Aalto University
School of Science

Dr. Sergey Gromov
Handling Topic Editor
Geoscientific Model Development

December 6, 2024

Dear Dr. Sergey Gromov,

We extend our gratitude for providing us with the reviewer reports. We also wish to express our appreciation to the referees for their evaluations. The manuscript has undergone revisions aimed at enhancing its clarity, and we are pleased to submit the revised version for your consideration.

Sincerely,
Prof. Patrick Rinke and Hilda Sandström

Response to comments by Handling Topic Editor

Dear Hilda Sandström and Patrick Rinke,
Thank you for submitting your manuscript entitled "Similarity-Based Analysis of Atmospheric Organic Compounds for Machine Learning Applications" to EGUSphere/GMD. I agree that it is reasonable to transfer the submission from ACP to GMD, as the MS topic fits the latter much better. I acknowledge the MS careful preparation and will be happy to oversee the discussion process. However, there are a few technical corrections required prior to the publication of the MS for the discussion in GMDD.

1. *The major correction concerns complying with the GMD Code and Data Policy (see* `https://www.geoscientific-model-development.net/policies/code_and_data_policy.html` *) which the journal is very strict to follow. In particular, we require that archived public versions of the code and data associated with the paper must be made available. The code and workflow used for data analysis you provide via a GitLab repository, which does not comply with these (e.g. there is no DOI, no guaranteed long-term archiving, no versioning, etc.) Therefore, please prepare the actual used code and workflow as a publication at one of suitable platforms (I recommend Zenodo) and reference this publication in the "Code Availability" section of the manuscript instead.*

   Thank you for bringing this to our attention. In response to yours and the Chief Editor's comments, we have now uploaded the versions of the code used in our manuscript to Zenodo, and you can access them through the following links:

   (a) Code Repository: `DOI:10.5281/zenodo.14224079`, `https://zenodo.org/records/14224079`
   (b) Additional Code: `DOI:10.5281/zenodo.14007835` , `https://zenodo.org/records/14007835`

   Furthermore, the datasets utilized for our analysis are freely available, as detailed in our README.md file at this link: `DOI:10.5281/zenodo.14224079` (`https://zenodo.org/records/14224079`).

   We have also ensured that a proper license is included with the uploaded code, specifically the GPLv3 license, as recommended.

   Finally, we have updated the 'Code and Data Availability' section in the manuscript to include the DOI of the newly uploaded code and any other necessary details:

"The datasets used for this analysis are all freely available from original publications or the database website (see Table 1). Code and workflow used for data analysis are freely available in Zenodo. Code repository: DOI:10.5281/zenodo.14224079. Additional Code: DOI:10.5281/zenodo.14007835."

2. *Figures 4a and 5a – form the overlapping histogram bars with transparent colours it is nearly impossible to distinguish which ones belong to which dataset, especially for people with impaired colour perception. I suggest that you re-render these figures either with connected step-lines of different colours/patterns, or plot the bars for each dataset next to each in a given bin.*

   We now updated Figures 4a and 5a so that the bars for each dataset are next to each other for improved clarity.

3. *Sect. 3.2.1 mentions Supplementary information; however, it appears to be the Appendix to the MS you imply. Please note that in Copernicus journals the former is published as a standalone document (not included in the main MS), in contrast to the Appendix. Please amend the last paragraph of Sect. 3.2.1 accordingly (i.e. refer to Appendix A).*

   We now modified the sentence in Section 3.2.1 to refer to Appendix A instead of Supplementary information:

   "We tested the robustness of our t-SNE analysis with respect to different perplexity hyperparameter values (Appendix A, see Figure A1 and Figure A2 and refer to Methods for a brief explanation)."

Response to comments by Reviewer #1

H. Sandstrom and P. Rinke conducted a study focused on the similarity-based analysis and its various datasetes containing organic compounds. They highlighted the challenges posed by the lack of curated datasets for atmospheric molecules and aimed to connect atmospheric compounds with existing large molecular datasets. Their investigation revealed that atmospheric molecules have limited overlap with non-atmospheric compounds due to distinct functional groups and atomic compositions. They utilized two molecular similarity metrics, specifically t-SNE and the Tanimoto similarity index, to compare atmospheric datasets (Wang, Gecko, and Quinones) between themseves and with non-atmospheric datasets (including drug-like and metabolite compounds). Their findings emphasize the need for collaborative efforts to improve dataset curation in order to enhance machine learning applications in atmospheric sciences.
From my point of view, their manuscript is well-written. All methods are well explained and referenced, and the text is easy to read and understand. The data manipulation and presented results are sufficiently explained. I do have minor (or rather nitpicking) suggestions for improving the manuscript (see below). Nevertheless, I am very pleased to recommend this manuscript for publication.

1. *Regarding Equation 1, it appears to be incorrect. Since the surrounding text and graphs make sense, I assume this is just a typo. Nevertheless, the correct equation should be:*
   either: $|A \bigcap B|/|A \bigcup B|$
   or: $|A \bigcap B|/(|A| + |B| - |A \bigcap B|)$
   but not: $|A \bigcap B|/(|A \bigcup B| - |A \bigcap B|)$

   We thank the reviewer for this correction. Indeed, the Tanimoto index should consider the fraction of shared features compared to the number of features in the combined set.

We have corrected the equation and text to read: "In contrast, the Tanimoto index, $S_{A,B}$, offers a quantitative measure of similarity. $S_{A,B}$ is calculated as the fraction of present features (represented by non-zero bits) that are shared compared to the total number of present features in molecules A and B, according to

$$S_{A,B} = \frac{\sum_{A \cap B} 1}{\sum_{A \cup B} 1}. \quad (1)$$

"

2. *The Tanimoto similarity distribution does indeed provide some information on the similarity between the two datasets. However, would it not be even more relevant for machine learning applications to compare the distributions of the highest Tanimoto similarity indices, taken between compounds from the analyzed dataset and all compounds in the reference dataset?*

   Thank you for raising this suggestion. We acknowledge that analyzing the highest similarity per compound can indeed provide valuable insight, particularly for certain machine learning applications. This approach has been especially useful for assessing model applicability in specific cases (Liu et al., 2018; Moret et al., 2023). However, the primary aim of our study is to evaluate the relevance of established datasets or databases for atmospheric science. Consequently, while a highest similarity per compound analysis is informative, it does not alter our central conclusion: that large molecular datasets currently available are generally not well suited for atmospheric science (as further elaborated in the newly added Tables C1-C4 in Appendix C).

   In response to your suggestion, we have incorporated additional plots to examine the highest similarity between compounds in our datasets and those in the reference datasets. Figures D1 and D2 present the highest similarity of compounds from our datasets to the reference datasets, while Figures D3 and D4 show the highest similarity of the reference compounds to all datasets.

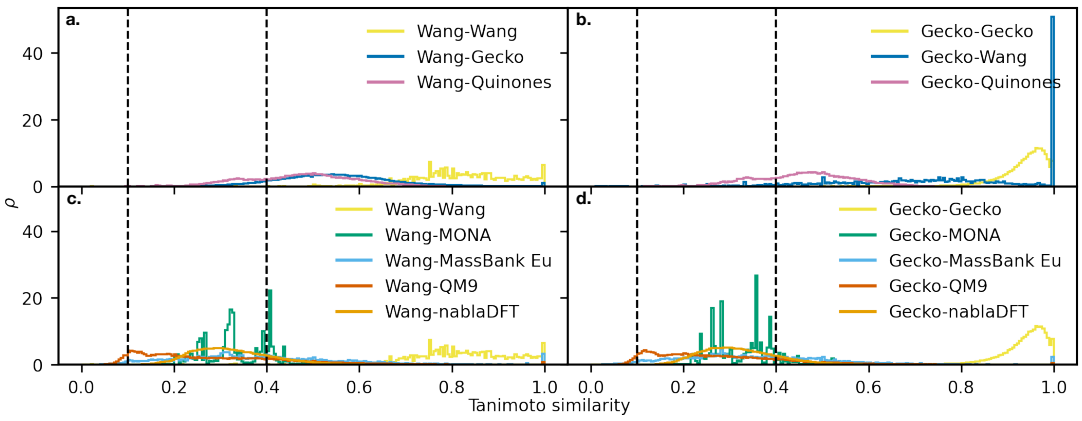

Figure D1: The distributions of maximum Tanimoto similarity are shown for non-reference compounds in different comparisons based on topological fingerprints. Panels (a) and (b) depict the distributions of maximum Tanimoto similarity between atmospheric molecules and Wang molecules (a) and Gecko molecules (b), respectively. Panels (c) and (d) present the distributions for comparisons between non-atmospheric molecules and Wang molecules (c) and Gecko molecules (d). Vertical lines at similarity values of 0.1 and 0.4 indicate reference values for low and high similarity, respectively. The histograms have been normalized such that the area under each curve equals 1.

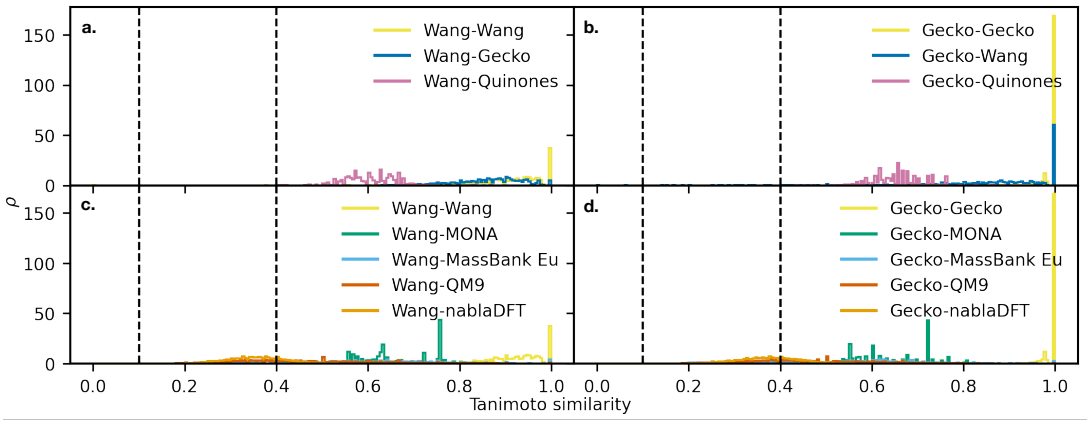

Figure D2: The distributions of maximum Tanimoto similarity are shown for non-reference compounds in different comparisons based on MACCS fingerprints. Panels (a) and (b) depict the distributions of maximum Tanimoto similarity between atmospheric molecules and Wang molecules (a) and Gecko molecules (b), respectively. Panels (c) and (d) present the distributions for comparisons between non-atmospheric molecules and Wang molecules (c) and Gecko molecules (d). Vertical lines at similarity values of 0.1 and 0.4 indicate reference values for low and high similarity, respectively. The histograms have been normalized such that the area under each curve equals 1.

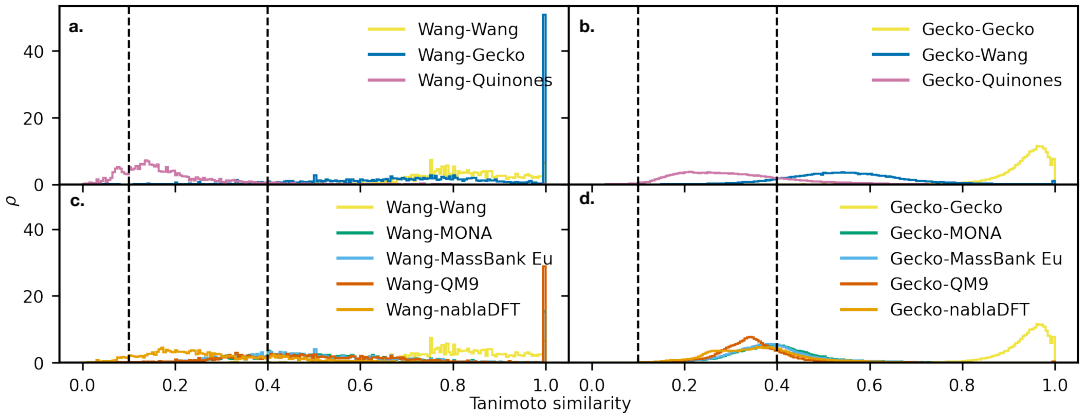

Figure D3: The distributions of maximum Tanimoto similarity are shown for reference compounds in different comparisons based on topological fingerprints. Panels (a) and (b) depict the distributions of maximum Tanimoto similarity between Wang molecules (a) and Gecko molecules (b) and atmospheric molecules, respectively. Panels (c) and (d) present the distributions for comparisons between Wang molecules (c) and Gecko molecules (d) and non-atmospheric molecules. Vertical lines at similarity values of 0.1 and 0.4 indicate reference values for low and high similarity, respectively. The histograms have been normalized such that the area under each curve equals 1.

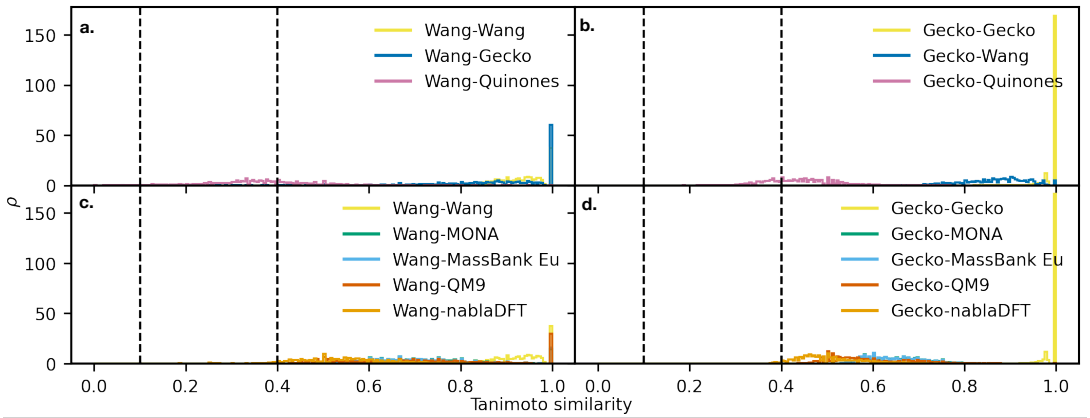

Figure D4: The distributions of maximum Tanimoto similarity are shown for reference compounds in different comparisons based on MACCS fingerprints. Panels (a) and (b) depict the distributions of maximum Tanimoto similarity between Wang molecules (a) and Gecko molecules (b) and atmospheric molecules, respectively. Panels (c) and (d) present the distributions for comparisons between Wang molecules (c) and Gecko molecules (d) and non-atmospheric molecules. Vertical lines at similarity values of 0.1 and 0.4 indicate reference values for low and high similarity, respectively. The histograms have been normalized such that the area under each curve equals 1.

These figures are included in Appendix D, with the following summary added to the main text in the Results section:

"In Appendix D, we investigate a subset of the Tanimoto similaries that belong to the nearest-neighbors (i.e. compounds with highest similarity). Such a comparison could reveal if the large datasets have local subsets in the high-similarity region. In Figure D1, the nearest-neighbor similarity for the topological fingerprint is shown. All atmospheric compounds have nearest neighbors in the high-similarity region within the reference datasets. In contrast, the majority of non-atmospheric datasets have nearest neighbors in the intermediate similarity region. Figure D2, which depicts the nearest-neighbor similarity for the MACCS fingerprints, reveals a similar trend: most datasets show nearest neighbors in the high-similarity region, with the exception of nablaDFT.

Figures D3 and D4 provide additional context. In Figure D3, which considers the topological fingerprint, most Wang compounds have nearest neighbors in the high-similarity region, whereas nablaDFT and Quinones are exceptions. Nearest neighbors of Gecko compounds predominantly fall in the intermediate similarity region, with the exception of Wang compounds. For the MACCS fingerprints in Figure D4, Wang and Gecko compounds both generally have nearest neighbors in the high-similarity region. However, nablaDFT and Quinones are notable exceptions, with Quinones being the only dataset where the majority of nearest neighbors fall below the high-similarity threshold. This result could be explained by the homogeneity of the Quinones dataset, which consists of a single compound class, limiting the structural diversity of potential nearest neighbors.

From these comparisons (Figures D1-D4), we observe that while some Wang and Gecko compounds have high-similarity nearest neighbors in non-atmospheric datasets, the overall suitability of existing datasets for atmospheric science remains limited. "

3. *The last sentence of Section 2.1 is hard to follow (during the first reading). Please try to*

be more descriptive.

We have updated the sentence for clarity. It now reads: "Moreover, a nearest neighbor similarity to the training set above 0.4 indicates enhanced prediction performance and increased machine learning model confidence (Liu et al., 2018; Moret et al., 2023)".

4. *Could you please elaborate on the role of dataset size and diversity? How would the similarity comparison change if, for example, the MONA dataset were removed from the t-SNE analysis? Also, have you tried shuffling the datasets and comparing again? Would you obtain the same conclusions? The size and distance in the t-SNE analysis are not informative—does it even make sense to use this analysis for similarity comparison or any filtering? I ask this to understand whether Figures 6a and 6b are truly different due to the choice of different representations, or if the differences arise because t-SNE is highly sensitive to initial conditions.*

   Thank you for your thoughtful questions. We emphasize that in the context of t-SNE, the distances between clusters should not be the primary focus. t-SNE is designed to preserve local structures and relationships, which means that while clusters can be visually identified, the absolute distances between them may not accurately represent meaningful differences or relationships in the data. Our analysis primarily relies on qualitative visual inspection to identify local shared versus non-shared clusters, rather than quantifying distances between different clusters.

   Data shuffling can influence t-SNE outcomes due to the algorithm's sensitivity to initialization. For both Figures 6a and 6b, we ensured consistency by parsing the data in the same order and using the same random seed for initialization. This approach helps minimize variability in the resulting visualizations. Thus, the differences seen between 6a and 6b are due to the different molecular representations. We have now added Figure A3 in Appendix A which shows the t-SNE results when the datasets are parsed in reverse order compared to the analysis presented in Fig. 6 in the main text. We observe the overall same separation of datasets and cluster formation. The main observed difference is that the overlap between the quinones and MONA in Figure A3b appears more limited than in Figure 6b when clustering the MACCS fingerprints.

   We have added Figure A4 in Appendix A which shows t-SNE clusters when MONA is excluded from the analysis. MONA has some overlap with the atmospheric datasets we are interested in, making it the most similar non-atmospheric dataset based on qualitative and quantitative similarity evaluations in this study. Yet, MONA's overall contribution to the conclusions from t-SNE clustering regarding the number of shared clusters is limited. Removing the MONA dataset did not change the number of shared clusters among the remaining datasets (Fig. A4). In summary, given that MONA is one of the largest datasets and occupies a substantial amount of space in the t-SNE plot removing the set will alter the appearance of the plot, but the remaining datasets cluster together or apart in a similar fashion.

5. *Nitpicking note on Figure 4 caption: Functional group which are at least in 10% of dataset are shown in c), but in the end you show even smaller fraction (i.e. peaks below 0.1), which just made me wonder whether I understand the graphs correctly.*

   We thank the reviewer for pointing out this typo. 1 % is the correct threshold used to generate the figure. We have now adjusted the captions of the figures mentioned.

   The captions now read: "Molecular structure analysis of the atmospheric molecules in terms of molecular size as represented by non-hydrogen atom count (a, histogram normal-

ized so bar heights sum to one), mean and standard deviation of atomic ratios (b), and functional groups (c, present in $\geq$ 1 % of dataset)." and "Molecular structure analysis of the non-atmospheric molecules in terms of molecular size as represented by non-hydrogen atom count (a, histogram normalized so bar heights sum to one), mean and standard deviation of atomic ratios (b), and functional groups (c, present in $\geq$ 1 % of dataset)."

6. *Another nitpicking note: It would be nice if the figures 4a and 5a use the same bin sizes (and scales).*

   We have now adjusted Figures 4a and 5a so that they both share the same scale on the x axis and bin width. Note that the data in the plots are mostly to the left of the plot as only MONA has some fraction of data at higher non-hydrogen atom counts (up to 230).

Response to comments by Reviewer #2

The manuscript by Sandström and Rinke investigates the similarity of organic compounds in multiple different datasets with focus on atmospheric oxidation products. In addition to comparing the molecular descriptors, the authors compared other molecular attributes between the datasets. The study shows how the compounds present in large data banks do not coincide with atmospherically relevant compounds. Therefore, these data banks are not sufficient training data for machine learning models in atmospheric studies. This is an important observation for future development of machine learning models and datasets compiled for the training of those models. I happily recommend that the article should be accepted after minor corrections.

1. *Related to the first paragraph of page 14, how is the size of the datasets taken into account in the Tanimoto similarity analysis? For example comparing Gecko and Wang, Gecko has 166434 compounds and Wang only 3414. It's obvious that 166434 easily contains more compounds that are similar to others, because the total number is just so big. If you were to take 3414 of the most different compounds from the Gecko dataset, would the result be similar to the Wang-Wang Tanimoto distribution? Or the opposite, if you would increase the size of the Wang dataset to 166434 compounds, would it be possible to create equally diverse set of atmospherically relevant oxidized organics? If the distributions were plotted without normalization, would the Gecko-Gecko distribution in the low similarity region still be higher in absolute values than the Wang-Wang distribution?*

   Our dataset analysis focuses on normalized Tanimoto similarity distributions, with relative rather than absolute numbers. While we appreciate the reviewer's point that larger datasets increase the likelihood of high similarity matches, we believe precursor diversity and the number of oxidation generations simulated are more relevant in explaining differences between Wang and Gecko. In atmospheric chemistry, varied precursors produce distinct oxygenated products, reflecting real-world chemical diversity. Although Gecko is larger, it was derived from only three precursors, limiting its structural diversity. Wang, despite being smaller, was derived from 143 precursors, yielding greater molecular variation (see e.g., elemental composition in Table 1). Meanwhile, Gecko's larger size resulted from simulations of successive reactions rather than from diverse inputs.

   A dataset as diverse as Wang and as large as Gecko could theoretically be constructed with a more varied set of precursors (e.g., 1,500), providing broader molecular diversity for robust model generalization. Thus, while filtering similar structures in Gecko may reduce redundancy, its diversity limitations stem mainly from the precursors, not just dataset size.

In response to another comment below, we have tabulated similarity percentages in low, intermediate, and high similarity regions (Appendix C, Tables C1-C4). Notably, the Gecko-Gecko distribution has 0% (0 million) and 3.1% (860 million) of pairs in the low similarity category for MACCS and topological fingerprint comparisons, respectively, compared to 5.7% (0.66 million) and 45.9% (5.3 million) for Wang-Wang. These values illustrate that absolute counts obscure how smaller datasets like Wang inherently contain more low-similarity pairs relative to their dataset size compared to larger datasets like Gecko.

2. *In the Tanimoto similarities, it would be interesting to see the percentages of the compounds in each of the similarity categories (low, intermediate, high).*

   This is an interesting point. We have now added Tables C1-C4 in Appendix C, which show what percentage of the comparisons fall in the low, intermediate and high similarity regions. However, we stress that these reference values for high and low similarity are only meant to serve as indicators for the reader, and are not established hard cutoffs.

3. *Page 1: "However, the underlying molecular-level processes involving organic molecules remain poorly understood, due to the vast number of organic compounds participating in atmospheric chemistry." For readers who are not familiar with atmospheric aerosol, add before this a sentence of how these organic compounds are connected to the aerosol particles you mention in the previous sentences (presumably SOA, since you talk about particle formation*

   We have now added a sentence on the relation between the secondary organic aerosol particles and organic compounds, and the corresponding section reads: "However, the underlying molecular-level processes involving organic molecules remain poorly understood, due to the vast number of organic compounds participating in atmospheric chemistry. Many of these particles are formed through the oxidation of volatile organic compounds, which leads to the formation of so called secondary organic aerosols in the atmosphere (Bianchi et al., 2019). This existing gap in knowledge hampers a comprehensive understanding of particle formation and growth in different environments (Masson-Delmotte et al., 2021; Elm et al., 2020)."

4. *Page 1: "human-based activities, like" -> "human-based activities, such as"*

   We have made the requested modification and the text now reads: "By doing so, we can provide a tool to tailor machine learning models for studies of aerosol particle formation and the effects human-based activities, such as industry and agriculture, on the formation process."

5. *Page 1: "Organic aerosol particle formation" -> "Secondary organic aerosol particle formation"*

   We made the suggested change and the sentence now reads: "Secondary organic aerosol particle formation is affected by atmospheric composition and molecular emissions into the atmosphere."

6. *Page 2: "datasets like Gecko" -> "datasets such as Gecko"*

   We have made the requested modification and the text now reads: "With such model simulations, atmospheric molecular datasets such as Gecko [...]"

7. *"degradation of 143 atmospheric compounds" Can you be more specific? Are these all organics? Hydrocarbons/VOCs or already oxidized species?*

Thank you for asking for this clarification. Currently, MCM includes the degradation of methane and 142 non-methane VOCs; specifically, hydrocarbons and oxygenated VOCs, primarily biogenic species: isoprene, three monoterpenes, one sesquiterpene, as well as one oxygenated VOC and one organosuphur species.

We have modified the sentence to read: "The Wang dataset (Wang et al., 2017) was constructed using MCM (Jenkin et al., 1997; Saunders et al., 2003) to simulate the atmospheric degradation of 143 atmospheric compounds (methane and 142 non-methane volatile organic compounds) by photolysis and reactions with OH, NO_3 and O_3."

8. *Page 3: "In recent years, machine learning methods have shown promise..." Hyttinen et al., 2022 doesn't use machine learning methods.*

   Thank you for pointing out this mistake. We have now replaced the erroneous reference and added the correct and intended paper by Hyttinen et al. from 2022: Hyttinen, N., Pihlajamäki, A., and Häkkinen, H.: Machine Learning for Predicting Chemical Potentials of Multifunctional Organic Compounds in Atmospherically Relevant Solutions, Journal of Physical Chemistry Letters, 13, 9928–9933, 2022.

9. *Page 6: "We tested three different perplexity values of 5, 50 and 100." Since perplexity is an important hyperparameter in t-SNE, a short explanation of its meaning here would be useful.*

   We agree with the reviewer that this is important. We note that the text on page 6 read:

   "We tested three different perplexity values of 5, 50 and 100. We pre-process the molecular fingerprints by performing a principal component analysis and select the 50 first components. **The t-SNE clusters depend on a perplexity hyperparameter which in brief balances the preservation of global and local aspects during projection from high to low-dimensional space.**".

   We consider the sentence emphasized in bold font to qualify as a short explanation of the meaning of perplexity. We now moved this explanation to earlier in the same paragraph for clarity:

   "t-SNE is an unsupervised machine learning method that embeds high-dimensional data into lower dimensions while preserving distances from the higher-dimensional space. The low dimensional embedding can be used to draw qualitative conclusions about data structure and similarity. The t-SNE clusters depend on a perplexity hyperparameter which in brief balances the preservation of global and local aspects during projection from high to low-dimensional space. We tested three different perplexity values of 5, 50 and 100. We pre-process the molecular fingerprints by performing a principal component analysis and select the 50 first components. Thereafter, we run the t-SNE clustering with random initialization for a maximum of 5000 iterations."

10. *Figures 4 and 5: Can you specify what the lines are in Figures 4b and 5b? Is the interval showing the range of ratios in the whole dataset? If yes, is the marker then the median? To my eye the markers seem to hit the center of the lines in all cases. Also, there are molecules in Gecko that have fewer O than C, right? If the lines are for the ranges, the O:C for Gecko seems off.*

    We thank the reviewer for asking for this clarification. Figures 4b and 5b show the mean value of the atomic ratios, and the lines correspond to the standard deviation. We have now added this explanation to the figure captions:

"Molecular structure analysis of the atmospheric molecules in terms of molecular size as represented by non-hydrogen atom count (a, histogram normalized so bar heights sum to one), mean and standard deviation of atomic ratios (b), and functional groups (c, present in $\geq$ 1 % of dataset)."

and

"Molecular structure analysis of the non-atmospheric molecules in terms of molecular size as represented by non-hydrogen atom count (a, histogram normalized so bar heights sum to one), mean and standard deviation of atomic ratios (b), and functional groups (c, present in $\geq$ 1 % of dataset)."

11. *Page 8: "Oxygen-carrying groups like hydroxyls" -> "Oxygen-carrying groups such as hydroxyls"*

    We have made the requested modification and the text now reads: "Oxygen-carrying groups such as hydroxyls, carbonyls, esters, and ethers appear in both atmospheric and non-atmospheric datasets (Panel c)."

12. *Page 8: "Functional groups like peroxides" -> "Functional groups such as peroxides"*

    We have made the requested modification and the text now reads: "Functional groups such as peroxides and nitrates are less prevalent in non-atmospheric than in atmospheric compounds."

13. *Figure 8: Can you comment on why the Tanimoto similarity distributions with the MACCS fingerprint are so much less smooth compared to the topological fingerprint? Is it related to the size of the fingerprint? And would a larger bin size in these histograms be more convenient? I assume that the "noise" in the distributions doesn't really give any important information about the similarities.*

    The differences in smoothness between the MACCS and topological fingerprint Tanimoto distributions stem from the fingerprint size and level of detail each fingerprint captures. The MACCS fingerprint, at only 166 bits, captures fewer structural features than the larger topological fingerprint, resulting in a "stepped" distribution with fewer possible similarity values. A larger bin size could indeed reduce this "noise" in the MACCS distribution, improving visual clarity without altering the interpretive value. However, we opt to keep the same bin size for both fingerprints for a clear comparison of the two types of similarity distributions.

14. *Page 12: "Our comparison of nitrogen-containing functional groups instead revealed a lack in amine and amide content in atmospheric compounds compared to the other compound classes." In datasets of atmospheric compounds compared to the other datasets, right? Now it sounds like there aren't amines and amides in the atmosphere.*

    That is an important distinction! The following sentences of the main text explain that we believe this to be an artifact of how these datasets were generated. We agree with the reviewer that a further clarification is needed. We have modified the sentence and the full paragraph now reads:
    "Our comparison of nitrogen-containing functional groups instead revealed a lack in amine and amide content in our atmospheric compound datasets compared to the other compound classes. We note that the atmosphere is known to contain numerous reduced nitrogen compounds (estimated to be at least hundreds (Ge et al., 2011). Yet, these compounds are typically presumed to quickly combine with acidic molecules or clusters to form aerosol particles in the atmosphere. Consequently, they are generally excluded

from gas-phase oxidation reactions in simulation models such as MCM or Gecko-A, which explains their absence in our study. These artificial biases in the computational generation of atmospheric compounds necessitate scrutiny and awareness when curating atmospheric datasets and developing models based on such datasets depending on application area."

15. *Page 13: "Furthermore, the similarity between molecular representations like fingerprints can unveil" -> "such as"*

    We have now made the requested change: "Furthermore, the similarity between molecular representations such as fingerprints can unveil whether compounds bear similarity to a machine learning model that utilizes such representations for molecular predictions."

16. *Page 15: "which can be characterized by properties like" -> "such as"*

    We have now made the requested change: "For instance, atmospheric particle formation involves compounds with low volatility, which can be characterized by properties such as extremely low vapor pressures."

17. *Figure 9: Add reference to GeckoQ. Also, use SI units instead of mbar in the x-axis label.*

    We have now changed the unit of the x axis to Pa. We assume the reviewer is asking for the reference to GeckoQ in the caption. The updated caption to Fig. 9 now reads: "Computationally predicted saturation vapor pressure of atmospheric compounds in a subset of the Gecko dataset studied here called GeckoQ (Besel et al., 2023) at 298 K (blue), and vapor pressures listed in the CRC Handbook of Chemistry and Physics (Rumble, 2023) Table entitled 'Vapor Pressure for Inorganic and Organic Substances at Various Temperatures' computed at 298 K using the Clausius-Clapeyron equation. "

18. *Page 16: "assessing not only the overlap of target values, but also to carefully examining" -> "not only assessing the overlap of target values, but also carefully examining"*

    We have now made the requested change: "Moreover, not only assessing the overlap of target values, but also carefully examining the target data type is crucial."

Response to comments by Prof. Jonas Elm

Sandström and Rinke investigate how closely atmospheric organic molecules resemble data in existing curated databases for machine learning (ML) applications. In particular they study the atmospheric Gecko dataset, the Wang dataset based on the master chemical mechanism (MCM) and a dataset consisting of quinones. These are compared to themselves and to the well-known QM9 dataset, as well as nablaDFT and MONA. To estimate the similarity between the datasets the authors apply a supervised ML method in the form of the Tanimoto index and an unsupervised ML method in the form of t-SNE clustering. Two different molecular representations are tested: The topological fingerprint and the MACCS fingerprint.
It is found that existing databases do not cover atmospheric organic molecules well. While this to some extent is no surprise, as these datasets are curated for vastly different purposes, it highlights the need for assembling specialized atmospheric databases in the future. Overall, this is very interesting work, that build upon the existing machine learning development in aerosol science and the conclusion that more specialized atmospheric datasets are needed is a welcoming appeal to the community.
The work is meticulously carried out, the manuscript is well-written and easy to follow. Overall, the work fits well in Geophysical Model Development, and I am happy to recommend the manuscript for publication, essentially as is. I only have a few minor comments. I emphasize

that these are not demands and the authors are free to dismiss the requests if they deem it necessary.

1. *Page 6: "We interpret our results by introducing a high and low similarity reference values. This choice is motivated by previous studies of Tanimoto similarity (Liu et al., 2018; Moret et al., 2023)."*
   I do not really have a gut feeling for the Tanimoto similarity values chosen as not similar (less than 0.1) and similar (0.4 or above). The authors mention that 0.4 or above has been shown to improve ML model performance. Can this value be quantified somehow in the form of the molecular structures? I.e. how similar/dissimilar should the structures be for these cut-off values? For instance, a simple example of some structures that corresponds to the different values would be helpful.

   This is an interesting question. We can provide examples of molecule pairs with similarities close to 0.1 or 0.4. However, we note that this is dependent on the molecular descriptors used. We have now added examples from our comparisons in Appendix B, Figure B1 for both the topological and MACCS fingerprints. We also added a reference to this figure in the beginning of section 3.2.2: "Figure B1 in Appendix B shows examples of molecules with similarities at these reference values."

2. *Page 6-7: "Both fingerprints have been used in atmospheric chemistry machine earning applications (Lumiaro et al., 2021; Besel et al., 2023, 2024) and are therefore pertinent for our comparison."*
   Figure 6 shows the difference between the two chosen representations. As both of the applied descriptors are fingerprints, it interesting to have perform similar analysis based on another descriptor with different architecture. In quantum chemical ML applications there are many possibilities such as coulomb matrix, SOAP, MBTR, FCHL, ect. Hence, could the authors speculate on how sensitive the similarity analysis is to the choice of descriptor architecture?

   This is an important point, connecting to the previous comments on how similarity values depend on descriptor type, size, and design. Here, we used Tanimoto similarity, which is well-suited for binary fingerprints. For continuous descriptors—such as Coulomb matrix, SOAP, MBTR, or FCHL—alternative similarity measures are common, like cosine similarity, dot product, or distance metrics (e.g., Euclidean or Wasserstein distance).

   Continuous descriptors could reveal different aspects of molecular similarity: while binary fingerprints capture certain structural similarity, it is often based on the two-dimensional molecular structure. On the other hand, continuous descriptors capture finer details and 3D properties, such as conformational differences. However, it is challenging to predict precisely how similarity will vary between different descriptors and metrics without performing the actual evaluations.

   The choice of descriptor thus shapes which similarity features are highlighted and which metrics are suitable. The relevance of a similarity measure or similarity of a certain descriptor depends on the specific machine learning applications. Here, we focused on binary fingerprints as they have been used for property prediction.

3. *Page 12: "Our comparison of nitrogen-containing functional groups instead revealed a lack in amine and amide content in atmospheric compounds compared to the other compound classes."*

This is simply a fact of the Gecko, Wang and Quinone datasets not including such compounds. Perhaps, further stress that this indicates that such species should be present in atmospheric databases to have a versatile and representative atmospheric dataset.

Yes, we agree with you. In response to this comment and that of Reviewer # 2 we have now modified this sentence to read:

"Our comparison of nitrogen-containing functional groups instead revealed a lack in amine and amide content in our atmospheric compound datasets compared to the other compound classes. We note that the atmosphere is known to contain numerous reduced nitrogen compounds (estimated to be at least hundreds (Ge et al., 2011). Yet, these compounds are typically presumed to quickly combine with acidic molecules or clusters to form aerosol particles in the atmosphere. Consequently, they are generally excluded from gas-phase oxidation reactions in simulation models such as MCM or Gecko-A, which explains their absence in our study. These artificial biases in the computational generation of atmospheric compounds necessitate scrutiny and awareness when curating atmospheric datasets and developing models based on such datasets depending on application area."

4. *Page 14: "In Figures 7 and 8, we observed that Gecko molecules exhibit greater similarity to each other, while the Wang compounds are more diverse."*

   Is this not simply related to the relative size of the two datasets? In addition, too many similar structures in the dataset just leads to redundant structures and essentially overtraining on specific molecular features. Would a cleaned-up version of the Gecko dataset, where structurally too similar molecules are removed, be a better fit for future training of ML models?

   We will respond in a similar manner as to a previous and similar comment by Reviewer #2.

   While dataset size can influence similarity patterns in dataset comparisons. We believe that for the Gecko and Wang datasets, precursor diversity is actually more important. In atmospheric chemistry, different precursors lead to distinct oxygenated products, capturing the chemical diversity of real-world processes. The Gecko dataset, although larger, was derived from only three precursors, limiting its structural diversity. In contrast, the smaller Wang dataset was generated from 143 precursors, naturally yielding greater molecular variation.

   In theory, a dataset as diverse as Wang and as large as Gecko could be constructed by using an even larger and more varied set of precursors (e.g., 1,500), assuming different precursors produce distinct reaction products. This diversity would better support model generalization without overfitting, as it would represent a broader range of molecular features.

   Therefore, while filtering out structurally similar molecules from Gecko could reduce redundancy, the dataset's current diversity limitations stem more from its precursor sources than from its size alone. Expanding precursor diversity would likely yield a more effective dataset for machine learning in atmospheric chemistry while maintaining a large dataset size.

5. *Page 17: "Examples of such initiatives have recently been developed, such as the Aerosolomics project (Thoma et al., 2022)." Perhaps explicitly specify that you are referring to experimental initiatives here. I would argue that our Atmospheric Cluster DataBase (ACDB)*

comprising the Clusteromics I-V and Clusterome datasets serve a similar purpose, but from a computational point of view.

That is a good point. We have now added references to the Clusteromics, Clusterome and also Goldstein libraries.

"Finally, our study underscores that focus should be given to initiatives aimed at sharing atmospheric molecular data in openly accessible repositories. Examples of such initiatives have recently been developed, such as the Clusteromics I-V and Clusterome datasets (Elm, 2021b, a, 2022; Knattrup and Elm, 2022; Knattrup et al., 2023; Ayoubi et al., 2023), the Aerosolomics project (Thoma et al., 2022) and repositories at University of California, Berkeley (Goldstein, 2024)."

Response to comments by Chief Editor

Dear authors,

Unfortunately, after checking your manuscript, it has come to our attention that it does not comply with our "Code and Data Policy".

`https://www.geoscientific-model-development.net/policies/code_and_data_policy.html`

You have archived your code on a Git repository. However, Git repositories are not suitable for scientific publication. This flaw in your manuscript was already pointed out by the Topical Editor when you submitted your manuscript and before the Discussions stage. Despite it, you have failed to address and solve the issue, which is specially disappointing.

Therefore, you must publish your code in one of the appropriate repositories and reply to this comment with the relevant information (link and a permanent identifier for it (e.g. DOI)) as soon as possible, as we can not accept manuscripts in Discussions that do not comply with our policy. Therefore, the current situation with your manuscript is irregular. Also, please include the relevant primary input/output data.

In this way, if you do not fix this problem in a prompt manner, we will have to reject your manuscript for publication in our journal. Therefore, please, I request you to reply to this comment before the end of the Discussions period with the information (link and DOI) for the new repository that complies with the policy.

Also, in the git repository no license is listed. If you do not include a license the code remains your property and nobody can use it. Therefore, when uploading the model's code to the new repository, you could want to choose a free software/open-source (FLOSS) license. We recommend the GPLv3. You simply need to include the file 'https://www.gnu.org/licenses/gpl-3.0.txt' as LICENSE.txt with your code. Also, you can choose other options that Zenodo provides: GPLv2, Apache License, MIT License, etc.

Also, you must include the modified 'Code and Data Availability' section in a potentially reviewed manuscript, the DOI of the code.

We have now uploaded the versions of the code used in our manuscript to Zenodo, and you can access them at the following links:

1. Code Repository: `DOI:10.5281/zenodo.14224079`,
   `https://zenodo.org/records/14224079`
2. Additional Code: `DOI:10.5281/zenodo.14007835` ,
   `https://zenodo.org/records/14007835`

Furthermore, the datasets utilized for our analysis are freely available, as detailed in our README.md file at this link: DOI:10.5281/zenodo.14224079 (https://zenodo.org/records/14224079).

We have also ensured that a proper license is included with the uploaded code, specifically the GPLv3 license, as recommended.

Finally, we have updated the 'Code and Data Availability' section in the manuscript to include the DOI of the newly uploaded code and any other necessary details:

"The datasets used for this analysis are all freely available from original publications or the database website (see Table 1). Code and workflow used for data analysis are freely available in Zenodo. Code repository: DOI:10.5281/zenodo.14224079. Additional Code: DOI:10.5281/zenodo.14007835."