https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

VaPOrS v1.0.1: An automated model for estimating vapor pressure of organic compounds using SMILES notation

Mojtaba Bezaatpour[1,a], Miikka Dal Maso [a], Matti Rissanen[2, a, b]

[a] *Aerosol Physics Laboratory, Tampere University, 33720 Tampere, Finland*

[b] *Department of Chemistry, University of Helsinki, 00014 Helsinki, Finland*

Abstract

Volatile organic compounds play a significant role in atmospheric chemistry, influencing air quality and climate change. Accurate prediction of their physical properties is essential for understanding their behavior. This paper introduces the VaPOrS (**Va**por **P**ressure in **Or**ganics via **S**MILES) as a comprehensive tool designed to process SMILES notation of organic compounds, identify key functional groups, and calculate their saturation vapor pressure and enthalpy of vaporization at any specified temperature. While this first study focuses on applying the SIMPOL method for parameterization, VaPOrS is inherently adaptable to other structure-based parameterization approaches, such as group additivity and volatility basis set (VBS) methods by extracting substructure information from each string that is meaningful to property predictive techniques. It can also be extended to any thermodynamic property that relies on structural group-based parameterizations. In its current version, the tool automates the detection of 30 critical structural groups and has been validated against manually counted functional groups and experimental saturation vapor pressure data for a diverse set of compounds. The results demonstrate high accuracy, with the tool correctly identifying the same functional groups, followed by providing prompt saturation vapor pressure predictions according to the SIMPOL parameterization. The developed method can be integrated into large-scale simulation models targeting secondary aerosol formation and involving thousands of organic species at once. Thus, the developed tool offers a robust computational approach for research in atmospheric chemistry and environmental science, allowing to streamline the

[1] Corresponding author email address: mojtaba.bezaatpour@tuni.fi (M. Bezaatpour)
[2] Corresponding author email address: matti.rissanen@tuni.fi (M. Rissanen)

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

analysis of a large collection of organic compounds, aiding in the assessment of their climatic impacts.

Keywords: Saturation vapor pressure, Volatile organic compounds, VaPOrS, SMILES notation, Functional group, Atmospheric chemistry, Secondary organic aerosol.

1. Introduction

Volatile organic compounds (VOCs) are a diverse group of organic chemicals that significantly impact atmospheric chemistry. Their volatility allows them to easily enter the atmosphere, where they participate in complex chemical reactions that influence air quality, climate, and human health (Mellouki, Wallington, and Chen 2015). VOCs are key precursors to both ground-level ozone and secondary organic aerosols (SOA). Ground-level ozone, a harmful air pollutant, forms through the photochemical reactions of VOCs with nitrogen oxides (NOx) in the presence of sunlight (Atkinson 2000). Ozone is a major component of urban smog and poses serious health risks, including respiratory problems and cardiovascular disease (WHO 2005). Secondary organic aerosols, on the other hand, result from the oxidation of VOCs, leading to the formation of particulate matter that can scatter sunlight and affect the Earth's radiative balance (Jimenez et al. 2009). These particles contribute to atmospheric haze, impact visibility, and have been linked to adverse health effects, including lung and heart diseases. Understanding the formation, transformation, and fate of VOCs is thus crucial for predicting their impact on both air quality and climate.

VOCs originate from a variety of sources, both natural and anthropogenic. Natural sources include biogenic emissions from vegetation, such as isoprene and monoterpenes, which are released in large quantities, especially in forested areas (Alex Guenther et al. 1995). Anthropogenic sources are primarily related to industrial activities, transportation, fuel combustion, and the use of solvents and consumer products (Goldstein and Galbally 2007; McDonald et al. 2018). The structural diversity of VOCs poses a challenge for their classification and analysis. They can range from simple hydrocarbons like methane to complex multifunctional molecules containing oxygen, nitrogen, sulfur, and, for example, halogens. This diversity affects their physical properties, reactivity, and environmental behavior, making it necessary to study their properties comprehensively.

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

The physical properties of VOCs, particularly saturation vapor pressure and enthalpy of vaporization, play a critical role in determining their volatility and phase behavior in the atmosphere. Saturation vapor pressure is a measure of a substance's tendency to evaporate, and it directly influences the distribution of VOCs between the gas and particulate phases (Mattila, Kulmala, and Vesala 1997). Compounds with high saturation vapor pressure are more likely to remain in the gas phase, whereas those with lower saturation vapor pressure can condense onto particulate matter, contributing to SOA formation (Donahue et al. 2006). Enthalpy of vaporization, the energy required to convert a substance from liquid to vapor, is an important property that influences the temperature dependence of saturation vapor pressure. Accurate knowledge of these properties is essential for modeling VOC emissions, their transport, and their transformation in the atmosphere.

The oxidation of VOCs significantly changes their chemical properties. An extreme example is provided by autoxidation, a chain-like oxidation process that propagates by sequential additions of molecular oxygen and resulting peroxy radical rearrangement reactions, which infuses multiple oxygen molecules to the hydrocarbon backbone. This process plummets the saturation vapor pressures of the participating VOC and ultimately forms so-called Highly Oxygenated organic Molecules (HOMs as expressed by Bianchi et al. (Bianchi et al. 2019; Crounse et al. 2013)), crucial intermediates in the formation of SOA. HOM formation is a common property of VOCs and is thus initiated by all common oxidants capable of forming alkyl radicals, including hydroxyl radicals (OH), ozone (O_3), and nitrate radicals (NO_3) (Zhao et al. 2021; Luo et al. 2023; Rissanen et al. 2014; Berndt et al. 2016) as well as by direct UV photolysis. The resulting HOMs can rapidly condense onto existing particles or sometime even form new particles through nucleation processes, thus contributing to the formation and growth of SOA. The formation of HOMs is heavily influenced by the structure of the precursor VOCs (Ehn et al. 2014). Recent advancements in mass spectrometry and atmospheric simulation techniques have provided valuable insights into the chemical pathways leading to HOMs formation and their role in atmospheric chemistry (Iyer 2023; Bianchi et al. 2019; Vereecken et al. 2018; Michael E. Jenkin et al. 2019; Iyer et al. 2021a; 2021b). Understanding the volatility of VOCs, and especially HOMs, is therefore essential for improving models of SOA formation.

Traditionally, the determination of saturation vapor pressure and enthalpy of vaporization has relied on experimental methods, such as gas chromatography-mass spectrometry (GC-MS) (Epping and Koch 2023). While these methods provide accurate measurements, they are time-consuming, expensive, and often impractical for large-scale studies involving thousands of

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

compounds. Moreover, many of the condensable chemicals relevant to atmospheric SOA formation are challenging to measure because their parent molecules are unstable, have never been synthesized, or, importantly, cannot be synthesized due to their inherent instability. These limitations highlight the critical need for approximative methods to estimate the physical properties of these compounds, as direct measurements are often unfeasible.

Given the limitations of experimental approaches, there is an obvious need for computational tools that can predict these properties efficiently. Such tools can complement experimental methods by providing rapid estimates for a wide range of compounds, facilitating large-scale atmospheric modeling studies. Various computational methods have been developed to estimate the saturation vapor pressure of organic compounds, complementing or replacing traditional experimental approaches. Among these, predictive models like COSMO-RS (COnductor-like Screening MOdel for Real Solvents) use quantum chemistry calculations to estimate thermodynamic properties, including saturation vapor pressure. COSMO-RS simulates the solvent environment and molecular interactions, providing a theoretical basis for property estimation. However, its computational intensity can be a drawback for large-scale applications (Klamt 1995; Eckert and Klamt 2002; Klamt et al. 1998). On the other hand, group contribution methods estimate saturation vapor pressure based on the contribution of structural groups within a molecule. These methods leverage empirical correlations derived from extensive datasets but often lack precision for complex or highly functionalized molecules (Joback and Reid 1987; Myrdal and Yalkowsky 1997). The Nannoolal method is a computational approach used to estimate the boiling points and saturation vapor pressures of organic compounds based on their molecular structure. This method utilizes group contribution techniques, where the overall properties of a molecule are determined by summing the contributions of its individual structural groups, such as functional groups and bonding patterns (Nannoolal, Rarey, and Ramjugernath 2008). The EVAPORATION method is a tool designed for estimating the saturation vapor pressure of organic molecules, accounting for contributions from the carbon skeleton, functional groups, and their interactions, and it adjusts for functionalized diacids with empirical modifications. It predicts the saturation vapor pressure of various compounds using only molecular structure as input, making it applicable to a wide range of molecules (Compernolle, Ceulemans, and Müller 2011). The SIMPOL method, developed by Pankow and Asher (Pankow and Asher 2008), is a widely used group contribution method that correlates the presence of specific functional groups in a molecule to its vapor pressure. By summing the contributions of individual functional groups, the method provides

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

an estimate of the compound's vapor pressure. The SIMPOL method has been validated against experimental data and is recognized for its accuracy and applicability to a wide range of organic compounds. Importantly, since all these methods employ group contribution techniques, they could potentially be integrated into automated tools like VaPOrS, enabling the efficient handling of large datasets and complex compounds with minimal manual effort.

The present study introduces a Python-based computational tool named VaPOrS (**Va**por **P**ressure in **Or**ganics via **S**MILES) to process SMILES (Simplified Molecular Input Line Entry System) notation of VOCs, identify key functional groups, and calculate their saturation vapor pressure and enthalpy of vaporization at any specified temperature. A key distinction between VaPOrS and existing tools such as UManSysProp (Topping et al. 2016) lies in their approach to parsing SMILES notation for functional group identification. While both tools employ group contribution methods for property estimation, VaPOrS explicitly searches for all possible patterns of a specific functional group (e.g., aldehyde) directly from the SMILES string, without relying on external libraries. More details on the specific patterns and their implementation are provided in the Methodology section. In contrast, UManSysProp utilizes SMARTS strings within the OpenBabel framework to identify molecular substructures relevant to predictive methods. The advantage of VaPOrS' approach is that it ensures full control over pattern-matching logic, potentially making it more adaptable to new group definitions without dependency on predefined SMARTS rules. Currently, the tool automates the SIMPOL method by detecting 30 functional groups, which are critical for determining the physical properties of VOCs. This initial implementation focuses on the SIMPOL method as a starting point; however, future iterations of VaPOrS aim to include additional parameterization approaches, such as group additivity and volatility basis set (VBS) methods, further expanding its utility and adaptability.

The development of this tool addresses several key challenges:

1. Automation of functional group detection: The tool eliminates the need for manual identification of functional groups, reducing the potential for human error and increasing efficiency.

2. Rapid and accurate property prediction: The tool leverages the SIMPOL method to provide rapid and accurate predictions of saturation vapor pressure and enthalpy of vaporization. This capability is particularly valuable for large-scale atmospheric simulations targeting SOA formation involving thousands of compounds.

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

3. Scalability and flexibility: The tool is designed to handle large datasets, making it suitable for high-throughput studies. It can process thousands of SMILES strings within seconds, providing quick insights into the properties of VOCs.

The computational analysis of VOCs is integral to several established atmospheric databases and models, which simulate the chemical and physical processes in the atmosphere. Incorporating the VaPOrS developed in this study into them can significantly enhance their efficiency and accuracy. As an instance, recently, Pichelstorfer et al. (Pichelstorfer et al. 2024) developed a close-to-mechanistic approach called auto-APRAM-fw for predicting the formation and general structure of HOMs during the autoxidation of initial radicals, outputting the molecular structures in SMILES notation. Integrating this vast array of HOMs structures into atmospheric simulation models presents significant challenges, particularly in identifying the functional groups for each HOM and calculating their saturation vapor pressures. This process is essential, as saturation vapor pressure is a critical parameter in understanding the volatility and partitioning behavior of HOMs in atmospheric conditions, influencing SOA formation and growth. The task, however, is resource-intensive when conducted manually or through non-specialized methods. By automatically detecting functional groups and calculating saturation vapor pressures for the generated SMILES strings, the VaPOrS tool can facilitate the efficient integration of HOMs data from auto-APRAM-fw into atmospheric models. This automation not only accelerates the data processing required for simulations but also improves the accuracy of saturation vapor pressure predictions, thus enhancing the reliability of SOA formation simulations. Consequently, VaPOrS holds significant potential for advancing atmospheric chemistry research, particularly in extending auto-APRAM-fw mechanistic model for broader applications under diverse atmospheric conditions. Some of the key atmospheric databases and models that could benefit from this tool are discussed in the Results and discussion section.

2. Methodology

The developed VaPOrS begins by processing the SMILES notation of the target chemical compounds. SMILES strings are used as inputs to identify the structural components of the molecules, such as number of carbon atoms, functional groups, and cyclic structures. Altogether 30 functional groups parameterized in the SIMPOL method are identified. Several functions are required to efficiently parse and interpret the structure of each compound. The model consists of two main parts. The first part is dedicated to identifying and counting the

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

number of functional groups within the target while the second part focuses on calculating the saturation vapor pressure as a function of temperature and providing the corresponding Antoine equation parameters.

2.1.Functional group identification

The first part of the model identifies and counts 30 distinct structural groups in organic compounds that are recognized by the SIMPOL method as influential on saturation vapor pressure. These groups include functional components like carbonyl groups (-C=O), hydroxyl groups (-OH), amines (-NR_3), ethers (-R-O-R-), and several others (see Figure 1). The model processes the SMILES string of each compound and applies functions to detect the presence and quantity of these groups. To achieve this, the SMILES string is parsed by several dedicated functions, each corresponding to a specific functional group (e.g., O-atom in a carbonyl or hydroxyl). The results are stored as variables that represent the number of occurrences of each group within the molecule. These counts are then used as input parameters in the saturation vapor pressure calculation. This identification process is critical, as the functional groups detected are the most impactful to the molecule's saturation vapor pressure. By systematically identifying the relevant groups for each compound, the model prepares the necessary data for the next phase of thermodynamic calculations. When using SMILES notations for molecular representation, it is important to exercise caution regarding the starting point of the notation. In particular, the SMILES string must begin from an atom that initiates a branching structure (e.g., position 1, 3, or 5 in Figure 1(a)) rather than from a middle atom (e.g., position 2 or 4 in Figure 1(a)). This ensures that the VaPOrS code correctly interprets the molecular topology and identifies functional groups accurately. This method of writing SMILES is the logical and most straightforward approach that is seen in most databases, sources, and online tools. The following subsections 2.1.1 to 2.1.19 provide detailed descriptions of the first steps of the model computations.

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

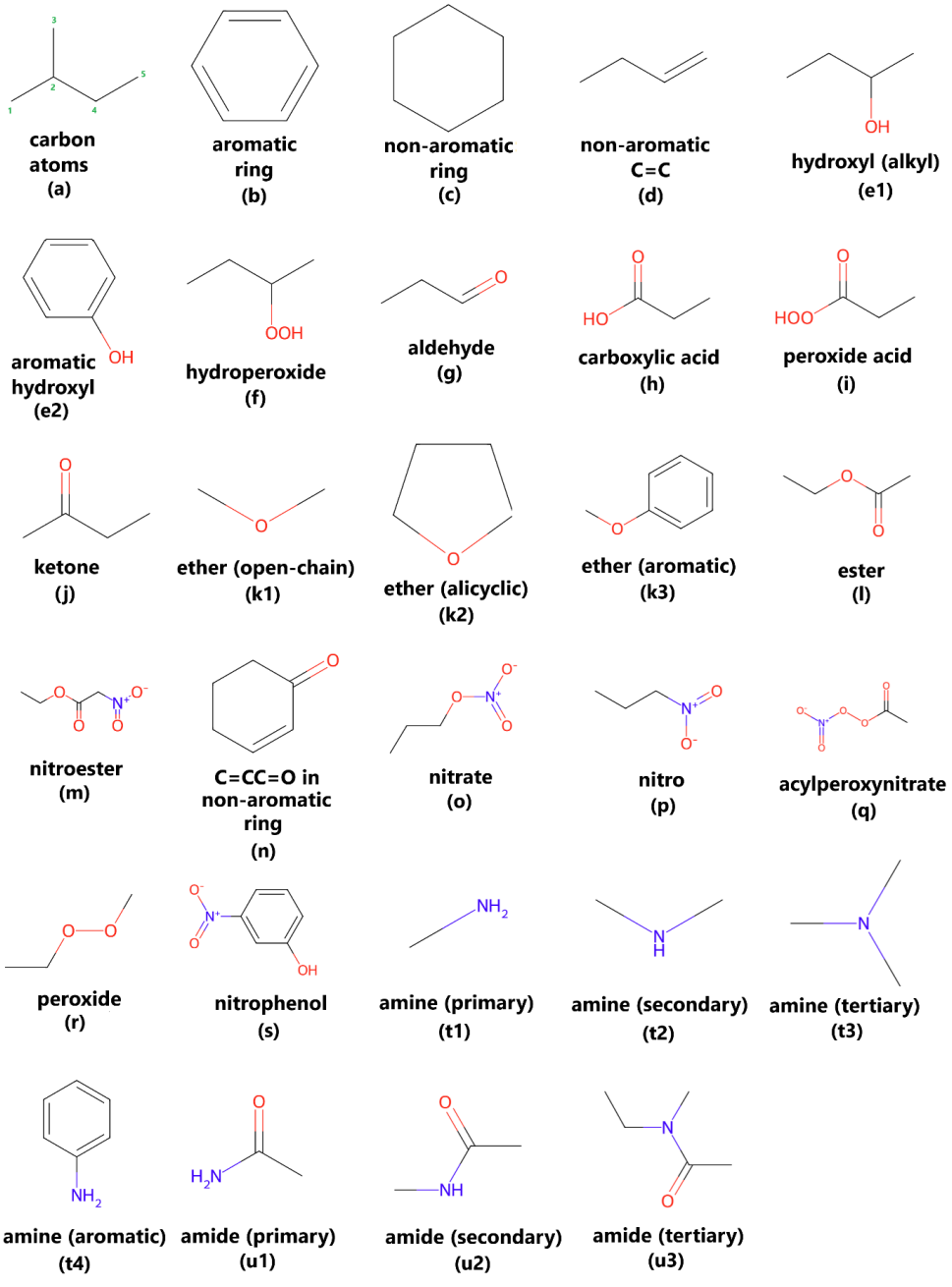


Figure 1. Overview of chemical functional groups identified and analyzed by the VaPOrS, including number of a) carbon atoms, b) aromatic and c) non-aromatic rings, d) non-aromatic double bonds (C=C), e) hydroxyl (alkyl (e1) and aromatic (e2)), f) hydroperoxide, g) aldehyde, h) carboxylic acid, i) peroxy acid, j) ketone, k) ethers (open-chain (k1), alicyclic (k2), and aromatic (k3)), l) ester, m) nitroester, n) C=CC=O in non-aromatic

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

rings, o) nitrate, p) nitro, q) acylperoxynitrate, r) peroxide, s) nitrophenol, t) amines (primary (t1), secondary (t2), tertiary (t3), and aromatic (t4)), and u) amides (primary (u1), secondary (u2), and tertiary (u3)).

2.1.1. Carbon atoms

Carbon atoms are an essential component in all organic molecules, and their number forms the basis for the calculations of molecular properties utilizing group additivity. The `carbon_number(s)` function counts the occurrences of both uppercase ('C', representing carbon atoms in alkyl chains) and lowercase ('c', representing carbon atoms in aromatic rings) characters in the SMILES string. The total number of carbon atoms is then calculated by summing these occurrences.

2.1.2. Identification of branches and rings in SMILES

Since SMILES notation uses parentheses to denote branching in molecular structures, and numeric indicators to specify ring closure, the model incorporates specialized functions to handle these aspects.

- The function `find_closing_parenthesis(s, open_index)` is designed to locate the corresponding closing parenthesis for a given open parenthesis in the SMILES string. This is crucial for identifying the boundaries of branched substructures. Similarly, `find_opening_parenthesis(s, close_index)` identifies the opening parenthesis corresponding to a closing parenthesis, enabling proper identification of substructures. These functions ensure that the branching parts of the molecules are accurately interpreted, which is essential for correctly assigning functional groups.
- The function `find_cycle_number(s)` is developed to parse through the SMILES string and identify the largest numeric index used for ring closure. This index represents the total number of rings in the compound. For example, if a compound contains three rings, its SMILES notation will include ring closure indicators numbered 1, 2, and 3. Each ring is assigned a specific index sequentially as it is encountered while writing the SMILES. Therefore, the largest numeric index in the SMILES indicates the total number of rings present in the structure.

To illustrate the importance of correctly matching parentheses in the algorithm, consider the following example. The pattern `O=C1C(=C...1...)...` at the beginning of a SMILES notation

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

may suggest the presence of a C=C–C=O group within a non-aromatic ring (such as in cyclohex-2-enone, see example (n) in Figure 1).The parentheses must correspond to each other in the above-mentioned pattern, enclosing the second occurrence of '1'. For instance, in the SMILES string **O=C1C(=C**C)CC(CC**1)**C, although it initially appears to match the pattern, the second '1' does not reside between the matching parentheses. As depicted in Figure 2(a), this structure places the C=C bond outside the ring, diverging from the characteristics of the intended functional group. By contrast, in the strings **O=C1C(=C**CCC**1)**C and **O=C1C(=C**(C(C)C(C)C)CCC**1)**C, the second '1' is properly enclosed within the matching parentheses (highlighted in bold), adhering to the pattern and satisfying the criteria for detecting the C=C–C=O group within a non-aromatic ring. Notably, the latter example contains more carbon atoms, making it easier to count and visually distinguish the structure's complexity. This highlights the versatility of the algorithm in handling SMILES strings of varying complexity, even when the carbon atom count increases.

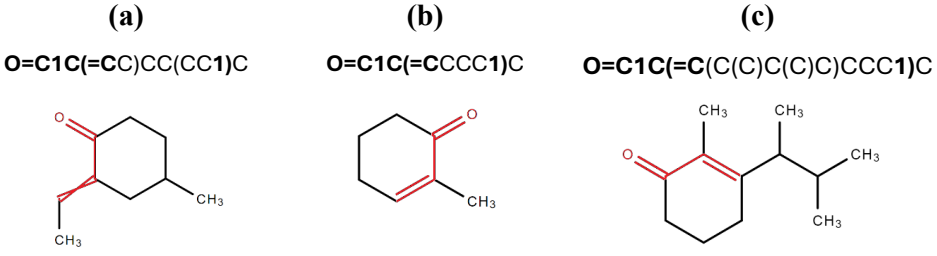


Figure 2. Structural representations of SMILES strings demonstrating the importance of correct parenthesis matching for the identification of C=C–C=O groups in non-aromatic rings. Structure (a) (O=C1C(=CC)CC(CC1)C) incorrectly places the C=C bond outside the ring, while structures (b) and (c) (O=C1C(=CCCC1)C and O=C1C(=C(C)C(C)C)CCC1)C) adhere to the correct pattern.

2.1.3. Aromatic and non-aromatic rings

Organic molecules often contain cyclic structures, which can either be aromatic (such as benzene rings) or non-aromatic (such as cyclohexane). The model uses two specific functions, `aromatic_ring(s)` and `non_aromatic_ring(s)`, to identify and quantify these rings based on the SMILES notation. The function `aromatic_ring(s)` utilizes the previously described `find_cycle_number(s)` function to identify the largest ring number within the SMILES

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

string. This number represents the highest numeric indicator for cyclic structures (e.g., `c1ccccc1` in SMILES denotes a benzene ring with the digit '1' marking the ring). The function then constructs a list of potential aromatic ring representations, such as `'c1', 'c2', …, 'cn'`, where `'n'` is the index of the last detected ring showing how many rings are in the structure. It then iterates over the SMILES string, counting the occurrences of these aromatic rings. To avoid double-counting, the final count is halved, since each ring closure is represented by two digits (e.g., `c1...c1`). The function returns the total number of aromatic rings present in the molecule.

Similarly, the `non_aromatic_ring(s)` function detects non-aromatic cyclic structures. Like the aromatic ring detection, the function constructs list of possible ring representations, including combinations such as `'C1', 'N1', 'O1'` and so on, based on the largest ring index. The function returns the total number of non-aromatic rings detected in the molecule.

2.1.4. Non-aromatic double-bonded carbon atoms

In addition to identifying cyclic structures, the model also detects non-aromatic double bonds between two carbon atoms, a common feature in organic molecules that strongly influences their chemical and physical properties. The function `double_bound_nonaromatic_carbons(s)` is designed to identify occurrences of double-bonded carbon atoms in non-aromatic structures. This function operates by scanning the SMILES string for occurrences of `=C`, which indicates a double bond involving a carbon atom. The following steps are used to ensure accurate detection of non-aromatic double bonds:

- Ring number detection: As with previous functions, `find_cycle_number(s)` is employed to identify the largest numeric ring closure indicator, which is used to distinguish between ring-bound and non-ring-bound carbons.
- Pattern matching: The function scans the SMILES string for `=C`, which denotes a double bond with a carbon atom. Upon finding such occurrences, additional checks are performed to ensure the carbon atoms involved in the double bond are not part of an aromatic system:

    i. Straight chains: If the string preceding the double bond (`=C`) contains another capital `C`, it is counted as a non-aromatic double bond (e.g., `C=C` in `CC=CCCC`).

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

ii. Ring systems: If the string preceding the double bond contains a numeric ring closure indicator (e.g., `C1=C in C1=CCCC1`), it is also recognized as part of a non-aromatic ring structure.

iii. Parentheses handling: The function is equipped to handle more complex structures, such as those involving nested parentheses (e.g., `C(...(...(...)...)...)=C` in `CC(C(C(CC)C)C)=CC`). In SMILES, nested parentheses represent branching in a molecule occurring when an atom in the main chain of the molecule is connected to one or more side chains. The parentheses indicate the start and end of each branch, and nesting occurs when a branch contains another branch. It uses the `find_opening_parenthesis` function to locate the corresponding opening parenthesis and verify that the bond belongs to a non-aromatic system.

This approach ensures that double bonds within non-aromatic systems are accurately counted, even in cases where the SMILES notation involves branching or ring structures.

2.1.5. Alkyl hydroxyl and hydroperoxide groups

The function `hydroxyl_group(s)` is designed to identify hydroxyl groups (–OH) present in the SMILES representation of a compound. To ensure accurate detection, the function incorporates specific conditions to avoid miscounting hydroxyl groups that are in specific arrangements that do not denote targeted simple hydroxyl groups. For example, it ensures that carboxylic acids with hydroxyl groups connected to carbonyls (C=O) are not mistakenly counted as plain hydroxyls. One important condition is that the `C(O)` pattern (a hydroxyl group in SMILES) must not be followed by `=O` or `(=O)`, which would indicate a carboxylic acid rather than a hydroxyl group. The following steps outline how this function operates to ensure accurate detection of hydroxyl groups:

- Ring number detection: The `find_cycle_number(s)` is utilized to determine the presence of cyclic structures to establish whether the hydroxyl group is part of a ring or a straight-chain structure.
- Pattern matching: The function examines the SMILES string for various patterns that denote hydroxyl groups, primarily focusing on terminal hydroxyls, where the function checks if the hydroxyl group appears at the end of the SMILES string, represented as

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

'O' or 'C–O' configurations (e.g., `CCCCO`) or branching with parenthesis (e.g., `CCC(C)O`). It evaluates whether the hydroxyl group appears within a branching structure or a cyclic component, recognizing patterns such as `C(...)O` or `C1(...)O`. Moreover, the function is equipped to deal with complex SMILES representations involving nested parentheses, ensuring that all possible hydroxyl configurations are evaluated (e.g., `(O)` in `CC(O)CC` and `CC(C)(O)CC`).

- Conditions for counting: Specific conditions are implemented to avoid miscounting hydroxyl groups that are either in specific arrangements that do not denote targeted hydroxyl group or are redundant in cyclic structures. In this regard, the function ensures that carboxylic acids with hydroxyl groups connected to carbonyls (C=O) are not mistakenly counted as plain hydroxyls. For example, in the SMILES representation `C(O)=O`, the hydroxyl group is not counted because it is directly bonded to a carbonyl group. As an example, one condition to take into consideration is that the `C(O)` pattern (a hydroxyl group in SMILES) does not proceed with `=O` and `(=O)`. Additional criteria influencing this classification are embedded in the code.

It's important to note that the detection of hydroperoxide groups (–OOH) in VaPOrS follows a similar pattern as that used for hydroxyl groups (–OH). The key difference in recognizing hydroperoxide groups is that instead of detecting single oxygen atoms (O), the algorithm would look for two consecutive oxygen atoms (OO). This adjustment is made throughout the function `hydroxyl_group(s)` and is provided in a new function as `hydroperoxide_group(s)` to recognize and count hydroperoxide groups in addition to hydroxyl groups.

In cyclic structures, hydroxyl groups are excluded if they are connected directly to aromatic rings (e.g., `c1ccc(O)cc1`). This is due to them being considered as another essential functional group called aromatic hydroxyl in the SIMPOL method. The sequential subsection explains how this functionality is described and detected by the developed algorithm.

2.1.6. Aromatic hydroxyl group

The function `aromatic_hydroxyl_group(s)` scans the SMILES string for the presence of hydroxyl groups attached to aromatic systems. The function looks for specific patterns that indicate an aromatic hydroxyl group, focusing on the position of the hydroxyl group in the SMILES string:

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

i. Hydroxyl at the end: The function checks if the SMILES string ends with `O`, ensuring that the preceding character(s) form part of an aromatic ring. For the SMILES string such as `c1cccc1O`, where the hydroxyl group is at the end of the SMILES, it detects the hydroxyl group attached to an aromatic ring (`c1`).
ii. Hydroxyl at the start: The function checks if the SMILES starts with the `Oc1` pattern, indicating a hydroxyl group attached to the first aromatic ring in the compound. `Oc1cc` would be recognized as having an aromatic hydroxyl group at the start.
iii. Hydroxyl in the middle: The function searches for occurrences of the pattern `c(O)` or `c(...O)` where a hydroxyl group is attached to an aromatic ring somewhere in the middle of the SMILES. For the SMILES `c1cc(O)cc1`, the hydroxyl group is identified within the ring. In strings such as `c1c(c...c1O)C`, the hydroxyl group connected to the aromatic ring `c1` is correctly identified.
iv. Hydroxyl at the end of a branch: The function checks for cases where a hydroxyl group is part of a branch but still attached to an aromatic ring. For example, in `c1c(c...)O`, where the hydroxyl group is at the end of a branch attached to an aromatic ring, it detects the correct structure.
v. Handling nested parentheses: If multiple parentheses are involved, the function ensures that the correct ring and attachment points are identified. For example, the pattern `...c1...(...c1)...)O)...` appearing in SMILES like `CC(c1cc(c(c1)CC(C=O)C)O)C` identifies the aromatic hydroxyl correctly.
vi. Hydroxyl connected to the index of aromatic carbon: The function ensures that hydroxyl groups are also counted if they are directly attached to index of carbon atom in an aromatic ring (represented by `c` in SMILES notation). In `c1(O)cccc1`, the hydroxyl group attached to `c1` is correctly counted.

As the final filtration step, after identifying a hydroxyl group in the aromatic ring, the algorithm also checks for the presence of a nitro group within the same aromatic ring. If a nitro group is found in addition to the hydroxyl group, the compound is no longer classified as an aromatic hydroxyl group; instead, it is categorized as a nitrophenol compound. A detailed explanation of this functional group is provided in a subsequent subsection.

2.1.7. Aldehyde group

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

The function `aldehyde_group(s)` is tasked with locating aldehyde groups (i.e., terminal carbonyl groups) within a given SMILES string. The following steps illustrate how this function operates:

- Pattern matching: The function checks the SMILES string for various patterns that indicate the presence of aldehyde groups, focusing on:
  i. Aldehyde at the beginning: The function checks if the SMILES starts with common aldehyde patterns. For instance, In `O=CC`, `C(=O)C`, `O=Cc`, and `C(=O)c` patterns appearing in the SMILES, the function counts these as aldehydes.
  ii. Aldehyde at the end: The function checks if the SMILES ends with the characteristic `C=O` pattern. For example, the last three characters are checked to be `C=O`. If preceded by a `C`, such as in `CC=O`, it counts as an aldehyde. As another example with cyclic structures, for `C1C=O` as the last characters of a string where the aldehyde is linked to a cyclic structure, this is also counted.
  iii. Aldehyde in the middle of the SMILES notation (not structure): The function scans for `C=O` patterns within the SMILES string. For the structure such as `C(C=O)...`, the carbonyl group may appear in the middle of the SMILES string. However, this still corresponds to a terminal carbonyl group in the molecular structure, i.e., an aldehyde and not a ketone. This distinction is important: while the SMILES position may suggest a non-terminal group, the actual bonding context in the molecular structure confirms its identity as an aldehyde.
  iv. Aldehyde at the end of a branch: The function examines occurrences of `C=O)` at the end of branches. For example, the appearance of `CC=O)...` pattern in the SMILES counts it as an aldehyde.
- Conditions for counting: Specific conditions are established to ensure accurate counting and avoid misidentification of aldehyde groups:
  i. Connected to carbon atoms: The function ensures that aldehydes are counted only if they are connected to carbon atoms directly. In `C(C=O)...` pattern, for example, the `C` preceding the `C=O` indicates a valid aldehyde.
  ii. Handling cyclic structures: The function accounts for rings, ensuring that aldehydes connected to cyclic structures are counted appropriately. For example, in `C1(...)C=O`, the presence of `C1` before the aldehyde indicates a connection to a ring, thus counting it as an aldehyde.

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

iii. Parentheses handling: The function effectively manages nested structures, checking for relevant connections before counting. As an instance, if `C=O` is branched from a carbon e.g., in `C(...)(C=O)...` pattern, it counts as an aldehyde.

2.1.8. *Carboxylic and peroxy acid groups*

The function `carboxylic_acid_group(s)` is designed to identify the presence of carboxylic acid groups within a compound. The detection is briefly broken down into the following steps:

- Carboxylic acid at the beginning: The function begins by checking if the SMILES string starts with the characteristic patterns for a carboxylic acid group. Allowed prefixes are `O=C(O)` or `OC(=O)`. For example, in the SMILES string `O=C(O)CC`, the carboxyl group `O=C(O)` at the beginning of the molecule is detected.
- Carboxylic acid as the last characters: The function checks if the SMILES string ends with patterns such as `C(=O)O` and `C(O)=O`, indicating a carboxylic acid group at the end of the molecule. As an instance, for the compound `CCC(=O)O`, the carboxylic acid group at the end is correctly identified as `C(=O)O`.
- Carboxylic acid in the middle of the SMILES: The function searches for occurrences of the carboxylic acid group in the middle of the SMILES string using the patterns like `C(=O)O)` and `C(O)=O)`. Each time one of these patterns is found, the function increases the count of carboxylic acid groups. For instance, in the SMILES `CC(C(=O)O)C(C(O)=O)C`, the carboxylic acid groups `C(=O)O` and `C(O)=O` in the middle are detected.

The detection of peroxy acid groups follows a similar pattern to that used for carboxylic acid groups within the model. The key difference in recognizing peroxy acid groups lies in the algorithm's adjustment to look for two consecutive oxygen atoms e.g., `C(=O)OO` and `C(OO)=O` instead of single oxygen atoms found in carboxylic acids e.g., `C(=O)O` and `C(O)=O`. By implementing this modification throughout the model, peroxy acid functional groups can be counted as well.

2.1.9. *Ketone group*

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

The function `ketone_group(s)` is designed to identify and count the presence of ketone groups in the SMILES string of a compound. These patterns can be located at the start, middle, or end of the SMILES string, as well as within rings. The detection process is divided into several steps:

- Ketone as the first characters: The function checks if the SMILES string begins with a ketone pattern (e.g., `O=C(C...)C` in `O=C(CCC)C` or `O=C(c...)C...` in `O=C(c1ccccc1)CCO`). These patterns represent the ketone group at the beginning of the SMILES string, followed by either a non-aromatic or aromatic carbon. The function also handles ketones connected to rings, such as `O=C1...` in `O=C1CCCC1` indicating a carbonyl group attached to the first position of a ring. Note that although the ketone group appears at the start of the SMILES string, it may not be at the beginning of the molecular structure itself. The pattern is recognized based on bonding context, not string position.
- Ketone as the last characters: The function checks if the SMILES string ends with `=O`, indicating a ketone group at the end of the molecule. If the ketone is ring-connected (e.g., `C1CCCC1=O`), additional checks ensure the presence of a carbonyl group within the ring. Similarly, the ketone group may appear at the end of the SMILES notation, but in the molecular structure, it could be part of a cyclic or internal configuration. The detection logic is based on chemical connectivity, not linear SMILES order.
- Ketone in the middle of the SMILES: The function searches for ketone groups within the middle of the SMILES string using patterns such as `C(=O)`, representing a carbonyl group between two carbons. Also, the function carefully checks if the ketone is branched, ensuring accurate identification of the ketone group in the middle. As an example, in the SMILES string `CC(=O)C`, the middle ketone group `C(=O)` is detected between two carbon atoms.
- Handling parentheses: The function accounts for complex SMILES structures that contain nested parentheses. It ensures that ketone groups within branches (e.g., `...C(C(C...)=O)...` in `CC(C(CCC)=O)CC`) are properly identified by finding the matching opening and closing parentheses.

2.1.10. Open-chain, alicyclic, and aromatic ether groups

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

Three functions `open_chain_ether(s)`, `alicyclic_ether(s)`, and `aromatic_ether(s)` are defined to distinguish between ethers present in the SMILES notation (open-chain, alicyclic, and aromatic ethers). Below is a breakdown of the major components and logical flow within this algorithm.

A key part of the function's operation involves detecting in-ring ethers. The function starts by searching for the sequence "`OC`" within the SMILES string, which is the primary sequence of the ether functional group. The other carbon atom bonded to the oxygen atom of the "`OC`" sequence (e.g., `COC` or `C(OC…)`) is then searched to identify if the ether function is alicyclic or open-chain. For each occurrence, the algorithm checks if the sequence is part of a non-aromatic ring by comparing its position relative to the numerical ring indicators. This is done by locating the positions of ring closure numbers (e.g., `C1...C1`) relative to the ether group. If the ether group lies outside the ring closure points (e.g., `C1CCCC1COCC`), it is classified as an open-chain ether. On the other hand, if the ether group is found within the two identical ring numbers (e.g., `C1...COC..C1`), the algorithm hesitates if the ether is part of the ring or an open-chain type. Therefore, the function employs a nested structure-parsing approach to clarify this issue. It uses parentheses to detect branching points or nested structures within the molecule. The algorithm carefully traces the boundaries of rings and other nested structures simultaneously. If the ether group is embedded within parentheses and surrounded by ring numbers (e.g., pattern `C1...(...COC...)...C1` in SMILES `C1CC(COC)CC1`), it is open-chain ether, while if it follows patterns, such as `...C1...(...COC...C1...)...`, it is alicyclic (e.g., SMILES `C1C(COCC1C)CC`). On the other hand, the function also checks if the ether is located near the start or end of a ring closure, which would indicate that the ether is alicyclic (e.g., pattern `C1...C(...)O1`), otherwise, it would be an open-chain ether. In compounds containing multiple rings, the function iterates through each ring. It systematically searches for ether groups within and around each ring, ensuring that all possible locations are checked for ether group presence. If the ether group is attached to a non-aromatic ring, special handling is performed to ensure accurate detection.

On the other hand, for aromatic ethers, the model looks for occurrences of `'Oc'` within the SMILES string. The logic checks different cases based on what precedes `'Oc'`:

- Simple alkyl group (`'C'` before `'Oc'`): This detects linear aromatic ethers where the oxygen is connected directly to an alkyl group (e.g., `COc1ccccc1`).

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

- Nested structures with parentheses: The model handles cases where the ether group is part of a complex structure. If the ether group is surrounded by parentheses, the model traces the original alkyl chain, ensuring it connects to an aromatic ring (e.g., patterns `C(Oc` and `C(...)Oc` in SMILES `CC(Oc1cccc1)CCC` and `CC(C(O)=O)Oc1cccc1`).
- In case the ether group is attached to one aromatic and one non-aromatic ring, the model iterates over possible ring numbers to find ether groups (e.g., `'C1CCCC1Oc1ccccc1'` where `'C1'` is part of a ring and `'Oc1'` is the ether group). The same logic is applied to nested structures with parentheses, ensuring that the correct connection between alkyl and aromatic groups is maintained.
- The model repeats the search, but this time looking for `'OC'`, where the aromatic group comes first (e.g., `'c1cccc1OC'` where `'c1'` indicates an aromatic ring and `'OC'` is the ether). Similar checks are performed for direct aromatic ether group connection (`'c1OC'`), complex nested structures (e.g., `c1(OC)ccc1` and `c1cc(OC)cc1`), and parentheses-based structures (e.g., `c1cc(ccc1)OC`).

2.1.11. Ester and nitroester groups

The developed function, `ester_group(s)`, identifies and quantifies ester groups within a given SMILES string. The ester group is characterized by the bonding of a carbonyl group (C=O) to an alkoxy group (O–R). This pattern is typically (not always) represented in SMILES as `C(=O)O`, and variations in its placement within the molecule must be accounted for. For instance, when analyzing the ester group, if the alkoxy group (-OR) shows up in the SMILES first and then the carbonyl group (C=O), common patterns, such as `...C(...)(...)OC(=O)C...` (e.g., in `CC(CO)(C)OC(=O)CC` with bolded characters) and `...C(...)(OC(=O)C...)...` (e.g., in `CC(CO)(OC(=O)CC)C` with bolded characters) are identified. Conversely, if the SMILES notation proceeds from the acid-side carbon (i.e., carbon attached directly to the carbonyl group in the ester bond –C(O)OR), ester groups may be represented as `...C(...)(...)C(=O)OC...` (e.g., in `CC(CO)(C)C(=O)OCC` with bolded characters) or `...C(...)(C(=O)OC...)...` (e.g., in `CC(CO)(C(=O)OCC)C` with bolded characters). These patterns ensure that ester groups are recognized irrespective of their position within a molecule.

The above-mentioned patterns with several other ones related to ester group may be determined whether at the start, middle, or end of the SMILES string:

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

- If the ester group is at the beginning, the function looks for patterns such as `O=C(C...)OC...` and `O=C(OC...)C` (e.g., in SMILES `O=C(CCO)OCCC` and `O=C(OCCC)CCO`), where the group starts with the double-bond oxygen atom.
- For esters embedded within the middle of a SMILES string, the function searches for the ester signatures such as `C(=O)O` and `OC(=O)`, confirming that the carbonyl group is bonded to the oxygen atom of the alkoxy group and followed by a suitable molecular fragment. For instance, in a SMILES notation like `CCC(=O)OCC` and `CCOC(=O)CC`, the ester group is correctly identified as part of the main chain.
- If an ester group is located at the end of the SMILES string, the function identifies patterns such as `)=O` sequence as an indicator of the terminal carbonyl group, followed by the appropriate bonding structure. This ensures that molecules like `CC(OCC)=O` are correctly parsed with the ester group assigned to the end of the chain.
- In cases of aromatic esters, where alternating single and double bonds are common, the function adapts the detection logic to properly handle aromaticity. For example, it correctly identifies the ester group in a molecule like `O=C(OCCC)c1ccccc1`, where the ester is attached to an aromatic benzene ring.
- The algorithm is designed to detect esters even in highly branched molecules. For instance, in a molecule like `CCC(C(=O)OCCC)(C)CC`, the ester group is part of a branching structure, and the function ensures it is accurately parsed by considering the nested arrangement of atoms.

On the other hand, the model locates the starting and ending positions of the acid-side branch for the ester group and searches for the presence of a nitro group (N(=O)=O) in the branch. If a nitro group is detected, the compound is classified as a nitroester, and the `nitroester_number` is incremented. In cases where no nitro group is present, the compound is classified as a regular ester, and the `ester_number` is incremented accordingly.

2.1.12. C=CC=O in non-aromatic rings

The function `nonaromatic_CCCO(s)` is designed to quantify C=C–C=O substructures in a molecular SMILES notation. This section explains how the function works.

Since the substructure of interest occurs within rings, the function first checks whether the molecule contains any cyclic structures. The function uses `find_cycle_number(s)` to determine if any rings exist in the molecular structure. If no rings are found, the function

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

terminates early. When rings are detected, the function proceeds to locate their positions within the SMILES string.

The next step is to detect the C=C–C=O group within the identified rings. This involves scanning the part of the SMILES string that represents each ring and checking for the specific pattern of atoms and bonds. Several important aspects are considered during this analysis:

- Pattern search: Within the extracted portion, the function looks for patterns that match the C=C–C=O group. For example, the presence of `C=O` at the start of a ring followed by a conjugated double bond (`C=C`) within the ring (e.g., pattern `O=C1...C(...)=C1` bolded in SMILES **`O=C1`**`CC`**`C(OC)=C1`**`C`), or variations where the C=C and C=O groups may be spaced by additional atoms, or where they may appear in different locations within the ring (e.g., pattern `...1...C(=O)C(...)=C...1...` bolded in SMILES `O=C`**`1`**`C`**`C(=O)C(CC)=C`**`C`**`1`**`C`).
- Handling complex ring structures and nested rings: The function accounts for these complexities by carefully navigating through the parentheses and ensuring that the entire cyclic structure is examined for the C=C–C=O group (e.g., pattern `...C(=C(C(...1...)=O)...)...` bolded in SMILES `CC1`**`C(=C(C(CC1C)=O)C)`**`CC`).

2.1.13. Nitrate group

The `nitrate_number(s)` function identifies and quantifies nitrate functional groups within the provided SMILES notation. This function examines specific configurations characteristic of nitrate, including the standard nitrate structure `ON(=O)=O`, the quaternary form `O[N+](=O)[O-]`, the N-nitro structure `O=N(=O)O`, and variations like `O=[N+]([O-])O` and `[O-][N+](=O)O`, which collectively reflect diverse bonding scenarios in nitrate chemistry. The function calculates the count of each of these configurations and aggregates them to derive the total number of nitrate groups present.

2.1.14. Nitro group

The identification of nitro groups is executed through the `nitro_group(s)` function, which analyzes the SMILES notation to quantify nitro functional groups. This function employs a series of search operations to locate specific nitro structures, notably `N(=O)=O`, `O=N(=O)`, and various ionic forms such as `[N+](=O)[O-]`, `O=[N+][O-]`, and `[O-][N+](=O)`. Each search utilizes a loop that not only finds occurrences of these structures but also ensures that adjacent

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

atoms do not disrupt the nitro configuration—specifically, it checks that there is no oxygen atom directly connected to the nitro group, which would suggest an alternative bonding scenario (i.e., nitrate group). As the final filtration step, the algorithm verifies whether the identified nitro group is not part of an aromatic ring (i.e., benzene). If this condition is met, it then checks for the presence of a hydroxyl group in the ring according to section 2.6. If no hydroxyl group is found, the nitro group is counted independently. However, if a hydroxyl group is present, the compound is categorized as a nitrophenol group. A detailed explanation of this functional group is provided in a subsequent subsection.

2.1.15. Nitrophenol

The function `nitrophenol_group(s)` is designed to identify and count the number of nitrophenol groups. After identifying aromatic hydroxyl group, the function then checks for the presence of a nitro group in the same aromatic ring within the SMILES string:

- As the last character sequence: It inspects whether the string ends with the motif `N(=O)=O`, which signifies a nitrophenol. The function checks if it is connected to the aromatic cyclic structure by examining the characters before it. For example, if the character is an aromatic ring identifier, the counter is incremented accordingly (e.g., `c1c`**`c(O)`**`cc`**`c1N(=O)=O`**).
- As the first character sequence: The function checks if the string starts with the pattern `O=N(=O)c1`, indicating a nitrophenol positioned within the SMILES string (e.g., **`O=N(=O)c1c`**`c`**`(O)`**`ccc1`). If this pattern is detected, the counter is incremented.
- As middle character sequences: A loop is employed to search for occurrences of patterns such as `c(N(=O)=O)`, which indicates that nitrophenol is positioned within the structure (e.g., `c1c`**`c(N(=O)=O)`**`cOc`**`c1O`**). Each time this pattern is found, the counter is increased. The function also iterates through the previously established list of rings to search for patterns of the form `j(N(=O)=O)` (where `j` is an aromatic carbon identifier). Whenever a match is found, the counter is incremented (e.g., **`c1(N(=O)=O)`**`c`**`c(O)`**`ccc1`).

2.1.16. Acylperoxynitrate group

The function `carbonylperoxynitrate_group(s)` is designed to detect the presence of acylperoxynitrate groups within a given SMILES string `s`. This function systematically counts occurrences of several distinct structural motifs characteristic of acylperoxynitrates, including

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

`OON(=O)=O`, `OO[N+](=O)[O-]`, `O=N(=O)OO`, `O=[N+]([O-])OO`, and `[O-][N+](=O)OO`. The function aggregates the counts of these motifs into a single variable, `carbonylperoxynitrate`, which represents the total number of acylperoxynitrate groups identified.

2.1.17. Peroxide group

The function `peroxide_group(s)` is engineered to identify and quantify peroxide groups. The function first determines the number of cyclic structures within the SMILES string by calling the `find_cycle_number(s)` function. If cyclic structures are found, a list of ring identifiers, comprising both carbon (`C`) and aromatic (`c`) rings, is generated to facilitate later checks. Subsequently, the function employs a loop to search for occurrences of the `OOC` motif, which signifies the presence of peroxy groups. Each iteration of the loop calls `s.find('OOC', ...)` to locate the next occurrence of the motif. Multiple conditions are assessed to ensure that the identified `OOC` is correctly positioned relative to other atoms or rings:

- Adjacent carbon or aromatic carbon atoms: If the character preceding `OOC` is `C`, `Cx` or `cX` (`x=1,2,…` is cyclic index), indicating that `OOC` is bonded to a carbon atom, the `peroxide_number` is incremented (e.g., `CCCOOCC`, `C1CCCC1OOCC` and `Cc1cccc1OOC`).
- Branching structures: If the character before `OOC` is a parenthesis, the function retrieves the index of the last corresponding parenthesis before `OOC` and verifies that the atom preceding this parenthesis is a carbon atom or an aromatic ring. If so, the counter is increased. For example, when the character before `OOC` is a closing parenthesis, the function checks whether the `OOC` is connected to a carbon atom in a similar manner as described previously. This involves searching back to the last opening parenthesis and ensuring the atom connected to that parenthesis is a carbon atom or part of a cyclic structure (e.g., pattern `...C(...)OOC...` bolded in SMILES `CCC(CO)OOCCC`).

These checks comprehensively ensure that only valid peroxide groups are counted, accounting for the complex connectivity possible within SMILES representations. The function ultimately returns the `peroxy_number`, providing a quantitative measure of peroxide groups within the molecular structure.

2.1.18. Aromatic amine group

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

Following the identification of aromatic rings, the function `aromatic_amine_group(s)` locates nitrogen atoms (`N`) within the SMILES string and determines their bonding to aromatic carbons.

- Direct bond to aromatic carbon (`cN`): If nitrogen (`N`) is found immediately following an aromatic carbon (`c`), it is counted as part of an aromatic amine group. To ensure the nitrogen atom belongs to an aromatic amine and not a nitro group (-NO2), the model incorporates an additional condition.
- Nitrogen in numbered rings (`c1N`, `c2N`, etc.): If nitrogen is attached to a carbon in a numbered ring, such as `c1N`, the nitrogen is identified as part of an aromatic amine group.
- Parenthetical structures (`c(...)N`): In cases where nitrogen is attached within parentheses following a cyclic group, the function checks if the nitrogen is part of an aromatic ring by verifying the bonding pattern of the cyclic carbon to the nitrogen. Parentheses in SMILES represent branching, and this function ensures that any branching nitrogen groups attached to aromatic carbons are also detected.

The function iterates through the SMILES string to ensure that all occurrences of nitrogen atoms are evaluated. The bonding pattern of each nitrogen atom is assessed against the aromatic rings identified in the first step. If the nitrogen atom is confirmed to be attached to an aromatic ring and not part of a nitro group, it is counted as part of an aromatic amine group. The total count of such groups is stored in the variable `aromatic_amine_number` and returned as the output of the function.

2.1.19. Primary, secondary, and tertiary amide and amine groups

The model is designed to detect and count primary, secondary, and tertiary amide and amine groups in a molecule represented by a SMILES string. A primary amide has the functional group structure –C(=O)NH2, and the model identifies both simple and branched forms of this group. To differentiate between primary amides and primary amines, the model specifically excludes patterns without a double-bonded oxygen, i.e., –C(...)NH2 where '...' is not =O, thereby ensuring correct identification of amide groups versus amine counterparts.

- The function first identifies primary amide groups located at the beginning of the SMILES string by searching for patterns such as `O=C(N)` or `NC(=O)`, as seen in SMILES

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

O=C(N)CCCO and NC(=O)CCCO). It also detects primary amides at the end of the SMILES string by tracing patterns like C(=O)N and C(N)=O, as in CCOCC(=O)N and CCOCC(N)=O. To identify primary amine groups at the beginning or end of the strings, the function checks for similar patterns but without a carbonyl group (=O). In other words, if a nitrogen is bonded to a carbon that does not carry a =O, it is interpreted as an amine rather than an amide. For example, C(N)CCCO, NCCCCO, CCOCCN, and CCOCCN are recognized as containing primary amine groups.

- The model also identifies primary amides within branches or internal positions in the molecule by searching for specific patterns, such as C(=O)N) and C(N)=O, as seen in the SMILES strings CC(C(=O)N)CC and CC(C(N)=O)CC. To identify primary amine groups in similar positions, the model checks for the absence of the double-bonded oxygen (=O) on the carbon adjacent to the nitrogen. This results in patterns like CN) in CC(CN)CC and CC(CN)CC.

For secondary and tertiary amides, the model similarly searches for patterns where the nitrogen atom is bonded to two or three carbon atoms, respectively. Secondary amides have the structure –C(=O)NR, where R represents an alkyl group starting with a carbon atom attached to the nitrogen atom, and tertiary amides have the structure –C(=O)N(R)R', where both R and R' are alkyl groups starting with carbon atoms attached to the nitrogen atom. The model identifies these structures by recognizing the presence of additional carbon attachments to the nitrogen atom. For secondary amides, the function searches for patterns that indicate the nitrogen is bonded to one additional carbon group, distinguishing them from primary amides by checking for two single bonds to nitrogen, along with the carbonyl group. Similarly, for tertiary amides, the function detects two alkyl groups attached to the nitrogen atom in addition to the carbonyl group. Once these amide patterns are identified, the model applies the same exclusion method for the double-bonded oxygen, converting these amides into their corresponding secondary and tertiary amines. For secondary amines, the nitrogen is attached to two carbon atoms, and for tertiary amines, the nitrogen is bonded to three carbon atoms. This method ensures that the correct amide or amine group is identified and classified, whether it is primary, secondary, or tertiary, based on the number of carbon attachments to the nitrogen atom.

Finally, the model counts the number of carbon atoms in the secondary and tertiary amides that are not part of the R and R' groups in the structure. This count is considered as the number of carbons on the acid side of the amides. For primary amides, since there are no additional alkyl

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

groups attached to the nitrogen atom, all carbon atoms in the structure are considered to be on the acid side of the amide. This ensures accurate categorization and counting of carbon atoms associated with the amide's acid side, contributing to the overall structural analysis of the molecule.

2.2.Saturation vapor pressure calculation

After identifying the functional groups, the detection functions return integer values representing the occurrence of each functional group, which are stored in an array. These values are critical input parameters for the subsequent saturation vapor pressure calculation, as the saturation vapor pressure is quantified as the sum of functional group contributions in the SIMPOL method. Thus, each SMILES string (e.g., for compound i) is processed to retrieve the saturation vapor pressures according to the SIMPOL method. The SIMPOL method defines how each functional group contributes to the saturation vapor pressure—at different temperatures. In the model:

i. Matrix B is read from a pre-defined text file containing coefficients of $B_{k,1}$, $B_{k,2}$, $B_{k,3}$, and $B_{k,4}$ for each functional group k, according to Table 5 of (Pankow and Asher 2008). A select functional group is assigned a value for contribution to saturation vapor pressure in the i^{th} SMILES string (such as hydroxyl, aldehyde, and ketone groups, etc.).
ii. Then, $b_k(T)$ and $P^0_{L,i}(T)$ are calculated for any given temperature according to Equations 1 and 2. The total liquid (saturation) vapor pressure, $P^0_{L,i}$ (atm), is calculated as the sum of all functional group contributions.

$$b_k(T) = \frac{B_{k,1}}{T} + B_{k,2} + B_{k,3}\,T + B_{k,4}\,\ln T \tag{1}$$

$$\log_{10} P^0_{L,i}(T) = \sum_k \nu_{k,i}\, b_k(T) \quad \mathrm{k} = 0, 1, 2, \ldots \tag{2}$$

where $\nu_{k,i}$ is the number of groups of type k, $b_k(T)$ is the contribution to $\log_{10} P^0_{L,i}(T)$ by each group of type k, and T is the temperature. Also, *0* and *L* show the reference and Liquid.

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

iii. To fit the saturation vapor pressure data to the Antoine equation (Equation 3) to enable further use of the saturation vapor pressure values in different applications, the saturation vapor pressure is calculated at 1000 temperature points across a wide range of temperatures from 220 K to 450 K according to Equation 2. The obtained data are then used in a non-linear least squares fitting procedure, which minimizes the difference between the data and the saturation vapor pressure values predicted by the Antoine equation. The Antoine equation parameters (i.e., A, B, and C) are then obtained for each compound.

$$\log P_{sat} = A - B/(T + C) \tag{3}$$

iv. After obtaining the Antoine equation parameters, the vaporization enthalpy relationship can be derived using the Clausius-Clapeyron equation:

$$\frac{d \log P_{sat}(T)}{d(\frac{1}{T})} = -\frac{\Delta H_{vap}(T)}{2.303R} \tag{4}$$

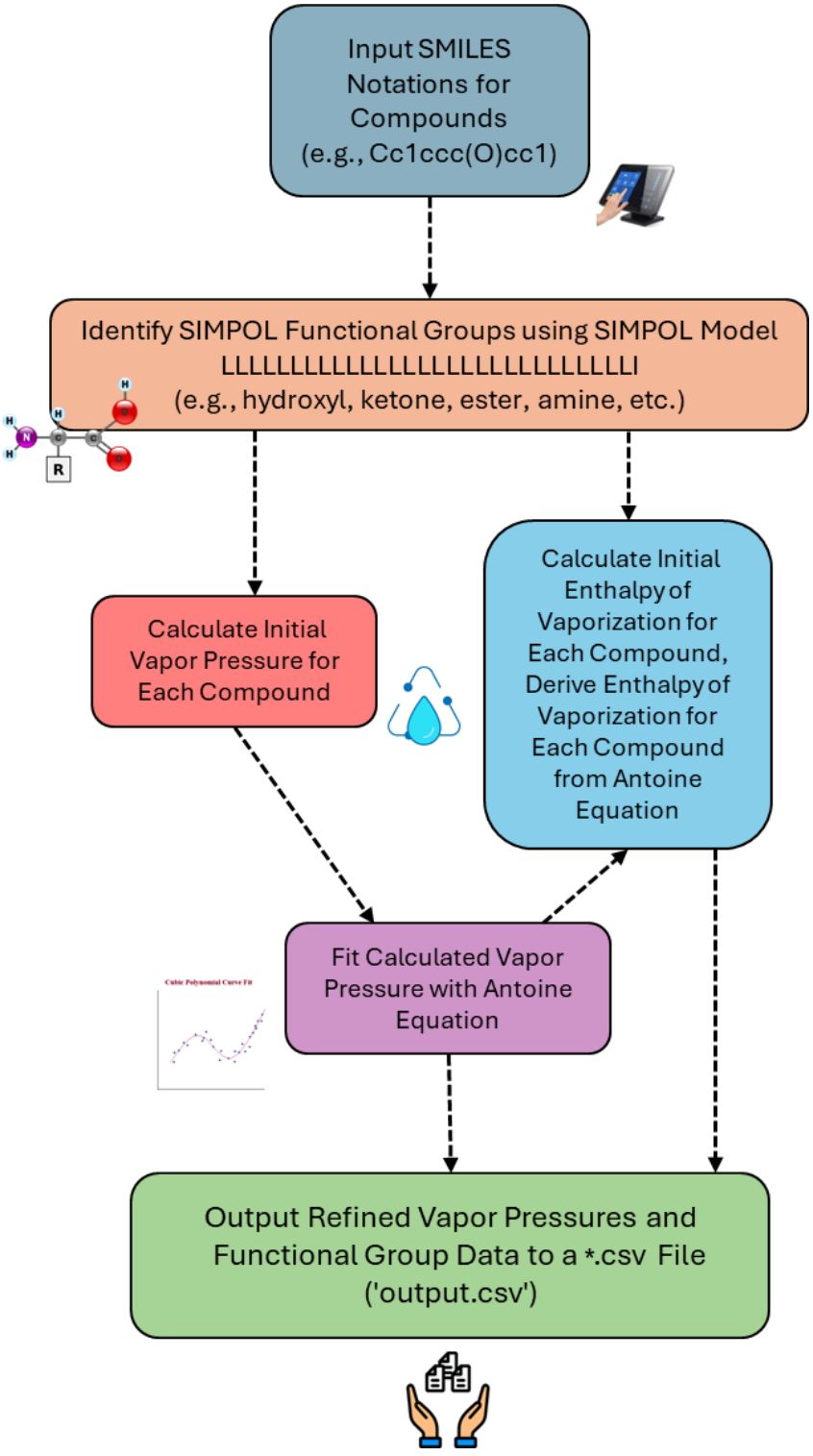
$$\Delta H_{vap}(T) = -2.303R\left(\frac{d \log P_{sat}(T)}{d(\frac{1}{T})}\right) \tag{5}$$

Here, $\Delta H_{vap}(T)$ is the temperature-dependent enthalpy of vaporization, and R is the universal gas constant. This expression relates the slope of the logarithm of the saturation pressure with respect to the inverse of temperature to the enthalpy of vaporization. The Antoine equation provides a framework to calculate saturation vapor pressure at any given temperature, and this relationship extends the utility of the model by allowing the determination of thermodynamic quantities such as vaporization enthalpy.

v. The output is written to a CSV file named by the user (e.g., output.csv), where each line corresponds to the i^{th} compound and its associated data, including its SMILES string, the count of each functional group in its structure, its saturation vapor pressure at 300 K, and its fitted Antoine equation parameters.

Figure 3 illustrates a flowchart of the VaPOrS from input to output.

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025



https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

Figure 3. Flowchart of the automated process of VaPOrS, beginning with SMILES notation input, followed by the identification of functional groups, initial saturation vapor pressure and enthalpy of vaporization calculations, their fitting to the Antoine equation, and final output.

3. Results and discussion

3.1.Saturation vapor pressure

To evaluate the accuracy of the developed automated VaPOrS in predicting the saturation vapor pressures of organic compounds using the SIMPOL method, the model was applied to a subset of the original dataset used to develop the SIMPOL parameterization. Saturation vapor pressures for 224 organic compounds were calculated using VaPOrS and compared against experimental data to assess the accuracy of the implementation and ensure consistency with established results.

Figure 4 presents a comparative analysis between the saturation vapor pressures computed by the VaPOrS and those reported by the SIMPOL method and measurement at a specific temperature (i.e., 333.15 K). The x-axis represents the experimental saturation vapor pressures, while the y-axis represents the numerical values calculated by the VaPOrS (filled symbols with no edge color) and the SIMPOL method (black-edged hollow symbols). Due to a complete overlap in the data points, the black-edged symbols from the SIMPOL method can be seen over the filled symbols from VaPOrS. A diagonal line is shown in the figure, indicating the ideal correlation where the computed values would perfectly match the measured saturation vapor pressures. Data points positioned close to this diagonal demonstrate a high level of agreement between the two approaches.

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

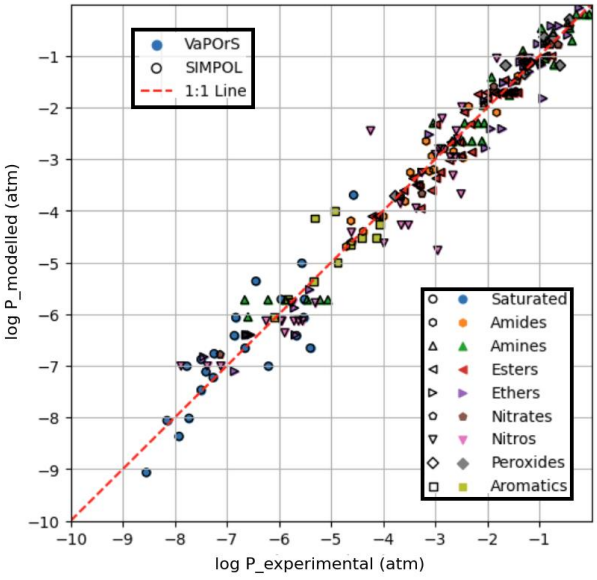


Figure 4. Comparison of saturation vapor pressure values calculated by the VaPOrS (filled symbols with no edge color) and the SIMPOL method (symbols with no face color and black edges) against measured saturation vapor pressures at 333.15 K. The diagonal line represents the ideal 1:1 correlation, and the proximity of data points to this line indicates the accuracy of both methods in predicting saturation vapor pressures. The overlap in symbols is visible due to the black-edged SIMPOL markers covering the filled VaPOrS symbols.

Figure 5 illustrates the saturation vapor pressure results obtained using the VaPOrS model, the SIMPOL method, and experimental measurements at six different temperatures: 273.15 K, 293.15 K, 310.15 K, 333.15 K, 353.15 K, and 373.15 K. Similar to Figure 2, the x-axis represents the measured saturation vapor pressures, while the y-axis displays the values calculated by VaPOrS (filled symbols with no edge color) and the SIMPOL method (black-edged hollow symbols). This figure includes a larger dataset, allowing for a more comprehensive assessment of model performance across a range of temperatures. Many data points are clustered close to this line, further confirming the effectiveness of the VaPOrS model in predicting saturation vapor pressures for various compounds across different temperatures as well. Some deviations from the line are observed e.g., the experimental values for the nitro and saturated compound saturation vapor pressures and *T*-dependencies have large uncertainties as seen in Figures 4 and 5. These discrepancies, which were present in the original dataset, do not undermine the overall trend, which demonstrates strong agreement among the

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

three methods and suggests that the VaPOrS code is reliable and robust across a wider temperature range. Most importantly, we see complete agreement in the output of VaPOrS and SIMPOL.

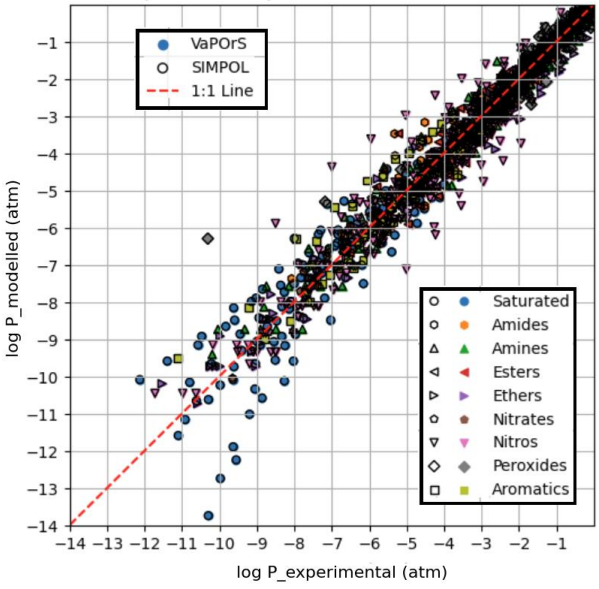


Figure 5. Comparison of saturation vapor pressures obtained from the VaPOrS model, the SIMPOL method, and experimental measurements at six different temperatures (273.15 K, 293.15 K, 310.15 K, 333.15 K, 353.15 K, and 373.15 K). The x-axis represents the measured saturation vapor pressures, while the y-axis shows the values calculated by VaPOrS (filled symbols) and the SIMPOL method (hollow symbols with black edges). The diagonal line indicates the ideal correlation, with points near the line demonstrating good agreement between the methods.

3.2.Enthalpy of vaporization

Figure 6 illustrates the relationship between the experimental vaporization enthalpy values and those calculated by both the VaPOrS and the SIMPOL method at 333.15 K. The x-axis represents the experimentally measured values, while the y-axis displays the calculated values. The results from the VaPOrS are represented by filled symbols without edge color, indicating a direct prediction from the present model. In contrast, the SIMPOL method results are depicted

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

as symbols with no face color and distinct black edges. The presence of overlapping symbols highlights instances where the calculated values from SIMPOL cover those from VaPOrS. According to Figures 4 to 6, many data points are clustered near the diagonal line, demonstrating the effectiveness of the model in predicting saturation vapor pressure and enthalpy of vaporization for several compounds. However, there are some outliers where the calculated values deviate from the experimental counterparts. These discrepancies could be attributed to the structural complexity of certain compounds making intramolecular interactions important and not amenable to simple group additivity predictions, which is also a known limitation in the SIMPOL method itself, or potentially, they could point out issues in the original experimental measurements.

The ability to estimate vaporization enthalpy alongside saturation vapor pressure enables a more comprehensive analysis of the compounds' behavior, especially for estimating the tendency to form aerosol, in various atmospheric and environmental conditions.

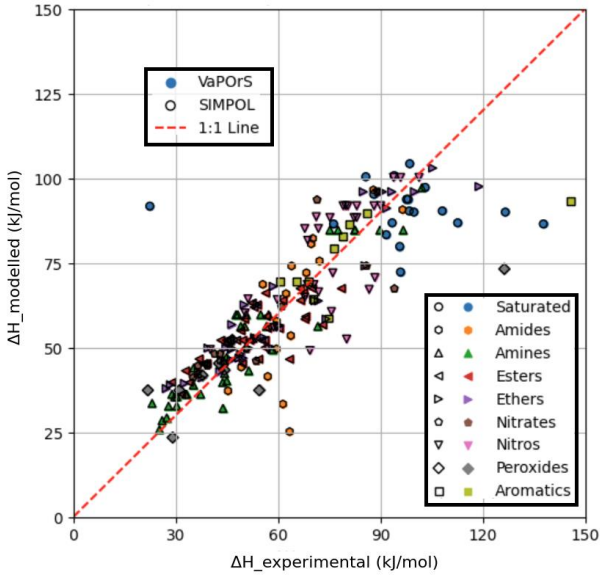


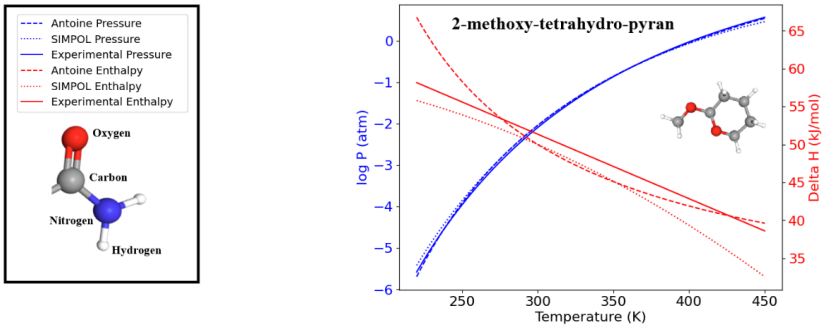
Figure 6. Comparison of enthalpy of vaporization values calculated by the VaPOrS (filled symbols without edge color) and the SIMPOL method (symbols with no face color and black edges) against experimental measurements at 333.15 K. The diagonal line indicates the ideal 1:1 correlation, showcasing the accuracy of the

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

models. Overlapping symbols are observed, with black-edged SIMPOL markers obscuring the filled VaPOrS symbols.

3.3.Antoine equation

The results of the fitting process for nine compounds, i.e., 2-methoxy-tetrahydro-pyran, decanedioic acid, methyl-benzoate, phenylamine, hexanamide, phenylmethyl-nitrate, 2-methyl-6-nitrobenzoic acid, diethyl-peroxide, and 2-napthol as representatives of ethers, saturated, esters, amines, amides, nitrates, nitro-compounds, peroxides, and aromatics, respectively, are visualized in Figure 7. The figure illustrates the temperature-dependent behavior of both pressure and enthalpy of vaporization for Antoine and SIMPOL relationships generated by VaPOrS and compares them with experimental data across varying temperatures. The left y-axis represents the logarithmic saturation vapor pressure in atmospheres, while the right y-axis shows the enthalpy of vaporization in kJ/mol. This dual-axis representation enables a direct visual comparison between pressure and enthalpy trends as the temperature increases. The fitting results demonstrate a high degree of agreement between the Antoine and SIMPOL curves for all compound classes, implying that the Antoine equation given by VaPOrS can be applied effectively to estimate saturation vapor pressures with good accuracy across a broad range of temperatures, enhancing the utility of the saturation vapor pressure data for various applications.



https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

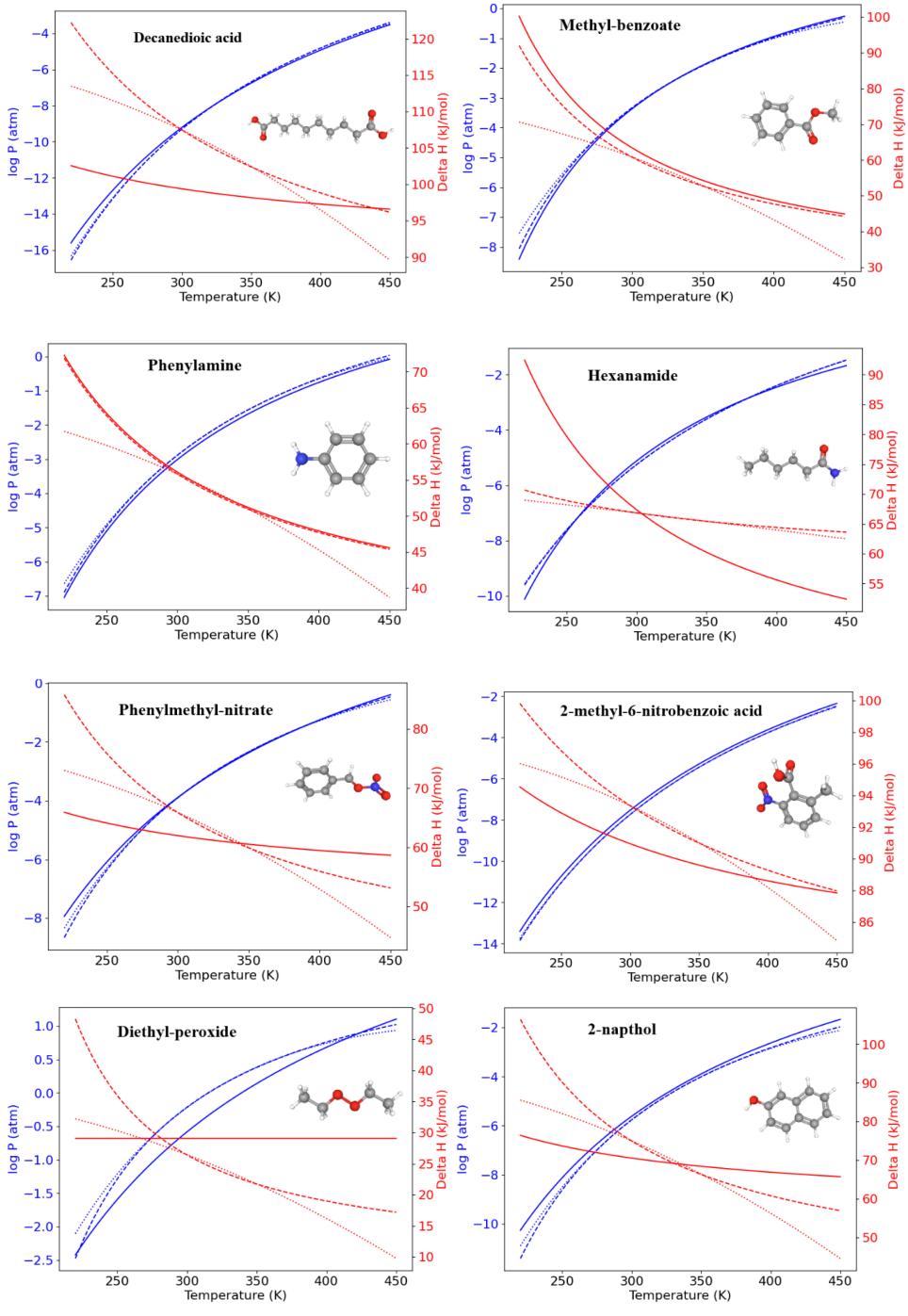


Figure 7. Temperature dependence of saturation vapor pressure and enthalpy of vaporization for nine representative organic compounds with nine distinct functional groups predicted by VaPOrS using the Antoine

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

and SIMPOL equations. The left y-axis shows the logarithmic saturation vapor pressure (in atm), and the right y-axis displays the enthalpy of vaporization (in kJ/mol). The results demonstrate the data generated by Antoine and SIMPOL methods across the temperature range, with experimental data closely matching the theoretical predictions.

3.4.VaPOrS application

Using VaPOrS, a detailed comparison was performed between the manual counting and calculation of functional groups and saturation vapor pressures and the automated results generated from the compounds' SMILES notations. The manual counting involved a systematic review of each compound's molecular structure, visually identifying and recording the functional groups, followed by calculating its saturation vapor pressure according to SIMPOL group contributions. This was then cross-referenced with the automated results generated by VaPOrS to ensure consistency. The procedure was performed in three steps described next.

3.4.1. MCM data

In the first step, a dataset of 126 primary VOCs sourced from the Master Chemical Mechanism (MCM) database are evaluated. While the MCM database provides detailed chemical mechanisms for atmospheric chemistry, its coverage of primary organic compounds is relatively limited compared to the vast diversity of VOCs present in the atmosphere. Nonetheless, it serves as a valuable resource for validation, given its detailed representation of key compounds. Notably, the MCM database not only provides the structures of these compounds but also includes their corresponding SMILES notation, facilitating an accurate assessment of functional group presence through the VaPOrS method.

Table 1 presents a sample comparison between the manual and automated counts of the functional groups for representative compounds from several categories in the MCM, including Alcohols and Glycols, Aldehydes, Alkanes, Alkenes, Alkynes, Aromatics, Dialkenes, Esters, Ethers and Glycol Ethers, Ketones, Monoterpenes and Sesquiterpenes, Organic Acids, and Unclassified compounds. As shown, the results from both methods are in complete agreement, with 0% discrepancy across all cases.

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

Table 1: Comparison of Manual and Automated Counts of Functional Groups for Representative Compounds across Various Categories in the MCM Database.

Category	Compounds	SMILES	Functionals	Consistency
Alcohols and Glycols	CYCLOHEXANOL	OC1CCCCC1	6 carbons, 1 nonaromatic ring, 1 hydroxyl	100%
Aldehydes	PROPENAL	C=CC=O	3 carbons, 1 C=C (non-aromatic), 1 aldehyde	100%
Alkanes	3-METHYLPENTANE	CCC(C)CC	6 carbons	100%
Alkenes	1-HEXENE	CCCCC=C	6 carbons, 1 C=C (non-aromatic)	100%
Alkynes	ETHYNE	C#C	2 carbons	100%
Aromatics	ETHYL BENZENE	CCc1ccccc1	8 carbons, 1 aromatic ring	100%
Dialkenes	1-3 BUTADIENE	C=CC=C	4 carbons, 2 C=C (non-aromatic)	100%
Esters	ETHYL ACETATE	CCOC(=O)C	4 carbons, 1 ester	100%
Ethers and Glycol Ethers	2-METHOXY ETHANOL	COCCO	3 carbons, 1 hydroxyl, 1 ether	100%
Ketones	CYCLOHEXANONE	O=C1CCCCC1	6 carbons, 1 nonaromatic ring, 1 ketone	100%
Monoterpenes and Sesquiterpenes	ALPHA-PINENE	CC1=CCC2CC1C2(C)C	10 carbons, 2 nonaromatic rings, 1 C=C (non-aromatic),	100%
Organic Acids	PROPANOIC ACID	CCC(=O)O	3 carbons, 1 carboxylic acid	100%
Unclassified	ETHYLENE OXIDE	O1CC1	2 carbons, 1 nonaromatic ring, 1 ether (alicyclic)	100%

In the second phase, alpha-pinene and benzene were selected as case studies to evaluate the species formed during their tropospheric degradation via gas-phase chemical processes, focusing on functional group occurrence and saturation vapor pressure. This analysis leveraged the detailed mechanism in the MCM to further validate the automated functional group detection system's accuracy in modeling atmospheric chemistry. For each species, the occurrences of functional groups were manually counted, and saturation vapor pressure was calculated using the SIMPOL method. Their SMILES notation was then input into the VaPOrS to automatically obtain functional group counts and saturation vapor pressures. The saturation vapor pressures obtained through both methods are compared in Figures 8 and 9 for alpha-pinene and benzene, respectively, where the y-axis represents the logarithmic saturation vapor pressure and the x-axis the molar mass of each species. Automated results are shown as color bars based on the number of detected functional groups, and their manual counterparts are displayed as points. The perfect alignment of points atop bars for each species indicates excellent agreement between both approaches. It is worth mentioning that C6H6N2O11 and CH2O (i.e., NNCATECOOH and HCHO in the MCM) were recognized as the least and most volatile species in the benzene oxidation process. On the other hand, C9H16O6 and CH3O

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

(i.e., C922OOH and CH3O in the MCM) were the least and most volatile species in the alpha-pinene oxidation process.

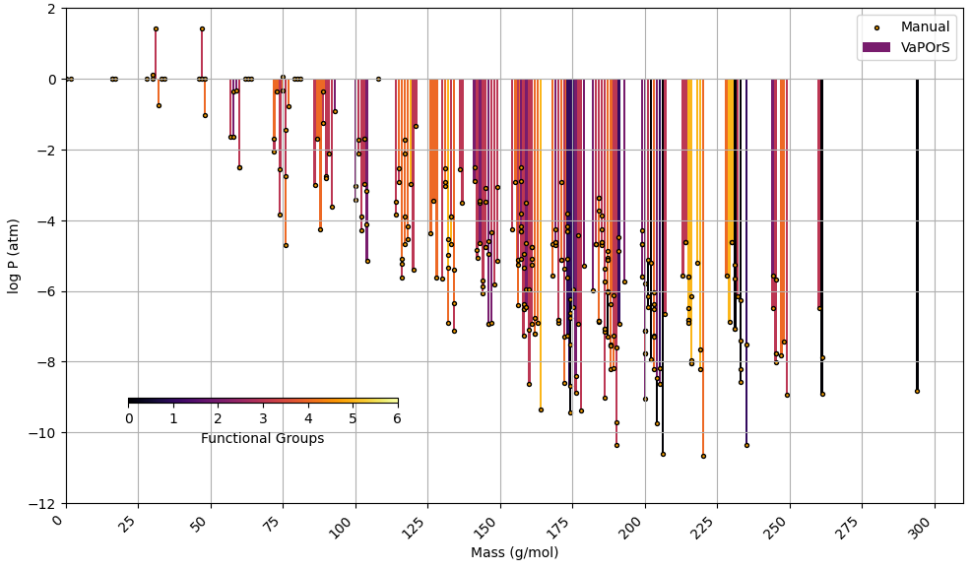


Figure 8. Comparison of saturation vapor pressure results for tropospheric degradation species of alpha-pinene. The figure presents the log saturation vapor pressure versus molar mass for species formed from the tropospheric oxidation of alpha-pinene according to MCM. Bars show results from the automated VaPOrS code, with colors based on detected functional group number involved in the chemical structure of species, and points reflect manually calculated values. The alignment of points atop bars demonstrates the perfect consistency between automated and manual calculations.

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

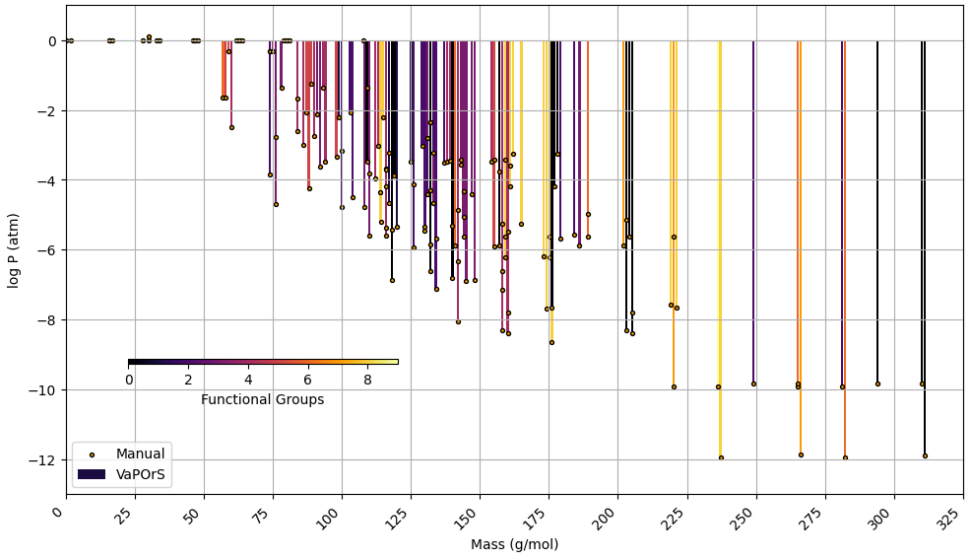


Figure 9. Comparison of saturation vapor pressure results for tropospheric degradation species of benzene. The figure illustrates the log saturation vapor pressure versus molar mass for species derived from benzene degradation according to MCM. The bars represent saturation vapor pressures calculated by the automated VaPOrS code, with colors based on detected functional group number involved in the chemical structure of species, while points indicate manually obtained values. The close alignment of points with bars highlights the accuracy of the automated method relative to manual calculations.

3.4.2. autoAPRAM-fw data

In the final stage of this analysis, the VaPOrS model was utilized to determine the saturation pressures of over 850 potential chemical species produced through the autoxidation of alkoxy and peroxy radicals, which emerge during benzene degradation. The initial radicals were defined by the MCM, with their respective saturation vapor pressures detailed in Figure 9. Conversely, the products of autoxidation were generated using the autoAPPRAM-fw tool (Pichelstorfer et al. 2024), with their potential structures represented by SMILES notation. Demonstrating its efficiency, VaPOrS analyzed all SMILES entries within a single second, accurately counting the required functional groups and calculating corresponding saturation vapor pressures. The results of these predictions are illustrated in Figure 10.

Figure 10 demonstrates the saturation vapor pressures achieved through the VaPOrS and compares them with their manually calculated counterparts. Automated results are shown as

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

colorful bars based on the number of detected functional groups reaching as high as 15 for some autoAPRAMfw products, and manual results are depicted as points. The compatibility of points atop bars for each species indicates excellent agreement between both approaches.

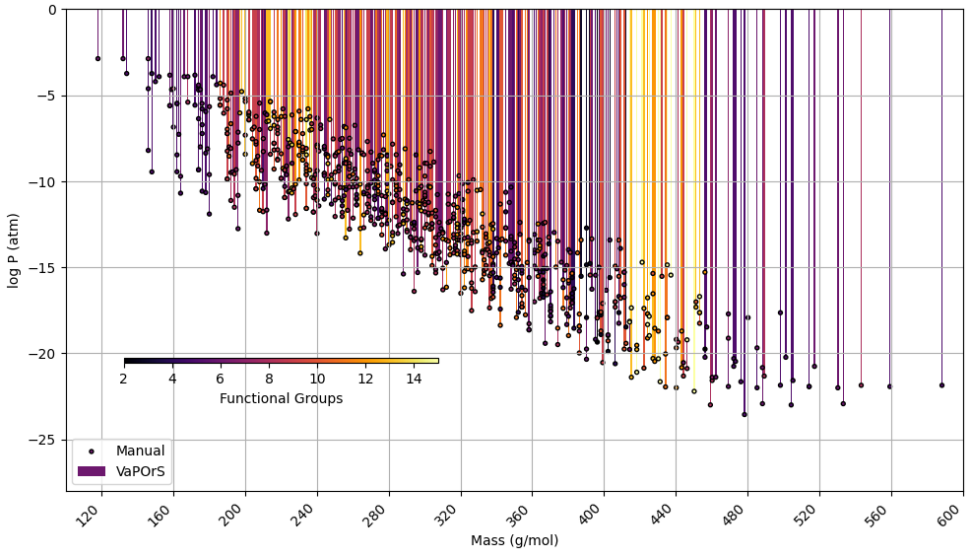


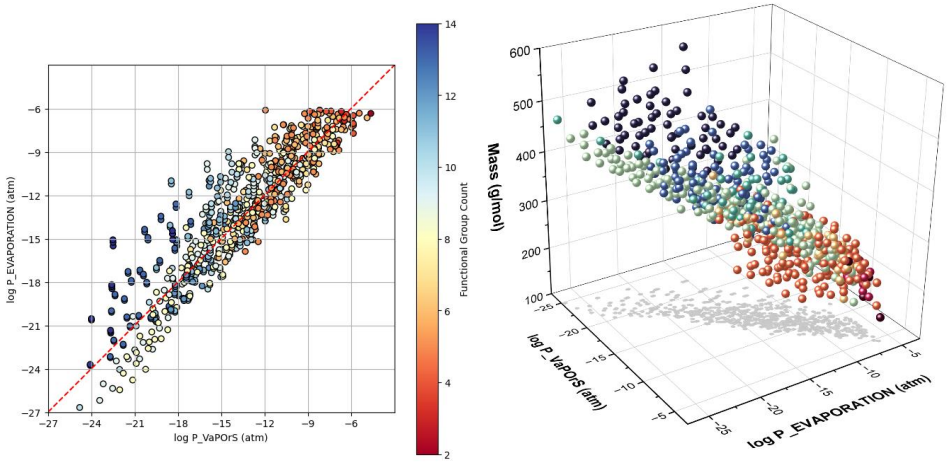
Figure 10. Comparison of saturation vapor pressure results for autoxidation species of benzene. The figure illustrates the log saturation vapor pressure versus molar mass for species derived from autoxidation of initial alkoxy and peroxy radicals of benzene degradation. The bars represent saturation vapor pressures calculated by the automated VaPOrS code, with colors based on detected functional group number involved in the chemical structure of species, while points indicate manually obtained values. The close alignment of points with bars highlights the accuracy of the automated method relative to manual calculations.

Figure 11 presents 2D and 3D scatter plots comparing the logarithmic saturation vapor pressures obtained from VaPOrS (log P_VaPOrS) with those predicted by the EVAPORATION (log P_EVAPORATION), Myrdal-Yalkowsky (log P_Myrdal_Yalkowsky), and Nanoolal (log P_Nanoolal) methods for autoxidation products. Each marker represents a compound, with color indicating the number of functional groups in its structure. The red dashed line in the 2D plots signifies the theoretical 1:1 relationship, where the saturation vapor pressures predicted by VaPOrS and the other models would be equivalent. Moreover, the 3D plots include the molar mass of species to give more details of the achieved results.

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

Further analysis shows that the current VaPOrS predictions based on SIMPOL parameterization align closely with those from the EVAPORATION method in the higher saturation vapor pressure range (approximately -6 to -15 on the logarithmic scale), with the EVAPORATION method tending to overestimate saturation vapor pressures for species with greater functional complexity. In contrast, VaPOrS demonstrates good agreement with the Nanoolal method in the lower saturation vapor pressure range (approximately -15 to -24), particularly for molecules with a high functional group count. The Myrdal-Yalkowsky method, however, totally overestimates saturation vapor pressures across the board compared to VaPOrS, with deviations increasing as functional group complexity rises. It is important to note that the current values are essentially SIMPOL-based predictions, so while these comparisons are informative, the general trends have been discussed in previous studies. However, as highlighted by (Isaacman-VanWertz and Aumont 2021), a combination of existing methods, potentially an average of them, has been suggested to yield the most reliable saturation vapor pressure estimates. Acknowledging this, a future refinement of the current approach could involve assessing whether incorporating such a hybrid method improves agreement with experimental data.

The VaPOrS model thus distinguishes itself by providing structure-based estimations for saturation vapor pressure predictions, offering a valuable tool for the assessment of complex organic compounds in atmospheric chemistry. This first application of the code utilized SIMPOL group contribution method for saturation vapor pressure and enthalpy of vaporization estimation, yet the code can be extended to work with any structure based thermodynamic property estimator, thereby streamlining work like secondary aerosol modelling considerably.



https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

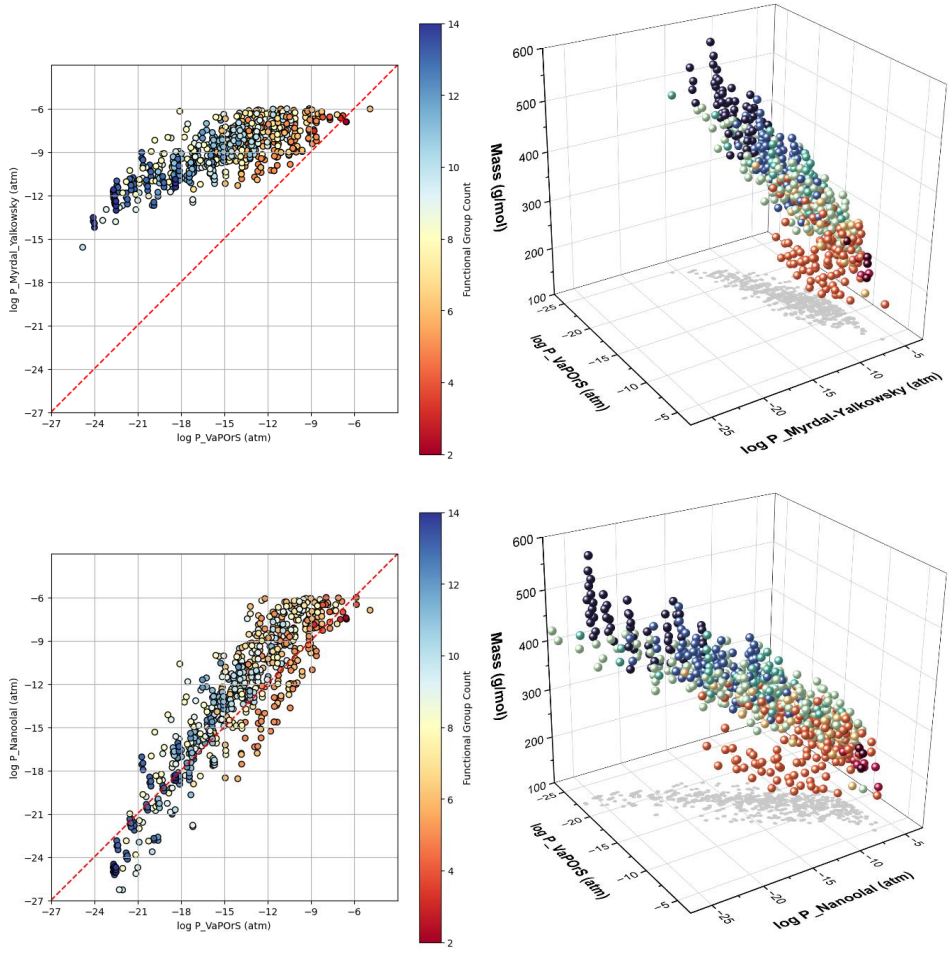


Figure 11. Comparison of logarithmic saturation vapor pressures predicted by the VaPOrS model (log P_VaPOrS) with those from the EVAPORATION (log P_EVAPORATION), Nanoolal (log P_Nanoolal), and Myrdal-Yalkowsky (log P_Myrdal_Yalkowsky) methods for autoxidation products derived from benzene degradation. Each data point represents a compound, with color indicating the functional group count. The red dashed line in the 2D plots represents the theoretical 1:1 relationship, where predictions from VaPOrS and other methods would be equivalent. The 3D plots include molar mass variation as well to give a comprehensive view of relationships between the parameters.

The versatility of VaPOrS lies in its potential for seamless integration into a range of widely used atmospheric models, enhancing their predictive capabilities. As seen, the Master Chemical Mechanism (MCM), with its comprehensive SMILES database of VOCs, can directly leverage VaPOrS to automate the generation of temperature-dependent saturation

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

vapor pressure equations. This automation ensures uniformity in saturation vapor pressure predictions, which is critical for large-scale atmospheric models to simulate chemical reactions and transport processes accurately. By providing a consistent and efficient means of handling VOC property data, VaPOrS enhances MCM's ability to support atmospheric chemistry research on gas-to-particle partitioning and secondary organic aerosol formation (Saunders et al. 2003; M. E. Jenkin et al. 2003).

In MEGAN (A. Guenther et al. 2006), VaPOrS can improve biogenic VOC emission estimates by rapidly providing saturation vapor pressure and enthalpy of vaporization data. Similarly, LOTOS-EUROS (Schaap et al. 2008) can benefit from more accurate secondary organic aerosol formation and gas-to-particle partitioning predictions. Global-scale models like GEOS-Chem (Bey et al. 2001) and WRF-Chem (Grell et al. 2005) are supported by VaPOrS's ability to process large datasets of VOCs efficiently, enabling more reliable simulations of atmospheric chemical processes. Regional models such as CMAQ (Byun and Schere 2006) are also enhanced, with VaPOrS contributing to improved particulate matter formation predictions.

Additionally, in SAPRC (Carter, n.d.) mechanisms, VaPOrS can automate the calculation of key VOC properties, streamlining computational workflows. Beyond these applications, specialized models like ADCHAM and ADCHEM (Roldin et al. 2014; 2011) can leverage VaPOrS to refine aerosol growth rate predictions and radiative forcing estimations, thus advancing studies on aerosol-cloud interactions and climate feedback mechanisms. These integrations highlight VaPOrS's broad applicability and its role in improving the accuracy and efficiency of atmospheric models, ultimately contributing to more informed strategies for air quality management and climate change mitigation.

It is worth mentioning that while group contribution methods, such as SIMPOL and VaPOrS, offer a reliable approach for estimating saturation vapor pressure by summing the contributions of individual functional groups, their accuracy tends to decrease as molecular complexity increases. This decline in predictive capability is particularly noticeable in highly functionalized compounds, where interactions between multiple functional groups may deviate from the assumed additive behavior. Previous studies have demonstrated that the presence of numerous functional groups does not always lead to a proportional reduction in saturation vapor pressure, as steric effects, intramolecular hydrogen bonding, and other molecular interactions can alter the expected contributions. This limitation highlights the need for caution when applying these models to large multifunctional molecules and suggests potential areas

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

for future refinement, such as incorporating correction factors or machine-learning approaches to account for non-additive effects.

4. Conclusion

This study presents a comprehensive analysis of the performance of the VaPOrS model, developed for identifying and quantifying functional groups in organic compounds with a specific focus on saturation vapor pressure calculations based on the SIMPOL method. The research highlights three critical aspects: the identification and counting of 30 structural groups recognized as influential on saturation vapor pressure, the subsequent calculation of saturation vapor pressure and enthalpy of vaporization for a range of organic compounds, and introduction of a temperature-dependent relationship for the saturation vapor pressure and enthalpy of vaporization. This study primarily utilizes the SIMPOL method for parameterization; however, the VaPOrS framework is inherently flexible and can be adapted to various other structure-based parameterization methods. These include approaches such as group additivity and volatility basis set (VBS) models. Furthermore, VaPOrS can be expanded to estimate a range of thermodynamic properties that depend on structural group-based parameterization, broadening its applicability beyond saturation vapor pressure alone. The validation process involved a meticulous comparison of manual and automated counts and calculations for over 1,000 organic compounds sourced from the Master Chemical Mechanism (MCM) database and recently introduced autoAPRAMfw autoxidation code. The perfect agreement demonstrates the accuracy for functional group identification and counting, with subsequent semi-automated saturation vapor pressure determination by SIMPOL through VaPOrS. This new methodology will be predominantly useful for researchers analyzing organic compounds, particularly in fields related to atmospheric chemistry and especially relating to aerosol formation. In conclusion, the findings from this research validate VaPOrS as a robust computational tool for estimating saturation vapor pressures while providing a systematic approach to functional group analysis. Its accuracy in functional group quantification and saturation vapor pressure prediction demonstrates substantial potential for advancing our understanding of organic compound behaviors in atmospheric and environmental applications. Future work will focus on enhancing the model's capabilities, exploring additional functional groups, and refining the saturation vapor pressure estimation model to further improve its applicability and precision in real-world scenarios.

Data availability

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

The raw data supporting the figures in the manuscript are openly available on Zenodo at: https://doi.org/10.5281/zenodo.15688105.

Code availability

The VaPOrS code used in this study is publicly available on Zenodo (Mojtaba Bezaatpour 2025). This archive includes a Jupyter notebook (VaPOrS.ipynb), a standalone Python script (VaPOrS.py), an input file containing SMILES strings (SMILES.txt), and example output files in both .txt and .csv formats. The repository is licensed under the MIT License and is fully open for use and redistribution under the conditions specified therein.

Author contributions

M.B. conceptualized the study and developed the Python code; M.B. and M.R. conducted the functional group analysis, validated the tool against existing data and contributed to data visualization; M.B. prepared the manuscript, and M.D.M. and M.R. reviewed and approved the final version of the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme under Grant No. 101002728 (ERC Consolidator grant ADAPT) and 101096133 (PAREMPI). This work is also funded by the Research Council of Finland (Grant Nos.: 331207, 336531, 346373, 353836). The AI-based tools were used for language editing to improve the readability of the manuscript.

References

Atkinson, R. 2000. "Atmospheric Chemistry of VOCs and NOx." *Atmospheric Environment* 34 (12–14): 2063–2101. https://doi.org/10.1016/S1352-2310(99)00460-4.

Berndt, Torsten, Stefanie Richters, Tuija Jokinen, Noora Hyttinen, Theo Kurtén, Rasmus V. Otkjær, Henrik G. Kjaergaard, et al. 2016. "Hydroxyl Radical-Induced Formation of Highly Oxidized Organic Compounds." *Nature Communications* 7 (1): 13677. https://doi.org/10.1038/ncomms13677.

Bey, Isabelle, Daniel J. Jacob, Robert M. Yantosca, Jennifer A. Logan, Brendan D. Field, Arlene M. Fiore, Qinbin Li, Honguy Y. Liu, Loretta J. Mickley, and Martin G. Schultz. 2001. "Global

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

Modeling of Tropospheric Chemistry with Assimilated Meteorology: Model Description and Evaluation." *Journal of Geophysical Research: Atmospheres* 106 (D19): 23073–95. https://doi.org/10.1029/2001JD000807.

Bianchi, Federico, Theo Kurtén, Matthieu Riva, Claudia Mohr, Matti P. Rissanen, Pontus Roldin, Torsten Berndt, et al. 2019. "Highly Oxygenated Organic Molecules (HOM) from Gas-Phase Autoxidation Involving Peroxy Radicals: A Key Contributor to Atmospheric Aerosol." *Chemical Reviews* 119 (6): 3472–3509. https://doi.org/10.1021/acs.chemrev.8b00395.

Byun, Daewon, and Kenneth L. Schere. 2006. "Review of the Governing Equations, Computational Algorithms, and Other Components of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System." *Applied Mechanics Reviews* 59 (2): 51–77. https://doi.org/10.1115/1.2128636.

Carter, William P L. n.d. "Documentation of the SAPRC-99 Chemical Mechanism for VOC Reactivity Assessment, Report to the California Air Resources Board". CA: University of California, Riverside, USA, 2000, https://intra.engr.ucr.edu/~carter/pubs/s99doc.pdf.

Compernolle, S., K. Ceulemans, and J.-F. Müller. 2011. "EVAPORATION: A New Vapour Pressure Estimation Methodfor Organic Molecules Including Non-Additivity and Intramolecular Interactions." *Atmospheric Chemistry and Physics* 11 (18): 9431–50. https://doi.org/10.5194/acp-11-9431-2011.

Crounse, John D., Lasse B. Nielsen, Solvejg Jørgensen, Henrik G. Kjaergaard, and Paul O. Wennberg. 2013. "Autoxidation of Organic Compounds in the Atmosphere." *The Journal of Physical Chemistry Letters* 4 (20): 3513–20. https://doi.org/10.1021/jz4019207.

Donahue, N. M., A. L. Robinson, C. O. Stanier, and S. N. Pandis. 2006. "Coupled Partitioning, Dilution, and Chemical Aging of Semivolatile Organics." *Environmental Science & Technology* 40 (8): 2635–43. https://doi.org/10.1021/es052297c.

Eckert, Frank, and Andreas Klamt. 2002. "Fast Solvent Screening via Quantum Chemistry: COSMO-RS Approach." *AIChE Journal* 48 (2): 369–85. https://doi.org/10.1002/aic.690480220.

Ehn, Mikael, Joel A. Thornton, Einhard Kleist, Mikko Sipilä, Heikki Junninen, Iida Pullinen, Monika Springer, et al. 2014. "A Large Source of Low-Volatility Secondary Organic Aerosol." *Nature* 506 (7489): 476–79. https://doi.org/10.1038/nature13032.

Epping, Ruben, and Matthias Koch. 2023. "On-Site Detection of Volatile Organic Compounds (VOCs)." *Molecules* 28 (4): 1598. https://doi.org/10.3390/molecules28041598.

Goldstein, Allen H., and Ian E. Galbally. 2007. "Known and Unexplored Organic Constituents in the Earth's Atmosphere." *Environmental Science & Technology* 41 (5): 1514–21. https://doi.org/10.1021/es072476p.

Grell, Georg A., Steven E. Peckham, Rainer Schmitz, Stuart A. McKeen, Gregory Frost, William C. Skamarock, and Brian Eder. 2005. "Fully Coupled 'Online' Chemistry within the WRF Model." *Atmospheric Environment* 39 (37): 6957–75. https://doi.org/10.1016/j.atmosenv.2005.04.027.

Guenther, A., T. Karl, P. Harley, C. Wiedinmyer, P. I. Palmer, and C. Geron. 2006. "Estimates of Global Terrestrial Isoprene Emissions Using MEGAN (Model of Emissions of Gases and Aerosols from Nature)." *Atmospheric Chemistry and Physics* 6 (11): 3181–3210. https://doi.org/10.5194/acp-6-3181-2006.

Guenther, Alex, C. Nicholas Hewitt, David Erickson, Ray Fall, Chris Geron, Tom Graedel, Peter Harley, et al. 1995. "A Global Model of Natural Volatile Organic Compound Emissions." *Journal of Geophysical Research: Atmospheres* 100 (D5): 8873–92. https://doi.org/10.1029/94JD02950.

Isaacman-VanWertz, Gabriel, and Bernard Aumont. 2021. "Impact of Organic Molecular Structure on the Estimation of Atmospherically Relevant Physicochemical Parameters." *Atmospheric Chemistry and Physics* 21 (8): 6541–63. https://doi.org/10.5194/acp-21-6541-2021.

Iyer, Siddharth. 2023. "Molecular Rearrangement of Bicyclic Peroxy Radicals: Key Route to Aerosol from Aromatics," February. https://doi.org/10.5281/ZENODO.8214481.

Iyer, Siddharth, Matti P. Rissanen, Rashid Valiev, Shawon Barua, Jordan E. Krechmer, Joel Thornton, Mikael Ehn, and Theo Kurtén. 2021a. "Molecular Mechanism for Rapid Autoxidation in α-Pinene Ozonolysis." *Nature Communications* 12 (1): 878. https://doi.org/10.1038/s41467-021-21172-w.

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

———. 2021b. “Molecular Mechanism for Rapid Autoxidation in α-Pinene Ozonolysis.” *Nature Communications* 12 (1): 878. https://doi.org/10.1038/s41467-021-21172-w.

Jenkin, M. E., S. M. Saunders, V. Wagner, and M. J. Pilling. 2003. “Protocol for the Development of the Master Chemical Mechanism, MCM v3 (Part B): Tropospheric Degradation of Aromatic Volatile Organic Compounds.” *Atmospheric Chemistry and Physics* 3 (1): 181–93. https://doi.org/10.5194/acp-3-181-2003.

Jenkin, Michael E., Richard Valorso, Bernard Aumont, and Andrew R. Rickard. 2019. “Estimation of Rate Coefficients and Branching Ratios for Reactions of Organic Peroxy Radicals for Use in Automated Mechanism Construction.” *Atmospheric Chemistry and Physics* 19 (11): 7691–7717. https://doi.org/10.5194/acp-19-7691-2019.

Jimenez, J. L., M. R. Canagaratna, N. M. Donahue, A. S. H. Prevot, Q. Zhang, J. H. Kroll, P. F. DeCarlo, et al. 2009. “Evolution of Organic Aerosols in the Atmosphere.” *Science* 326 (5959): 1525–29. https://doi.org/10.1126/science.1180353.

Joback, K.G., and R.C. Reid. 1987. “ESTIMATION OF PURE-COMPONENT PROPERTIES FROM GROUP-CONTRIBUTIONS.” *Chemical Engineering Communications* 57 (1–6): 233–43. https://doi.org/10.1080/00986448708960487.

Klamt, Andreas. 1995. “Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena.” *The Journal of Physical Chemistry* 99 (7): 2224–35. https://doi.org/10.1021/j100007a062.

Klamt, Andreas, Volker Jonas, Thorsten Bürger, and John C. W. Lohrenz. 1998. “Refinement and Parametrization of COSMO-RS.” *The Journal of Physical Chemistry A* 102 (26): 5074–85. https://doi.org/10.1021/jp980017s.

Luo, Hao, Luc Vereecken, Hongru Shen, Sungah Kang, Iida Pullinen, Mattias Hallquist, Hendrik Fuchs, et al. 2023. “Formation of Highly Oxygenated Organic Molecules from the Oxidation of Limonene by OH Radical: Significant Contribution of H-Abstraction Pathway.” *Atmospheric Chemistry and Physics* 23 (13): 7297–7319. https://doi.org/10.5194/acp-23-7297-2023.

Mattila, Timo, Markku Kulmala, and Timo Vesala. 1997. “On the Condensational Growth of a Multicomponent Droplet.” *Journal of Aerosol Science* 28 (4): 553–64. https://doi.org/10.1016/S0021-8502(96)00458-2.

McDonald, Brian C., Joost A. De Gouw, Jessica B. Gilman, Shantanu H. Jathar, Ali Akherati, Christopher D. Cappa, Jose L. Jimenez, et al. 2018. “Volatile Chemical Products Emerging as Largest Petrochemical Source of Urban Organic Emissions.” *Science* 359 (6377): 760–64. https://doi.org/10.1126/science.aaq0524.

Mellouki, A., T. J. Wallington, and J. Chen. 2015. “Atmospheric Chemistry of Oxygenated Volatile Organic Compounds: Impacts on Air Quality and Climate.” *Chemical Reviews* 115 (10): 3984–4014. https://doi.org/10.1021/cr500549n.

Mojtaba Bezaatpour. 2025. “Mojtababzp/VaPOrS: VaPOrS v1.0.1.” Zenodo. https://doi.org/10.5281/ZENODO.15222175.

Myrdal, Paul B., and Samuel H. Yalkowsky. 1997. “Estimating Pure Component Vapor Pressures of Complex Organic Molecules.” *Industrial & Engineering Chemistry Research* 36 (6): 2494–99. https://doi.org/10.1021/ie9502421.

Nannoolal, Yash, Jürgen Rarey, and Deresh Ramjugernath. 2008. “Estimation of Pure Component Properties.” *Fluid Phase Equilibria* 269 (1–2): 117–33. https://doi.org/10.1016/j.fluid.2008.04.020.

Pankow, J. F., and W. E. Asher. 2008. “SIMPOL.1: A Simple Group Contribution Method for Predicting Vapor Pressures and Enthalpies of Vaporization of Multifunctional Organic Compounds.” *Atmospheric Chemistry and Physics* 8 (10): 2773–96. https://doi.org/10.5194/acp-8-2773-2008.

Pichelstorfer, Lukas, Pontus Roldin, Matti Rissanen, Noora Hyttinen, Olga Garmash, Carlton Xavier, Putian Zhou, et al. 2024. “Towards Automated Inclusion of Autoxidation Chemistry in Models: From Precursors to Atmospheric Implications.” *Environmental Science: Atmospheres* 4 (8): 879–96. https://doi.org/10.1039/D4EA00054D.

Rissanen, Matti P., Theo Kurtén, Mikko Sipilä, Joel A. Thornton, Juha Kangasluoma, Nina Sarnela, Heikki Junninen, et al. 2014. “The Formation of Highly Oxidized Multifunctional Products in

https://doi.org/10.5194/egusphere-2025-2564
Preprint. Discussion started: 26 June 2025

the Ozonolysis of Cyclohexene." *Journal of the American Chemical Society* 136 (44): 15596–606. https://doi.org/10.1021/ja507146s.

Roldin, P., A. C. Eriksson, E. Z. Nordin, E. Hermansson, D. Mogensen, A. Rusanen, M. Boy, et al. 2014. "Modelling Non-Equilibrium Secondary Organic Aerosol Formation and Evaporation with the Aerosol Dynamics, Gas- and Particle-Phase Chemistry Kinetic Multilayer Model ADCHAM." *Atmospheric Chemistry and Physics* 14 (15): 7953–93. https://doi.org/10.5194/acp-14-7953-2014.

Roldin, P., E. Swietlicki, G. Schurgers, A. Arneth, K. E. J. Lehtinen, M. Boy, and M. Kulmala. 2011. "Development and Evaluation of the Aerosol Dynamics and Gas Phase Chemistry Model ADCHEM." *Atmospheric Chemistry and Physics* 11 (12): 5867–96. https://doi.org/10.5194/acp-11-5867-2011.

Saunders, S. M., M. E. Jenkin, R. G. Derwent, and M. J. Pilling. 2003. "Protocol for the Development of the Master Chemical Mechanism, MCM v3 (Part A): Tropospheric Degradation of Non-Aromatic Volatile Organic Compounds." *Atmospheric Chemistry and Physics* 3 (1): 161–80. https://doi.org/10.5194/acp-3-161-2003.

Schaap, Martijn, Renske M.A. Timmermans, Michiel Roemer, G.A.C. Boersen, Peter J.H. Builtjes, Ferd J. Sauter, Guus J.M. Velders, and Jeanette P. Beck. 2008. "The LOTOS EUROS Model: Description, Validation and Latest Developments." *International Journal of Environment and Pollution* 32 (2): 270. https://doi.org/10.1504/IJEP.2008.017106.

Topping, David, Mark Barley, Michael K. Bane, Nicholas Higham, Bernard Aumont, Nicholas Dingle, and Gordon McFiggans. 2016. "UManSysProp v1.0: An Online and Open-Source Facility for Molecular Property Prediction and Atmospheric Aerosol Calculations." *Geoscientific Model Development* 9 (2): 899–914. https://doi.org/10.5194/gmd-9-899-2016.

Vereecken, L., B. Aumont, I. Barnes, J.W. Bozzelli, M.J. Goldman, W.H. Green, S. Madronich, et al. 2018. "Perspective on Mechanism Development and Structure-Activity Relationships for Gas-Phase Atmospheric Chemistry." *International Journal of Chemical Kinetics* 50 (6): 435–69. https://doi.org/10.1002/kin.21172.

WHO. 2005. "Air Quality Guidelines: Global Update 2005. Geneva: World Health Organization."

Zhao, Defeng, Iida Pullinen, Hendrik Fuchs, Stephanie Schrade, Rongrong Wu, Ismail-Hakki Acir, Ralf Tillmann, et al. 2021. "Highly Oxygenated Organic Molecule (HOM) Formation in the Isoprene Oxidation by NO₃ Radical." *Atmospheric Chemistry and Physics* 21 (12): 9681–9704. https://doi.org/10.5194/acp-21-9681-2021.