Response to Reviewers

We sincerely thank the Editor and the reviewers for their valuable time and insightful comments on our manuscript. We appreciate the constructive feedback, which has helped us significantly improve the quality and clarity of the paper. All comments have been carefully addressed, and the manuscript has been revised accordingly. Below, we provide a point-by-point response to each comment raised by the reviewers. All changes in the manuscript are highlighted in yellow color.

Reviewer 1

Comment: Many thanks for your comprehensive reply. I appreciate the time and effort that has been dedicated to both the original paper and the response. The evidence presented in the response does demonstrate the significance of VaPOrS, in contrast to my evaluation in my original review. I will contact the editor to ask whether another review, in light of the response, is allowed/wanted. If permitted, I will suggest the supplied response be included as a major revision to the original paper.

Response: *We sincerely thank you for your constructive follow-up and for reconsidering your initial evaluation in light of our response. We greatly value the time and effort you have devoted to reviewing our work and for acknowledging the significance of VaPOrS in the revised context. Your thoughtful feedback has been instrumental in strengthening the manuscript, and we are grateful for your support in moving this work forward. We included the supplied response as a new subsection of the 'Results and discussion' (see subsection 4.3) in the revised manuscript. Also, we allocated the last paragraphs of the 'Introduction' section to this matter, as follows:*

Despite the usefulness of existing tools such as UManSysProp (Topping et al. 2016) for vapor pressure estimation, they show limitations in correctly identifying functional groups in a range of organic molecules. These shortcomings can lead to significant deviations in predicted vapor pressures, particularly for multifunctional species relevant to atmospheric oxidation and SOA formation. Our observations of such discrepancies provided the main motivation for this work. To address these deficiencies, we developed a Python-based computational framework named VaPOrS (**Va**por **P**ressure in **Or**ganics via **S**MILES) to process SMILES (Simplified Molecular Input Line Entry System) notation of VOCs, automatically identify functional groups, and apply group-contribution methods for property estimation.

The core innovation of VaPOrS lies in its self-contained SMILES parsing and group recognition algorithm, which eliminates reliance on external cheminformatics libraries such as OpenBabel. Instead of depending on SMARTS-based pattern matching, VaPOrS explicitly searches for all possible patterns of each functional group directly from the SMILES string, ensuring full control over the detection logic. This approach makes the tool both transparent and adaptable, enabling straightforward extension to new group definitions and methods without external dependencies. In its current version, VaPOrS implements the SIMPOL method by detecting 30 functional groups required for estimating saturation vapor pressure and enthalpy of vaporization. However, this is only the first demonstration of the framework: the same group recognition functions can be applied to other parameterization schemes (e.g., group additivity, volatility basis set (VBS) models, partition coefficients, Henry's law constants), making VaPOrS a general platform for group-contribution modeling rather than a tool restricted to vapor pressure prediction. Therefore, the development of VaPOrS addresses several challenges:

1. Automated and auditable functional group detection: Eliminating manual identification, reducing error potential, and providing full transparency in detection logic.
2. Computational efficiency: By bypassing Pybel's SMILES-to-graph conversions, VaPOrS reduces overhead, enabling much faster execution for large-scale atmospheric simulations involving thousands of compounds across many steps and grid cells.
3. Scalability and flexibility: Capable of processing thousands of SMILES strings within seconds, with design features that make it portable to high-performance computing (HPC) environments and easily translatable to compiled languages such as Fortran for integration into large-scale chemical transport and climate models.

Reviewer 2

This manuscript introduces VaPOrS v1.0.1, a Python-based tool developed to estimate saturation vapor pressure and enthalpy of vaporization for organic compounds using their SMILES representations. A key feature of VaPOrS is its built-in capability to detect functional groups directly from SMILES strings, eliminating the need for external cheminformatics libraries and manual SMARTS definitions. The tool relies on the SIMPOL group contribution method developed by Pankow and Asher for its vapor pressure predictions and, at the moment is able to recognize only the 30 functional groups needed to apply this method.

The authors validated VaPOrS against the original SIMPOL dataset and demonstrated perfect agreement between the two approaches. Further testing on an external dataset (i.e., MCM database) showed strong correlation with manually derived SIMPOL predictions and other established models such as EVAPORATION and Nanoolal. The methodology is sound, and the tool appears robust and computationally efficient and has potential to be further expanded.

I recommend the manuscript for publication after revisions are made to address the concerns outlined below.

Comment: I think the main contribution of this work is the development of the functions to detect functional groups. This is done in a efficient way directly in the tool without relying on external libraries. I know how frustrating can be installing and setting up dependencies between different libraries and tools and I value a self-contained tool that can be adapted to include different methods. So, I think the strongest point of the paper is the SMILES parser and groups identification. The authors present a first implementation of the SIMPOL methods and highlight that the tool can be expanded to include more group contribution methods to predict saturation vapour pressure. However, this feels somewhat restrictive, as the group recognition framework developed in VaPOrS is broadly applicable and could be adapted to predict a wider range of physicochemical properties (e.g., partition coefficients) and not only vapor pressure. I think this point should be stressed more in the manuscript and the tool should be presented as a general tool for SMILES parsing and group contribution method application. Conversely the authors mainly focus on describing the SIMPOL implementation for the prediction of VP. This is an established method developed by other scientists. At a first reading it appears the VaPOrS just apply the SIMPOL method without apporting any contribution, thus I understand the comments of the first reviewer. The main contribution of the paper is the automatization of the fragments recognition in an efficient way and I think this should be stressed more.

Related to this, since the real novelty are the SMILES parsing functions, I think a substantial validation of the group recognition method is missing in the paper. Section 3.4.1 (MCM data) briefly describes as the SMILES parsing functions have been tested on 126 external compounds. I think this should be one of the main sections of the paper demonstrating that the functions are able to correctly recognize the functional groups needed by SIMPOL (or any other group contribution method implemented) on an external dataset. The authors have provided supplementary material in response to a previous reviewer's comment, comparing their approach with UManSysProp and highlighting cases where UManSysProp fails to correctly identify certain groups, leading to inaccurate predictions. This comparison is highly relevant and should be integrated into the main text to underscore the robustness and reliability of VaPOrS. The authors criticize the SMART pattern recognition in OpenBabel, so a comparison between fragments identified by VaPOrS and fragments identified by OpenBabel should be included to highlight the strength of VaPOrS related to OpenBabel and justify the development of ad-hoc functions in a new method.

Response: *We sincerely thank the reviewer for the constructive and thoughtful comments, which helped us to better highlight the true novelty and contributions of our work. We agree with the reviewer that the core innovation of our study lies in the development of efficient, self-contained functions for functional group recognition directly from SMILES, without relying on external libraries. We revised the manuscript to emphasize this point more clearly. In particular, we framed VaPOrS not only as a tool implementing the SIMPOL method, but more broadly as a general framework for SMILES parsing and group contribution method applications, adaptable to the prediction of a wider range of physicochemical properties such as partition coefficients, Henry's law constants, and activity coefficients. We acknowledge that in the current version, the manuscript places more focus on the SIMPOL implementation, which may give the impression that the work simply replicates an established method. In the revised version, we shifted the emphasis toward the novelty of the group recognition framework and its versatility, presenting the SIMPOL implementation as a case study to demonstrate the functionality and accuracy of the approach. We addressed this in different sections of the revised manuscripts as follows:*

In the Abstract:

While this first implementation focuses on the SIMPOL method for estimating saturation vapor pressure and enthalpy of vaporization, the framework is readily extendable to other group-contribution schemes and thermodynamic properties (e.g., partition coefficients, volatility basis set models, solubility, Henry's law constants).

In the 'Introduction' section:

Despite the usefulness of existing tools such as UManSysProp (Topping et al. 2016) for vapor pressure estimation, they show limitations in correctly identifying functional groups in a range of organic molecules. These shortcomings can lead to significant deviations in predicted vapor pressures, particularly for multifunctional species relevant to atmospheric oxidation and SOA formation. Our observations of such discrepancies provided the main motivation for this work. To address these deficiencies, we developed a Python-based computational framework named VaPOrS (**Va**por **P**ressure in **Or**ganics via **S**MILES) to process SMILES (Simplified Molecular Input Line Entry System) notation of VOCs, automatically identify functional groups, and apply group-contribution methods for property estimation.

The core innovation of VaPOrS lies in its self-contained SMILES parsing and group recognition algorithm, which eliminates reliance on external cheminformatics libraries such as OpenBabel. Instead of depending on SMARTS-based pattern matching, VaPOrS explicitly searches for all possible patterns of each functional group directly from the SMILES string, ensuring full control over the detection logic. This approach makes the tool both transparent and adaptable, enabling straightforward extension to new group definitions and methods without external dependencies. In its current version, VaPOrS implements the SIMPOL method by detecting 30 functional groups required for estimating saturation vapor pressure and enthalpy of vaporization. However, this is only the first demonstration of the framework: the same group recognition functions can be applied to other parameterization schemes (e.g., group additivity, volatility basis set (VBS) models, partition coefficients, Henry's law constants), making VaPOrS a general platform for group-contribution modeling rather than a tool restricted to vapor pressure prediction. Therefore, the development of VaPOrS addresses several challenges:

1. Automated and auditable functional group detection: Eliminating manual identification, reducing error potential, and providing full transparency in detection logic.
2. Computational efficiency: By bypassing Pybel's SMILES-to-graph conversions, VaPOrS reduces overhead, enabling much faster execution for large-scale atmospheric simulations involving thousands of compounds across many steps and grid cells.
3. Scalability and flexibility: Capable of processing thousands of SMILES strings within seconds, with design features that make it portable to high-performance computing (HPC) environments and easily translatable to compiled languages such as Fortran for integration into large-scale chemical transport and climate models.

In the 'methodology' section:

Once functional groups are identified, VaPOrS converts the detection results into integer values representing the frequency of each group in a given compound. These values are stored in an array and serve as the critical input for property prediction. In the present implementation, saturation vapor pressure is calculated using the SIMPOL group-contribution method, in which the total logarithmic vapor pressure is expressed as the sum of contributions from all relevant functional groups as a function of temperature.

In the 'Conclusion' section:

While the present work focused on SIMPOL, the underlying framework of VaPOrS is flexible and can be extended to other structure-based approaches, including group additivity and volatility basis set (VBS) methods. Its design also makes it readily adaptable for estimating additional thermodynamic properties beyond vapor pressure, broadening its utility in atmospheric and environmental chemistry.

Regarding validation, we fully agree with the reviewer that a substantial demonstration of the robustness of the group recognition method is essential. To address this, we reorganized the sections of the revised manuscript, allocating a main section to validation of VaPOrS, where we evaluated its accuracy in predicting the saturation vapor pressures and enthalpy of vaporization of 224 organic compounds listed in the SIMPOL-pertained study (see section 3). Moreover, we expanded the main sections of the 'Results and discussion', demonstrating the capability of VaPOrS to correctly recognize the functional groups needed by SIMPOL on external datasets provided by MCM and autoAPRAM-fw (see sections 4.1 and 4.2). Finally, we integrated into the main text the comparison with UManSysProp, currently included in the supplementary material, as it clearly illustrates how VaPOrS overcomes the limitations of existing tools, specifically those relying on OpenBabel and SMARTS-based approaches, in group recognition and vapor pressure prediction.

We believe these revisions clarify the originality of the work, underscore the robustness of the developed framework, and address the reviewer's concern that the manuscript currently underemphasizes its main contributions.

Comment: Page 5, line 4, […tools like VaPOrS, enabling…] VaPOrS acronym has not been established yet.

Response: *We thank the reviewer for pointing this out. In the revised manuscript, we have now introduced the full name of the tool at its first occurrence as VaPOrS (Vapor Pressure of Organics via SMILES), after which the acronym is used consistently (see line 12, page 5).*

Comment: Page 7, lines 19-21 [In particular, the SMILES string must begin…], the authors write MUST BEGIN implying that the SMILES need to be provided in a specific way. A SMILES for a chemical can be written in many different variation (e.g, canonical vs kekulized). To be valid, the tool must be able to recognize functional groups even for different variation of the same SMILES. Given that SMILES syntax can vary depending on generation method or canonicalization, it is important to demonstrate that the tool yields consistent fragment counts regardless of input variation. This would strengthen confidence in the robustness of the group recognition algorithm and its suitability for large-scale automated analyses. The manuscript should also clarify whether VaPOrS includes a SMILES standardization step prior to functional group parsing. Standardization is essential to ensure reproducibility in fragment recognition.

Response: *We thank the reviewer for highlighting the importance of SMILES variation and standardization. At the current stage, VaPOrS operates using canonical SMILES as input. This choice was made because canonical SMILES are widely used in chemical databases and ensure a unique representation of each molecule, which simplifies functional group detection and validation. We acknowledge that the lack of an internal standardization step limits flexibility when handling alternative SMILES notations. We plan to extend future versions of VaPOrS to include this functionality. This will require additional development time, as the current algorithm was designed and validated specifically with canonical SMILES. To clarify this point for readers, we have revised the manuscript to explicitly state that VaPOrS presently requires canonical SMILES as input. This has been brought in the 'Methodology' section as follows:*

At the current stage, VaPOrS operates using canonical SMILES as input. This choice was made because canonical SMILES are the most common representation in chemical databases and ensure a unique representation of each molecule, which simplifies functional group detection and validation. Although alternative SMILES forms (e.g., kekulized or non-canonical variations) can represent the same molecule, these are not yet supported in the present version. To maintain consistency, users should therefore provide canonical SMILES when running VaPOrS.

Comment: Furthermore, the manuscript should address how VaPOrS handles tautomeric variability in SMILES representations. Tautomers are chemically equivalent but structurally distinct forms of the same chemical that can be encoded differently. This variability can significantly impact functional group recognition and, consequently, the accuracy of property predictions. It is unclear whether the authors have tested the tool's consistency across different tautomeric forms of the same compound. I recommend including a discussion on this issue and, if not already performed, conducting a validation study to assess whether VaPOrS yields consistent fragment counts and predictions across tautomeric variants.

Response: *We thank the reviewer for raising the issue of tautomeric variability in SMILES representations. As noted, tautomerism can lead to structurally distinct encodings of the same compound, which in turn may alter the number and type of functional groups identified. VaPOrS is designed to parse SMILES and count these functional groups exactly as they are presented. Therefore, any variability in the predicted properties between tautomeric SMILES does not arise from VaPOrS itself, but rather from the group-contribution framework applied (in this case, SIMPOL). Since functional group–based methods such as SIMPOL assign unique parameters to each group (e.g., C=C, OH, and C=O), different tautomeric inputs naturally yield different parameterizations. Currently, SIMPOL and similar models do not provide distinct parameter sets for tautomeric forms, leaving this ambiguity unresolved. Should future studies propose a robust strategy, such as allocating distinct parameters for tautomeric groups, this could be readily implemented in VaPOrS. For now, VaPOrS follows the input SMILES representation provided by the user or database and detects C=C, OH, and C=O as distinct functional groups, while acknowledging that tautomerism is a limitation of group-contribution methods rather than of VaPOrS itself. We have now added a detailed discussion of tautomerism as a limitation of group-contribution approaches such as SIMPOL in the revised manuscript. This has been brought in the 'Results and discussion' section as follows:*

Another important limitation relates to tautomerism, in which a compound can exist in multiple chemically equivalent but structurally distinct forms, for example keto-enol tautomerism. These forms contain different functional groups. For instance, a keto tautomer may contain a carbonyl group (C=O), while its enol counterpart may contain one hydroxyl (OH) and one carbon-carbon double bond (C=C). Since functional group-based methods like SIMPOL assign unique parameters to each group, different tautomeric SMILES representations of the same compound can yield different vapor pressure predictions, despite the compound having a single experimentally measurable vapor pressure. VaPOrS does not attempt to canonicalize or normalize tautomeric forms; it faithfully parses and counts groups in the SMILES provided by the user or database. Thus, this ambiguity arises from the group-contribution framework itself rather than from VaPOrS. In practice, most studies adopt a convention of selecting the thermodynamically more stable tautomer, commonly the keto form in the gas phase, as the reference structure, although this is not universally standardized. Should future investigations establish robust strategies for handling tautomerism, for example by developing distinct parameters for different tautomeric states, such improvements could be readily integrated into VaPOrS. Addressing these limitations, whether through correction terms, hybridization with other frameworks, or leveraging data-driven approaches such as machine learning, represents an important avenue for future development.

Comment: Page 30, lines 14-17 [Many data points are clustered close to this line…], this is subjective comments. A more objective description would consider some metric like the R2 or the RMSE. Please provide some quantitative metrics to describe your correlation.

Response: *We appreciate the reviewer's suggestion. In the revised manuscript, we have replaced the subjective description with quantitative performance metrics (RMSE and R^2) to objectively evaluate the accuracy of the VaPOrS predictions. Specifically, for log P at 333.15 K (Figure 4), we report RMSE = 0.4232 and R^2 = 0.9648. Across six temperatures (Figure 5), the results yield RMSE = 0.5556 and R^2 = 0.9570. For vaporization enthalpies at 333.15 K (Figure 6), the comparison gives RMSE = 14.4740 and R^2 = 0.6146. These values are now included in the Results section to provide a clear, quantitative assessment of model performance.*

Comment: Page 32, lines 6-10 [These discrepancies could be attributed to the structural complexity…] this paragraph concern the applicability domain of the model. I know the SIMPOL model has not been developed by these authors, but could VaPOrS provide an applicability domain? Maybe something related to the groups count? For instance, does the presence of certain group together of the presence of too many instance of the same fragment result in more uncertain prediction?

Response: *We thank the reviewer for this insightful comment regarding the applicability domain of the SIMPOL model and the potential role of VaPOrS in this respect. While it is true* *that SIMPOL was not originally designed with an explicit applicability domain, the strength of VaPOrS lies in its ability to decompose molecular structures into functional groups and quantify them systematically. This feature could indeed be leveraged to flag molecules where the presence of multiple instances of the same fragment or unusual combinations of groups may result in higher prediction uncertainty. For example, compounds containing a large number of hydroxyl or carboxyl groups, or rare functionalities such as peroxides, tend to show greater discrepancies from experimental values. We have added a discussion in the manuscript to clarify that although VaPOrS itself does not yet provide a formal applicability domain, it could serve as a diagnostic tool by identifying structural features that are likely to fall outside the reliable range of SIMPOL predictions. This has been brought in the 'Results and discussion' section as follows:*

At the same time, it is important to recognize the limitations of group contribution approaches such as SIMPOL, which underpins this first implementation of VaPOrS. While reliable for many compounds, predictive accuracy declines as molecular complexity increases. Highly functionalized molecules may exhibit non-additive effects, such as steric hindrance, intramolecular hydrogen bonding, or electronic interactions, that are not captured by simple group summation rules. Previous studies have shown that such effects can dampen or amplify volatility changes in ways not accounted for by group contribution approaches.

.

.

.

Addressing these limitations, whether through correction terms, hybridization with other frameworks, or leveraging data-driven approaches such as machine learning, represents an important avenue for future development.

Comment: Figure 7, Antoine and SIMPOL methods seem to give a good agreement. However, there are instances where the two methods seem far from the experimental line (Decanedioic acid, Hexanamide, Diethyl-peroxide). Please comment.

Response: *We thank the reviewer for this valuable observation. We agree that while both the Antoine and SIMPOL methods generally show good agreement with experimental data, there* *are notable deviations for certain compounds. As shown in Figure 7, examples include Decanedioic acid, which contains two carboxyl groups, Hexanamide, where strong hydrogen-bonding interactions are important, and Diethyl-peroxide, which contains an uncommon peroxide group. These structural features make intramolecular interactions more significant and less amenable to simple group additivity approaches, which likely explains the deviations. Such cases highlight the limitations of SIMPOL method when applied to structurally complex or less common functionalities, and it also points to the potential of VaPOrS to help identify molecules that fall outside the reliable applicability domain of the SIMPOL framework. We added a discussion on this deviation in the manuscript, as follows:*

Although the SIMPOL model does not explicitly define an applicability domain, the structure-based framework of VaPOrS offers an opportunity to explore this aspect. Since the code identifies and counts functional groups for each molecule, it can be used to highlight cases where predictions may be less reliable. For instance, compounds such as Decanedioic acid, which contains two carboxyl groups, or Hexanamide, where strong hydrogen-bonding interactions may play a role, or Diethyl-peroxide, which contains a relatively uncommon peroxide group show noticeable deviations from experimental values (See Figure 7). These examples suggest that compounds with an unusually high number of hydroxyl or carboxyl groups, or those containing less common fragments such as peroxides, often exhibit larger deviations from experimental vapor pressures. Likewise, the co-occurrence of multiple reactive groups within the same molecule (e.g., carbonyl–peroxide combinations) may introduce additional uncertainties. While a systematic applicability domain analysis is beyond the scope of this study, we note that VaPOrS could be extended to provide such functionality, thereby guiding users in assessing the reliability of SIMPOL predictions for structurally complex molecules.

Comment: Figure 11, 3D graphs look cool on computer screen in interactive applications. When on paper are kind of hard to read. For example, I cannot see the depth on one of the axis. I see that the information on the Mass can be interesting, perhaps a 2D correlation between Mass and the groups count would be better for a printed version of the paper.

Response: *We appreciate your comment regarding Figure 11. We agree that 3D plots can sometimes be challenging to interpret in printed form. However, we believe the 3D* *representation provides valuable information by simultaneously illustrating the relationships among Mass, functional group count, and comparing vapor pressures, which would be difficult to capture in a purely 2D format. To enhance clarity and improve readability, we have therefore kept the 3D plot but also added an additional 2D plot (see Figure 12) showing the correlation between Mass and important features, as you suggested. This provides a clearer representation in print while retaining the multidimensional perspective of the original figure.*

We revised the manuscript accordingly. Using VaPOrS, the benzene-derived oxidation and autoxidation products were further classified into volatility categories based on their effective saturation mass concentration (C, µg m*$^{-3}$*) in the newly added Figure 12. These include ultra-low-volatility organic compounds (ULVOC,* $C^* \leq 3\times10^{-9}$ *µg m*$^{-3}$*), extremely low-volatility organic compounds (ELVOC,* $3\times10^{-9} < C^* \leq 3\times10^{-5}$ *µg m*$^{-3}$*), low-volatility organic compounds (LVOC,* $3\times10^{-5} < C^* \leq 3\times10^{-1}$ *µg m*$^{-3}$*), semi-volatile organic compounds (SVOC,* $3\times10^{-1} < C^* \leq 3\times10^{2}$ *µg m*$^{-3}$*), and intermediate-volatility organic compounds (IVOC,* $3\times10^{2} < C^* \leq 3\times10^{6}$ *µg m*$^{-3}$*). In this figure, we present the 2D relationship between molar mass and effective saturation concentration for these products.*