right axes) for (a) NO_x, (b) temperature, (c) RH, (d) solar radiation, and (e) $HCHO/NO_x$ ratio. The color of the markers represents the hour of day, as indicated by the respective blue and red color bars.

https://doi.org/10.5194/egusphere-2026-2647
Preprint. Discussion started: 18 June 2026

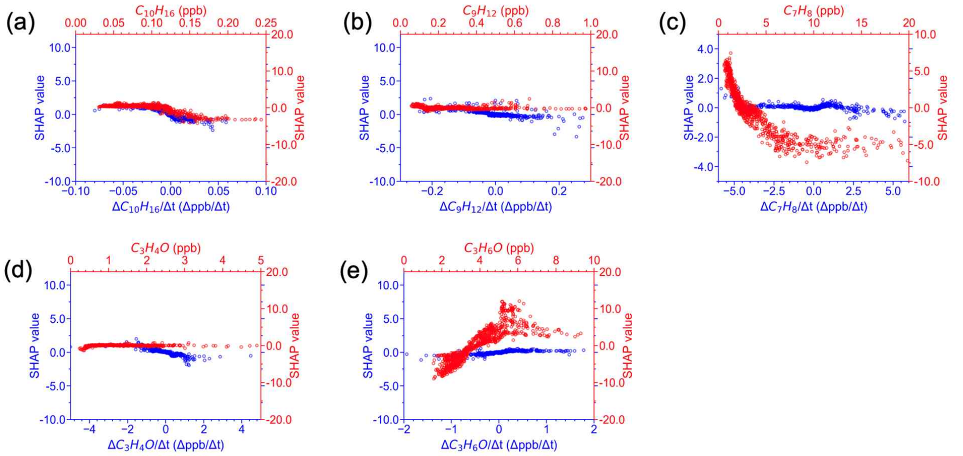

Figure 3: SHAP dependence plots comparing the concentration-based model (red markers, bottom/left axes) and the delta model (blue markers, top/right axes) for: (a) $C_{10}H_{16}$, (b) C_9H_{12}, (c) C_7H_8, (d) C_3H_4O, and (e) C_3H_6O.

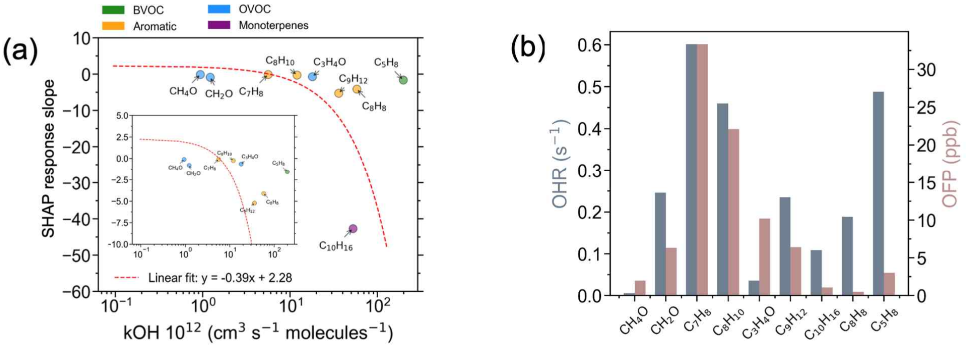

Figure 4: Relationship between SHAP-derived response slope (slope of SHAP dependence for ΔVOC) and chemical reactivity metrics: (a) OH reaction rate constant (kOH), and (b) OH reactivity (OHR, left axis) and ozone formation potential (OFP, right axis) using the same x-axis VOC ordering as in (a). Analysis includes only VOCs with negative SHAP slopes, where consumption correlates with ozone formation. kOH for isomeric groups are weighted averages from KORUS-AQ: C_8H_{10} (1.39×10^{-11}) for ethylbenzene and xylenes; C_9H_{12} (3.60×10^{-11}) for trimethylbenzenes; and $C_{10}H_{16}$ (6.63×10^{-11}) for a 1:1 α- to β-pinene ratio (Kenagy et al., 2020; Simpson et al., 2020).

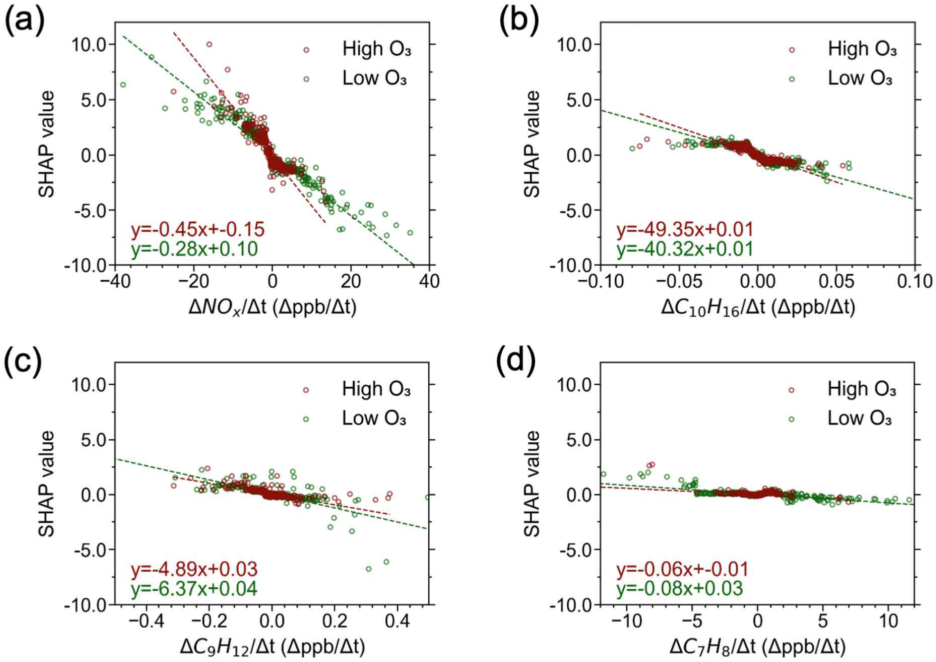

Figure 5: SHAP dependence plots for (a) NO_x, (b) $C_{10}H_{16}$, (c) C_9H_{12}, and (d) C_7H_8. Red and green markers represent high (O_3 > 45 ppb) and low (O_3 < 45 ppb) ozone periods, respectively. Dashed lines denote linear regressions, where the fitted slope indicates the ozone response slope to the precursor.