

Response to Reviewer #3

General Comment

In their manuscript 'Reactivity Over Abundance: Unveiling the True Kinetic Drivers of Urban Ozone Using Process-Informed Machine Learning', Hu et al. present a process-informed machine learning framework to predict urban ozone formation and demonstrate its use with data from a roadside air quality monitoring station in Seoul, South Korea. They demonstrate that this approach is more reliable in NO_x-saturated environments, where other methods relying on ozone precursors' concentrations suffer from "survivor bias".

I find that the results are presented clearly and that the arguments in the conclusions are satisfactory, demonstrating the usefulness of the proposed framework. However, while the substance of the manuscript is new and interesting, the form could be improved in my opinion. If the authors can address these minor comments, the manuscript should be accepted for publication in ACP.

We appreciate the reviewer's careful evaluation of the framework and presentation. We address the requested clarifications regarding the chemical-regime scope, SHAP interpretation, kinetic calculations, terminology, and reproducibility below.

Comment 1

As it seems a pretty important caveat that this framework is useful in NO_x-saturated (VOC-limited) environments, I wondered if it would be worth including this in the title somehow.

We thank the reviewer for this suggestion. We agree that the chemical-regime scope should be visible in the title and have revised it accordingly.

Line 1: Reactivity Over Abundance: Unveiling the True Kinetic Drivers of Urban Ozone Using Process-Informed Machine Learning **After:** "Reactivity Over Abundance: Unveiling the True Kinetic Drivers of Urban Ozone Using Process-Informed Machine Learning in a NO_x-Saturated Urban Environment"

Comment 2

The last part of the introduction (lines 60 to 75) seems to me to fit more as an abstract, while some sentences of the abstract are quite general and would fit better in the introduction to leave space in the abstract to describe the presented study and results.

We appreciate the reviewer's suggestion. We reorganized the Abstract around the actual Seoul application: it now identifies the NO_x-saturated roadside setting, defines ΔVOC as a net concentration change, reports the acetone-toluene contrast and the association between model response and OH reaction rate constants, and states the measured-suite and chemical-regime scope. We rewrote the final Introduction paragraph around the survivor-bias hypothesis, the comparison between concentration-based and process-informed attribution, and the independent chemical diagnostics used to test the interpretation. Species-level findings and proposed mechanisms are reserved for the Abstract and Results. The resulting 227-word Abstract is a concise, self-contained summary of the work, while the Introduction now concludes with the study hypothesis and objectives rather than the results.

- Line 14:** Ground-level ozone remains a persistent challenge in East Asian urban centers, where concentrations continue to rise despite reductions in precursor emissions. Designing effective mitigation strategies is complicated by the non-linear relationship between precursor abundance and reactivity. Standard data-driven approaches can undervalue kinetic drivers through "survivor bias," in which models assign low importance to highly reactive precursors that are rapidly depleted. In this study, we introduce a process-informed machine learning framework that utilizes net chemical consumption (ΔVOC) rather than ambient concentrations to resolve this attribution failure. We developed a process-informed machine-learning framework that uses 18 min net VOC changes (ΔVOC) to predict concurrent O_3 changes (ΔO_3) in sub-hourly observations at a Seoul roadside site during April 2022. **After:** "O₃ control in NO_x (nitrogen oxides) saturated urban air depends on identifying volatile organic compounds (VOCs) that sustain radical propagation. Ambient concentration can undervalue highly reactive VOCs depleted before sampling. We developed a process-informed machine-learning framework that uses 18 min net VOC changes (ΔVOC) to predict concurrent O_3 changes (ΔO_3) in sub-hourly observations at a Seoul roadside site during April 2022."
- Line 18:** Integrating 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 ozone formation potential (OFP) rank precursors by their stoichiometric capacity under idealized accumulated conditions, the process-informed framework reveals 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; reactive aromatics and biogenics emerge as the rate-limiting controls on peak ozone in fresh-emission environments. 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. **After:** "The process-informed XGBoost model achieved a mean 10-fold cross-validated R² of 0.79. The concentration-based model ranked acetone above toluene, whereas the process-informed model assigned the strongest responses to reactive aromatics and biogenic VOCs. Independently, the MCM-based observation-based model estimated approximately an order of magnitude more direct peroxy-radical initiation from toluene than from acetone. This contrast shows that the concentration ranking did not reflect their relative direct radical-initiation fluxes. Across the measured VOC suite, process-informed SHAP response slopes were associated with OH reaction rate constants. Departures from this relationship were consistent with organic-nitrate termination and monoterpene ozonolysis. During high-O₃ periods, the aggregate response to reactive precursors strengthened while the negative NO_x response persisted, consistent with stronger kinetic selectivity in the observed regime. Together, these results show that short-term precursor turnover, interpreted with independent chemical diagnostics, can reveal kinetic O₃ drivers that abundance-based attribution obscures. The species ranking is specific to the measured precursor suite and chemical conditions sampled at this site, but the framework provides a testable approach for diagnosing turnover-sensitive VOC attribution in comparable urban environments."
- Line 57:** 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 Δ VOC 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 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, ClNO₂), or computational cost. **After:** "We hypothesize that selective chemical depletion in fresh, NO_x-saturated urban air causes concentration-based models to underrepresent reactive VOCs associated with short-term O₃ changes. To test this hypothesis, we developed a process-informed machine-learning framework that uses net VOC concentration changes over a sub-hourly interval (Δ VOC) to predict concurrent net O₃ changes (Δ O₃). We apply the framework to an April 2022 roadside dataset from Seoul and compare its attribution with that of a conventional concentration-based model. We then assess the chemical consistency of the process-informed attribution using VOC reaction kinetics, conventional reactivity indicators, photochemical-clock behavior, and an observation-based model using the MCM. This design tests whether short-term net concentration changes recover kinetic VOC influences that are obscured when standing abundance is interpreted alone."

Comment 3

Section 2.2.2: The configuration of the model and the introduction of shapley additive explanations (SHAP) is condensed in one sentence/two lines. SHAP being such an essential part of the manuscript (every single plot includes SHAP values), it might be worth describing the concept more in the main text. I would suggest bringing the text S7 to the manuscript, including Equation (2) (1?). (Note that it is the only Supplementary Text that is not referred to in the main text. More technical observations below.) Especially for a journal like ACP whose readership might not be entirely familiar with machine learning concepts, it would be useful.

We thank the reviewer for noting that SHAP required a fuller explanation for ACP readers. We expanded the Methods to explain how SHAP decomposes each prediction into a baseline and additive feature contributions and how the sign and magnitude of those contributions are interpreted. The revised text also distinguishes model attribution from causal inference.

Line 109 Added: "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) (Lundberg et al., 2020). The SHAP output can be expressed in a linear additive form, as shown in Equation (1). These SHAP values represent model attributions and were interpreted together with the kinetic, photochemical-clock, and MCM-based OBM diagnostics described below.

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

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."

Comment 4

Section 3.1: The authors describe how they use a fixed ratio from another campaign for C₈H₁₀ as PTR-ToF-MS cannot chemically resolve ethylbenzene and xylenes. Table S2 indicates ranges of kOH values for xylenes, trimethylbenzenes, and monoterpenes and in the text there is a mention of kOH of "36.0 x 10⁻¹²" as "average of three isomers", but it is not clear if it is a weighted average or how the average was calculated as it does not seem consistent with Table S2. Also, it is not clear for monoterpenes which kOH is used and why and how it affects the results. Clarifying these numbers would help readers who would like to apply the framework in a consistent way with their own data.

We thank the reviewer for requesting clarification. We now calculate effective kOH using KORUS-AQ Seoul constituent fractions from Simpson et al. (2020) and kinetic constants from Atkinson and Arey (2003). This gives 13.9, 36.0, and 63.5 × 10⁻¹² cm³ molecule⁻¹ s⁻¹ for C₈H₁₀, C₉H₁₂, and C₁₀H₁₆, respectively. Thus, 36.0 is a weighted mean within the Table S2 range, and the monoterpene value has been corrected from 66.3 to 63.5. These values were used only for post hoc kinetic comparisons and did not determine the SHAP slopes. Figure 4a now shows the corresponding composition-sensitivity whiskers. The fitted kOH–SHAP slope remains negative across their limits, although its magnitude varies. Table S2 provides the constituent fractions and kinetic data sources used in these calculations.

- **Line 554: Updated Figure 4a**

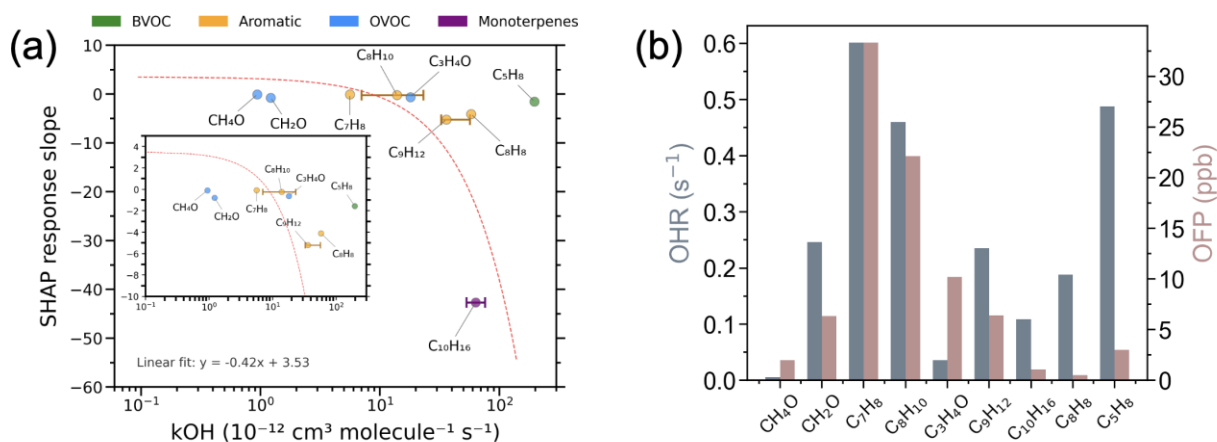


Figure 4: Relationship between SHAP-derived response slope (slope of SHAP dependence for ΔVOC) and chemical reactivity metrics: (a) OH reaction rate constant (kOH), and (b) OH reactivity (OHR, left axis) and ozone formation potential (OFP, right axis) using the same x-axis VOC ordering as in (a). Analysis includes only VOCs with negative SHAP slopes, where consumption correlates with ozone formation. Isoprene is shown in (a) but excluded from the regression fit due to concurrent emission-oxidation covariance. For unresolved PTR-ToF-MS channels, post hoc kOH values were calculated as constituent-fraction-weighted means using the KORUS-AQ Seoul composition (Simpson et al., 2020). The central values were 1.39×10^{-11} for C_8H_{10} , 3.60×10^{-11} for C_9H_{12} , and $6.35 \times 10^{-11} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$ for $\text{C}_{10}\text{H}_{16}$. Table S2 reports the constituent fractions and kinetic data sources. Horizontal whiskers show conservative composition-sensitivity ranges rather than statistical confidence intervals. kOH was not an XGBoost input and therefore did not determine the SHAP values.

- **Line 530 (Figure captions) and 559:** Fig. 4 caption. **After:** "For unresolved PTR-ToF-MS channels, post hoc kOH values were calculated as constituent-fraction-weighted means using the KORUS-AQ Seoul composition (Simpson et al., 2020). The central values were 1.39×10^{-11} for C_8H_{10} , 3.60×10^{-11} for C_9H_{12} , and $6.35 \times 10^{-11} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$ for $\text{C}_{10}\text{H}_{16}$. Table S2 reports the constituent fractions and kinetic data sources. Horizontal whiskers show conservative composition-sensitivity ranges rather than statistical confidence intervals. kOH was not an XGBoost input and therefore did not determine the SHAP values."

Comment 5

The 'Data availability' statement does seem to be sufficient according to ACP guidelines and the authors should justify why the data and/or model cannot be made available through a FAIR-aligned reliable public data repository.

We appreciate the reviewer's emphasis on FAIR data access. The observational dataset forms part of an ongoing institutional measurement program at Seoul National University, with related analyses still in progress, which precludes unrestricted public deposition at this stage. We have therefore retained request-based access, which is accepted under ACP's data policy for Research Articles. The specific dataset and

analysis code used in this study will be provided to any researcher upon request to the corresponding author for verification and reproduction purposes. We have clarified this in the Data availability statement.

- **Line 388:** The observational data and ML code used in this study are available from the corresponding author upon reasonable request. **After:** "The observational data and machine-learning code used in this study are available from the corresponding author upon reasonable request. The observational dataset forms part of an ongoing institutional measurement program at Seoul National University, and related analyses remain in progress. National air quality and meteorological observations are available through the AirKorea and Korea Meteorological Administration open-data portals, respectively."

Technical Comments

In addition, I found some small inconsistencies in the way abbreviations are used and introduced. Here a few examples of these technical corrections:

- In the abstract, Delta VOC is introduced as "net chemical consumption", but VOC is not defined. Nor is NO_x or OFP. In the introduction, Delta VOC is introduced as "net VOC consumption" and later as "net change in VOC concentration".
- "Delta" itself is used on its own once after 'processed-informed' and once as 'rate of change', which seems unnecessary to be mentioned like that. Also, "delta model" (not even Delta as defined earlier) is used as shorthand for "processed-informed model", which I would discourage. It would be better to keep a unified nomenclature throughout the manuscript.
- "OHR" is introduced on line 124, but only defined later on line 134. Of the compounds presented in the figures using chemical formulas, only monoterpenes seem to be defined in the main text and while Table S2 summarizes the compounds and abbreviations, it might be good to at least define them once in the main text as well. For instance, on line 93, the authors define "formaldehyde (HCHO)", but not "acetone".
- Some concepts are capitalized before the definition of their abbreviation (e.g. SHAP), but this is not according to ACP guidelines.

We thank the reviewer for identifying these terminology and abbreviation inconsistencies. We revised the first-use definitions of VOCs, NO_x, OHR, and OFP, standardized Δ VOC as the net VOC concentration change over the selected interval, removed standalone uses of " Δ ," and replaced "delta model" with "process-informed model." We also defined the compounds shown as formulas in Fig. 3 and corrected the capitalization of Shapley additive explanations.

References

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

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.