Journal of Aerosol Science 179 (2024) 106375

Contents lists available at ScienceDirect

Journal of Aerosol Science

journal homepage: www.elsevier.com/locate/jaerosci

The search for sparse data in molecular datasets: Application of active learning to identify extremely low volatile organic compounds

Vitus Besel [a,*], Milica Todorović [b], Theo Kurtén [a], Hanna Vehkamäki [a], Patrick Rinke [c]

[a] Institute for Atmospheric and Earth System Research, Gustaf Hällströmin katu 2, Helsinki, 00014, Finland
[b] Department of Mechanical and Materials Engineering, Vesilinnantie 5, Turku, 20014, Finland
[c] Department of Applied Physics, P.O. Box 11100, Espoo, 00076, Finland

ARTICLE INFO

Dataset link: https://doi.org/10.23729/dd0396b3-9017-40f2-ae4b-6876bf33dd08

Keywords:
Low volatile organic compounds
Machine learning
Oxygenated organic molecules
Particle formation
Molecular data

ABSTRACT

The formation of aerosol particles in the atmosphere is driven by the gas to particle conversion of *extremely low volatile organic compounds* (ELVOC), organic compounds with a particularly low saturation vapor pressure (p_{Sat}). Identifying ELVOCs and their chemical structures is both experimentally and theoretically challenging: Measuring the very low p_{Sat} of ELVOCs is extremely difficult, and computing p_{Sat} for these often large molecules is computationally costly. Moreover, ELVOCs are underrepresented in available datasets of atmospheric organic species, which reduces the value of statistical models built on such data. We propose an active learning (AL) approach to efficiently identify ELVOCs in a data pool of atmospheric organic species with initially unknown p_{Sat}. We assess the performance of our AL approach by comparing it to traditional machine learning regression methods, as well as ELVOC classification based on molecular properties. AL proves to be a highly efficient method for ELVOC identification with limitations on the type of ELVOC it can identify. We also show that traditional machine learning or molecular property-based methods can be adequate tools depending on the available data and desired degree of efficiency.

1. Introduction

Secondary Organic Aerosols (SOA) play a major role in atmospheric chemistry and physics. They reflect and scatter solar radiation, act as cloud condensation nuclei, and are a source of large uncertainties in current climate models (Arias et al., 2021). SOA formation is driven by gas-to-particle conversion of *oxygenated organic molecules* (OOMs) (Kerminen, Chen, Vakkari, Petäjä, Kulmala, & Bianchi, 2018; Kupc, Williamson, Hodshire, Kazil, Ray, Bui, Dollner, Froyd, McKain, Rollins, Schill, Thames, Weinzierl, Pierce, & Brock, 2020; Metzger et al., 2010-04-13; Yan et al., 2016; Zhang et al., 2004). A myriad of different OOMs can be found in the atmosphere as products of oxidation chains of organic molecules. Particularly interesting for SOA formation are OOMs that are so low in volatility that they inevitably condense in ambient conditions, even in the absence of pre-existing surfaces. Such OOMs are commonly referred to as *extremely low volatile organic compounds* (ELVOCs) (Donahue, Kroll, Pandis, & Robinson, 2012; Schervish & Donahue, 2020).

The volatility of a compound or its affinity to the condensed phase, and thus its potential to participate in particle formation, can be expressed by its saturation vapor pressure (p_{Sat}) or equivalently saturation mass concentration. The dependence of p_{Sat} on the

* Corresponding author.
E-mail address: vitus.besel@helsinki.fi (V. Besel).

https://doi.org/10.1016/j.jaerosci.2024.106375

Available online 30 March 2024
0021-8502/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

chemical and molecular structure of OOMs and ELVOCs helps us to better understand the origin and chemistry of molecules involved in organic particle formation in the atmosphere. However, p_{Sat} measurements of ELVOCs are complicated by the low gas phase concentrations and by the fact that the molecules easily condense onto the measuring device due to their low volatility. The number of successful measurements (Hyttinen et al., 2022) and thus the number of identified ELVOCs is therefore still small compared to the millions of different OOMs in the atmosphere. The lack of experimental data on the p_{Sat} of polyfunctional ELVOCs also means that empirical structure–property relationships for predicting p_{Sat} are unlikely to be reliable for these species, despite performing very well for chemically simpler and more volatile compounds. Chamber experiments (Hyttinen et al., 2022) can treat larger OOM numbers and, *e.g.*, recently suggested that nucleation of α-pinene, isoprene and β-caryophyllene oxidation products is driven by molecules with 15–20 carbon atoms (Dada et al., 2023). Another study found that relative humidity does not affect the ELVOC partitioning to the particle phase (Surdu et al., 2023). Yet, beyond probing overall condensation behavior, chamber experiments do not reveal the chemical identity or molecular structure of involved OOM. Quantum chemistry provides a direct route to calculate p_{Sat} from the molecular structure, (Kurtén, Hyttinen, D'Ambro, Thornton, & Prisle, 2018) without empirical system-specific parameters. Such calculations are computationally very costly, but have been applied recently to generate moderately sized OOM datasets (Besel, Todorović, Kurtén, Rinke, & Vehkamäki, 2023; Tabor et al., 2019).

A viable strategy to first identify and then study ELVOCs is to utilize these emerging OOM datasets (Besel et al., 2023; Wang et al., 2017). We have recently shown that machine learning (ML) models trained on OOM datasets can accurately predict the p_{Sat} of organic molecules from their geometric structure (Besel et al., 2023; Lumiaro, Todorović, Kurten, Vehkamäki, & Rinke, 2021-09-06). We could therefore envision applying such ML models to a vast pool of OOMs to identify ELVOCs by their predicted p_{Sat}. This strategy, although appealing, faces several challenges. (1) Current datasets are limited in size due to the aforementioned experimental limitations and the computational cost of quantum chemistry calculations. (2) ELVOCs are likely underrepresented in standard datasets, because the molecular generators that simulate OOM oxidation are truncated too early (Isaacman-VanWertz & Aumont, 2021) or lack key mechanisms such as accretion and autoxidation reactions. (3) Regression or classification models trained on datasets with scarcely represented ELVOCs might not be very accurate for ELVOC identification, as the predictive accuracy of machine learning models usually correlates with the presence of corresponding data. A direct consequence is that (4) OOM data sets would have to be large enough to feature enough ELVOCs for ML training, which could strain computational or experimental budgets.

In this article, we address these challenges by designing efficient ML strategies for the classification of OOMs as "ELVOCs" or "non-ELVOCs" merely based on their chemical structure. We use a large set of 157k OOMs (Isaacman-VanWertz & Aumont, 2021) with known chemical and molecular structure, but unknown p_{Sat}, *i.e.* a large set of unlabeled molecules.[1] Our objective is then to identify ML strategies that minimize the computational cost associated with labeling, that is, with computing many p_{Sat}. We will contrast two different strategies: (i) a large model trained on 24k molecules of the 157k OOMs, for which we have already computed the p_{Sat} (the GeckoQ dataset (Besel et al., 2023)); (ii) an active learning (AL) model that is initially trained on only 500 labeled molecules and then iteratively refined to target ELVOCs. Strategy (i) comes with a high initial cost, but might deliver a well-balanced ML model with sufficient classification accuracy. Strategy (ii) combines supervised and unsupervised machine learning to target ELVOCs, leading to a low initial cost, but might not deliver a very predictive ML model. In this work, we will assess the performance of both strategies in terms of the total number of correctly identified ELVOCs and the associated computational cost, and compare our ML methods to ELVOC classification with the empirical SIMPOL method (Pankow & Asher, 2008-05-19) and a rule-based approach.

The article is structured as follows. After a review of the applied methods and data, we introduce four performance measures. Thereafter, we compare our methods with respect to the performance measures, analyze AL in depth, and eventually, discuss the merits of the different methods in different scenarios of data availability.

2. Materials and methods

Below we summarize the methods used in this work. We first provide a robust definition of "ELVOC", and then briefly review quantum chemical p_{Sat} calculations. We also describe the employed dataset, introduce our ML approaches and discuss model training. We, furthermore, explain the identification of molecular functional groups and how these are used to compute p_{Sat} with the SIMPOL group contribution method or in a rule-based p_{Sat} classification method. Finally, we introduce four performance measures for comparing the different methods.

2.1. ELVOC definition and p_{Sat} computation

The term "ELVOC" is context-dependent and not unambiguously defined in the atmospheric science community (Bianchi et al., 2019). Previous work set the limit for the ELVOC saturation mass concentration to $3 \cdot 10^{-4}\ \frac{\mu g}{m^3}$ (Donahue et al., 2012). This limit was later reduced by a factor of ten, (Tröstl, Chuang, Gordon, Heinritzi, Yan, Molteni, Ahlm, Frege, Bianchi, Wagner, Simon, Lehtipalo, Williamson, Craven, Duplissy, Adamov, Almeida, Bernhammer, Breitenlechner, Brilke, Dias, Ehrhart, Flagan, Franchin, Fuchs, Guida, Gysel, Hansel, Hoyle, Jokinen, Junninen, Kangasluoma, Keskinen, Kim, Krapf, Kürten, Laaksonen, Lawler, Leiminger, Mathot,

[1] Generative algorithms that have been successfully used for targeted molecular discovery in other contexts (Westermayr, Gilkes, Barrett, & Maurer, 2023) will not be pursued in this work.

Möhler, Nieminen, Onnela, Petäjä, Piel, Miettinen, Rissanen, Rondo, Sarnela, Schobesberger, Sengupta, Sipilä, Smith, Steiner, Tomè, Virtanen, Wagner, Weingartner, Wimmer, Winkler, Ye, Carslaw, Curtius, Dommen, Kirkby, Kulmala, Riipinen, Worsnop, Donahue, & Baltensperger, 2016) and the term Ultra Low VOC (ULVOC) was coined (Schervish & Donahue, 2020). Because the relation between saturation vapor pressure and saturation mass concentration is molecular weight-dependent, we chose a p_{Sat} threshold instead of saturation mass concentration to avoid any ambiguity in ELVOC definitions and to simplify the analysis of our results. A saturation mass concentration of $3 \cdot 10^{-4}\ \frac{\mu g}{m^3}$ corresponds to a p_{Sat} of $2.66 \cdot 10^{-8}$ Pa for the lowest molecular weight (28 g/mol) present in the GeckoQ data set at 298.15 K. We rounded this up to a p_{Sat} threshold of $3 \cdot 10^{-8}$ Pa. By basing the threshold on the lowest molecular weight, rather than on a larger one, we obtain a comparably higher ELVOC threshold. This choice is balanced by the fact that p_{Sat} is generally temperature dependent and was computed for 298.15 K in our work. In real ambient conditions, the temperature is often lower, resulting in a lower p_{Sat} for all compounds.

In the following sections we will refer to any molecules with a p_{Sat} below the threshold of $3 \cdot 10^{-8}$ Pa as "ELVOC" and above as "non-ELVOC", for simplicity. To determine if a molecule is an "ELVOC", we computed its p_{Sat} as follows: first, an initial conformer-search was conducted with COSMO*conf* (Dassault Systèmes, 2022). Then the gas-phase ("vacuum") and liquid-phase energies were computed for all conformers. For the latter, we employed the conductor-like screening model for real solvents (COSMO-RS), (Klamt, Jonas, Bürger, & Lohrenz, 1998; Klamt & Schüürmann, 1993) a continuum solvation model. All energy calculations were carried out with density functional theory (B88-PW86 functional, (Becke, 1988; Perdew, 1986) TZVPD basis, using multipole accelerated RI-approximation) as implemented in Turbomole (Balasubramani et al., 2020). Next, we selected up to 40 of the energetically lowest conformers with a minimal number of intramolecular H-bonds, because it has been demonstrated that the inclusion of conformers with many internal H-bonds leads to larger errors in COSMO predictions (Kurtén et al., 2018). Finally, we computed the p_{Sat} based on the chosen conformers with COSMO*therm*(Klamt et al., 1998; Klamt & Schüürmann, 1993) for a standard temperature of 298.15 K. This p_{Sat}-computation workflow has been described in more detail in our previous work (Besel et al., 2023). COSMO-RS has been ascribed an accuracy within 0.5 log(MAE/Pa) relative to measured p_{Sat} (Eckert & Klamt, 2002). Neither the COSMO-RS accuracy nor the chosen temperature have an impact on the method comparison below, as long as all reference p_{Sat} have been computed consistently with the same quantum chemistry methodology. This study does not work with measured reference pressures, because, as noted in the introduction, currently it is not possible to measure them for ELVOC in high quantity and this kind of data is not available. In the following, we will refer to the quantum-chemical computation of a molecule's p_{Sat} simply as *labeling*.

2.2. Dataset

For the development of our ELVOC search procedures, we selected a dataset of 157,395 atmospheric organic molecules (containing C, H, O, and N; size range 4 – 45 atoms per molecule), generated by the chemical mechanism GECKO-A (Aumont, Szopa, & Madronich, 2005-09-22; Isaacman-VanWertz & Aumont, 2021) and post-processed as detailed in previous work (Besel et al., 2023). We will refer to the 157k molecules as the *raw Gecko data*. The aforementioned GeckoQ dataset is a labeled subset of the raw Gecko data, in which the p_{Sat} has been computed for 31,637 randomly chosen molecules following the procedure outlined in the previous section.

Inspection of the computed p_{Sat} values reveals that GeckoQ contains 1,608 ELVOCs, *i.e.* 5.1% of all the 31,637 molecules (cf. Fig. 1). That roughly corresponds to the 8% ELVOC yielded by the analysis of OOM field measurements (Zheng et al., 2023) with the group contribution method SIMPOL (Pankow & Asher, 2008-05-19). Since GeckoQ has been uniformly sampled from the raw Gecko data, we can apply the percentage of 5.1% to the raw Gecko data to estimate that 8,027 molecules in the raw data should be ELVOCs. Subtracting the 1,608 ELVOCs we already identified in GeckoQ, we are left with approximately 6,419 ELVOCs in the raw Gecko data that have not yet been found (*i.e.* not labeled as ELVOC). It is our objective in this work to find as many of these 6,419 ELVOCs as efficiently as possible.

In this work, we introduced a hard decision boundary for ELVOC classification, although in reality p_{Sat} distributes continuously, as Fig. 4 illustrates. A small p_{Sat} prediction error could therefore shift an ELVOC close to the threshold into the non-ELVOC region and vice versa. We assume that such mis-classification errors occur with the same frequency for false positives and false negatives and will therefore average out.

2.3. Machine learning approach

2.3.1. Regression model

As the p_{Sat} distribution of the data is continuous, we opted for a regression-based machine learning technique for ELVOC identification. With the regression model, we will make predictions for p_{Sat} upon which we distinguish between ELVOCs and non-ELVOCs. This enables a comparison between the p_{Sat} and corresponding predictions, which facilitates the interpretation of our results.

We used Gaussian process regression (GPR) as implemented in PyTorch for the p_{Sat} predictions. The employed product kernel was composed of the radial basis function (RBF) kernel and a multiplicative constant. For the GPR fitting, molecules need to be represented in a machine readable format, a so-called *descriptor* (Himanen et al., 2020; Langer, Goeßmann, & Rupp, 2022). We chose the topological fingerprint (TopFP) (James & Weininger, 1995; Landrum, Tosco, Kelley, Ric, Cosgrove, Sriniker, Gedeck, Vianello, NadineSchneider, Kawashima, N, Jones, Dalke, Cole, Swain, Turk, AlexanderSavelyev, Vaucher, Wójcikowski, Take, Probst, Ujihara, Scalfani, Godin, Lehtivarjo, Pahl, Walker, Berenger, Jasondbiggs, & Strets123, 2023) descriptor, in accordance with previous work (Besel et al., 2023; Lumiaro et al., 2021-09-06). GPR model choices and TopFP hyperparameters are discussed in the SI.

We monitored ML model performance with the mean absolute error (MAE) calculated for a test set, for which we randomly picked 2000 molecules from the raw Gecko data prior to any ELVOC search and computed their p_{Sat} with quantum chemistry. This test set was kept fixed and used throughout our whole study.

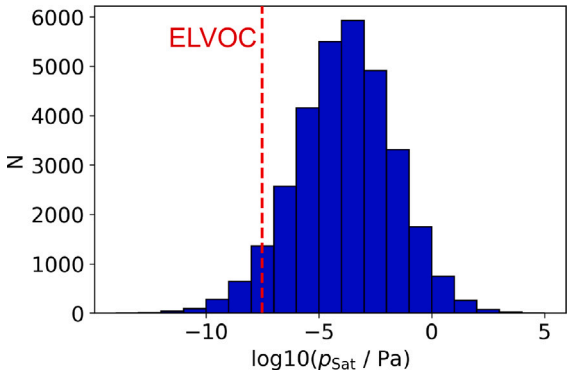

Fig. 1. p_{Sat} distribution in the GeckoQ data. 5.1% of all molecules are classified as ELVOC with the saturation vapor pressure threshold of $3 \cdot 10^{-8}$ Pa applied in this work.

Table 1
An overview of all applied machine learning classification strategies. "5x" indicates that five different sets of training sets were drawn and utilized. The trainingset of AL was extended each AL cycle.

Method	Abbreviation	Held-out set	Trainingset size
General single	GS	126k	24k
General ensemble	GE	126k	5x 24k
Active Learning	AL	157k	500–3000

2.3.2. ML ELVOC classification strategies

As alluded to in the introduction, we distinguish between two different regression types for our classification task (summarized in Table 1): (1) static global models that are trained once on a large training set and (2) active learning models that evolve iteratively from a small, initial training set. We will compare these two strategies against two empirical models: (3) the group contribution method SIMPOL (Pankow & Asher, 2008-05-19) and (4) a simple rule-based approach. All strategies are briefly described in the following.

Global models – We trained a ML model on 24k OOMs randomly chosen from the GeckoQ data. ELVOC classification with this model is referred to as the *GPR single* (GS) method. In addition, we build an ensemble, by training four additional ML models on different sets of 24k randomly chosen OOMs from the GeckoQ data. As the training sets are chosen from the same pool of overall 32k molecules, a training set size of 24k ensures that the models are not identical but incorporate some variability. The four models were combined with the GS model into an ensemble model termed *GPR ensemble* (GE).

ELVOC classification with the GS and GE was conducted as follows: We created a held-out set that consisted of all remaining unlabeled molecules in the raw Gecko data (*i.e.*, 125,758 molecules that were not in the test set or used for GS and GE training). For GS, we predicted the p_{Sat} of all held-out set molecules and classified those as ELVOC that fell below our p_{Sat} threshold. For GE, we repeated the procedure for the remaining four models in the ensemble. A molecule was then classified as ELVOC, if at least one of the five models predicted it as such. We then calculated the p_{Sat} of all molecules classified as ELVOC by GS and GE to check our predictions.

Active learning – Our AL approach proceeds iteratively and is depicted schematically in Fig. 2. Starting with an initial batch of labeled molecules and a GPR trained on this batch, further batches are selected iteratively from the held-out set and added to the training set. When a new batch has been added, we retrain the GPR.

Like for the GS and GE, we predicted the p_{Sat} of held-out set molecules with the GPR at each iteration. The held-out set comprises the whole raw Gecko data set minus the test set and the molecules in the 0th batch. Molecules below the ELVOC threshold were then clustered into 520 clusters to maximize molecular diversity. From each cluster we picked the molecule closest to the centroid and computed their p_{Sat} with quantum chemistry. These molecules were then added to the batch. The overhead of 20 molecules accounted for potential p_{Sat} calculation failures (*e.g.* non-convergence). In practice the number of failures never exceeded 20 and we randomly picked 500 molecules out of the successfully calculated ones.

For reference, we also executed the workflow with batches randomly picked (RND) from the held-out set to compute a baseline against which to measure potential active learning benefits. Both AL and RND were stopped after six iterations at which point 3000 molecules had been assembled, because after iteration 6 the AL model identified fewer than 520 new ELVOCs (cf. Table 3).

SIMPOL group contribution method – Functional groups (FG) are major defining factors for the chemical behavior of a molecule, and thus also its p_{Sat}. FG that are able to form strong intermolecular interactions, such as hydrogen-bonds (H-bonds), lower p_{Sat}, because the molecules are more strongly bound in the liquid phase. For example hydroxy, carboxylic acid, or hydroperoxide FGs, contain H-bond donors and H-bond acceptors, and can form strong intermolecular interactions by themselves and are typically abundant in low p_{Sat} molecules. Other FGs such as aldehydes, ketones, peroxides and esters can only act as H-bond acceptors and

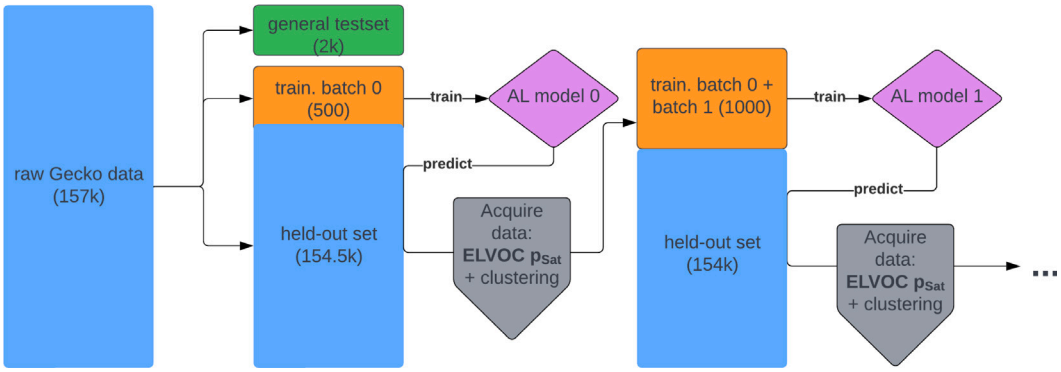

Fig. 2. The active learning scheme.

decrease p_{Sat} less than the groups that also include an H-bond donor. Finally, the nitrogen containing FGs, nitrate- and nitrogroups, have the smallest impact on p_{Sat} decrease.

With the FG contribution method SIMPOL (Pankow & Asher, 2008-05-19) the log10 of a p_{Sat} in Pa is calculated as

$$\log_{10} p_{Sat} = \sum_k \nu_k b_k, \tag{1}$$

where ν_k is the number of FGs of type k. The expansion coefficients b_k were fitted to reference data (Pankow & Asher, 2008-05-19). We chose SIMPOL as a computationally cheap, empirical reference method. We used our own Matlab (The MathWorks Inc., 2022) SIMPOL implementation applying the fitted SIMPOL coefficients. The FGs were obtained with the APRL Substructure Search Program (APRL-SSP), (Ruggeri & Takahama, 2016) which includes the most relevant FGs. APRL-SSP presently does not correctly identify carbonyl groups attached to a carbon that is also attached to a peroxy group and we corrected for this.

Since SIMPOL is already parameterized, we do not need to train it on labeled data. We then tested the SIMPOL ELVOC classification performance by applying SIMPOL to 42,156 molecules from the Gecko dataset (*i.e.*, the 32k GeckoQ molecules + 10k additional molecules that have been labeled over the course of this study). Our objective is not to draw a comparison between SIMPOL p_{Sat} and COSMO-RS p_{Sat}, as this has been done before Besel et al. (2023), Hyttinen et al. (2022, 2021).

Rule-based approach – During this work, we identified the molecular weight (MW; in g/mol), the average oxidation state ($\overline{OS}_C$) and specific FGs as valid indicators for p_{Sat} and decided to test their potential for ELVOC classification.

Fig. 3 displays the $\overline{OS}_C$– and MW–p_{Sat} relationships for 42,156 molecules from the Gecko dataset. The green curve shows molecules with fewer than three nitrogen atoms and the orange curve molecules with no nitrogen ("no N"). Fig. 3(a) confirms what is already known: p_{Sat} generally decreases with increasing $\overline{OS}_C$/MW, as more functional groups are introduced that can form hydrogen bonds and other types of intermolecular dipole–dipole interactions. In addition, Fig. 3 confirms that nitrogen containing groups have little influence on p_{Sat}, because the p_{Sat}–$\overline{OS}_C$/MW slope becomes more pronounced if it is only plotted for molecules without nitrogen-containing groups.

We further analyzed the FG distribution in Fig. 3(c). For this analysis, we chose the five most frequent (not nitrogen containing) FGs in GeckoQ (ketone, hydroxy, hydroperoxide, carboxylic acid, aldehyde groups) and distinguished between ELVOCs and non-ELVOCs. Fig. 3(c) demonstrates that hydroperoxide, carboxylic acid, and hydroxy groups are more frequent in ELVOCs. This is expected, since these groups lead to lower p_{Sat}. In contrast, ketone and aldehyde groups are more frequent in non-ELVOCs.

From the above considerations, we derived our rule-based classification. First, a molecule needs to have $\overline{OS}_C > 1.25$ and MW > 236 g/mol. If this is true, we check, if the molecule has more than one FG of the types hydroxy, hydroperoxide, or carboxylic acid. If this is also true, we classify the molecule as ELVOC. This approach will be abbreviated as "RULE" in the following.

2.4. Performance measures

For the comparison of our classification strategies, we apply four performance measures: *identification accuracy*, *identification cost*, *classification accuracy* and *prediction accuracy*.

- The identification accuracy quantifies how many ELVOCs were correctly identified. The target values (the number of expected ELVOC) differ slightly, because the different models are applied to different data volumes. For each method, we count the number of correctly identified ELVOCs and divide by the corresponding target reported in Table 2 to express the identification accuracy in percent.
- The identification costs measures the resource requirements. In our study, labeling is the most resource intensive step. The quantum mechanical p_{Sat} calculations take time and require computational resources (*e.g.*, computing hours on high-performance computing infrastructures). The resource requirements for each p_{Sat} calculation depend on the size and complexity of each molecule. To simplify, we assume a constant, molecule-independent time and computation cost, where the cost of determining X p_{Sat} values with quantum chemistry costs X. The labeling and thus the classification cost then becomes proportional to the number of molecules that need to be labeled.

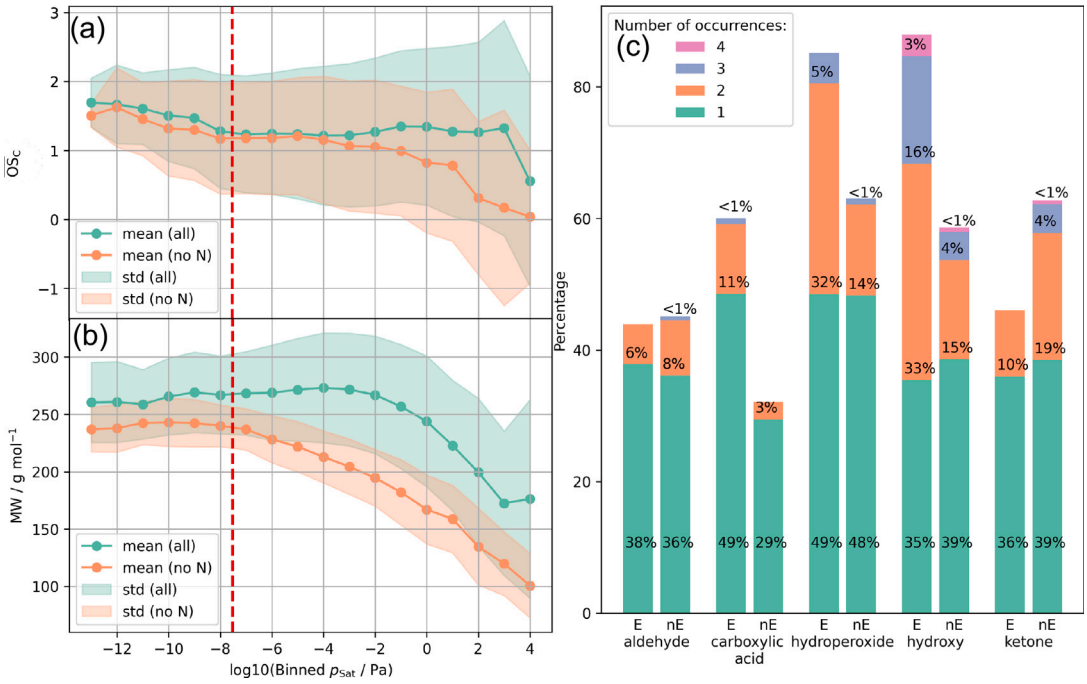

Fig. 3. Molecular properties of 42k molecules. (a) The average carbon oxidation state $\overline{OS}_C$ and (b) the molecular weight MW plotted against the binned log10 p_{Sat}. The orange curves only consider molecules without any nitrogen. Singular molecules had even lower p_{Sat}, but were excluded from the visualization for the sake of clarity. (c) Percentage of molecules that contain respective FG split by ELVOC ("E") and non-ELVOC ("nE"). Groups with less than 0.5% occurrences not depicted for clarity. "Number of occurrences" indicates the number of instances of the same FG if it was found multiple times in a single molecule. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- The classification accuracy provides information on the ability of our methods to classify a molecule correctly as ELVOC or non-ELVOC. For each method, we collect the predicted ELVOCs and non-ELVOCS and compare them to the available labeled molecules. The overlap for each category gives the number of true positives and negatives, respectively, whereas the cross-comparison returns the false positives and false negatives. With these numbers, we build confusion matrices for each method.
- The ML prediction accuracy quantifies how far the ML-predicted p_{Sat} values are from the quantum mechanical reference. We use the MAE as performance measure.

Our main objective is to find as many ELVOC as possible with the least amount of effort. We will therefore balance identification accuracy against cost. The classification accuracy provides further insight into the performance of the methods. Our secondary objective is to produce reasonable p_{Sat} predictor ML model. For this, we would strive for high prediction accuracy.

3. Results and discussion

In the following, we will compare our different ELVOC identification strategies for the different performance metrics to investigate their performance in terms of

3.1. ELVOC identification accuracy and cost

Table 2 illustrates the identification performance of the five strategies applied in this work. For each method, we list the number of correctly identified ELVOCs (N_E) and the size of the held-out set (N_{ho}) from which they were determined. For AL, GS and GE, the number of expected ELVOCs (N_{exp}) amounts to 5.1% of the held-out set size, following the GeckoQ estimate presented in Section 2.2. SIMPOL and RULE were only applied to labeled data, for which we know the number of ELVOCs as determined by COSMO-RS exactly.

3.1.1. Global models

The global strategies GS and GE find a similar number of ELVOCs (2088 and 2448, respectively). For statistical reasons, the GE ensemble method identifies 360 more ELVOCs. In GE, five models of GS size make predictions, which increases the likelihood of identifying ELVOCs. In principle, GE could identify even more ELVOCs with more inbuilt diversity (*i.e.* more training molecules in

Table 2
Summary of the identification performance and prediction accuracy of all investigated methods including the number of actual ELVOC found (N_E), held-out set size (N_{ho}), and the target number of expected ELVOCs (N_{exp}). The prediction accuracy, noted as MAE, is in log units of vapor pressure. *This cost cannot be quantified unambiguously and will be discussed in the text. Thus, there is no cost ratio.

	GS	GE	AL	SIMPOL	RULE
N_E	2088	2448	1606	1951	2136
N_{ho}	126k	126k	157k	42k	42k
N_{exp}	6414	6414	8027	5067	5067
Classification accuracy	66%	62%	54%	44%	52%
Identification accuracy	32%	38%	20%	39%	42%
Identification cost	27k	36k	3k	272*	42k*
cost ratio (ELVOC/cost unit)	0.077	0.069	0.54	–	–
MAE [log units] on all	0.84	0.83	1.19	–	–
MAE [log units] on ELVOC	1.78	1.82	2.36	–	–

Table 3
AL performance by batch: acquisition stage, the number of molecules predicted to be potentially ELVOC (N_{Pot}) in each batch, the number of subsequently labeled molecules N_{lab} and the number of actual ELVOC found (N_E) also expressed as percentage. AL model 5 was still trained, but we terminated the AL cycle after this stage, *i.e.* did not compute potential ELVOCs for a batch 6.

Acquisition stage	Batch no.	N_{Pot}	N_{lab}	N_E	Percentage per batch
AL random init.	0	–	500	27	5.4%
AL model 0	1	1520	500	292	58%
AL model 1	2	1467	500	331	66%
AL model 2	3	1196	500	331	66%
AL model 3	4	992	500	305	61%
AL model 4	5	717	500	320	64%
AL model 0–4	0–5	3439	3000	1606	54%
AL model 5	6	512	–	–	–

the ensemble). However, at a training set size of 24k out of 32k possible molecules, the statistical variation is limited and we would need larger datasets to improve the GE performance.

The identification success rates of GS and GE are 32% and 38%, respectively. While 38% is the 3rd best identification accuracy reported in Table 2, it is far from ideal. Including more molecules in the GS and GE training sets would increase the p_{Sat} prediction accuracy and therefore the identification accuracy, however, at the cost of having to label more molecules. Furthermore, the identification accuracy is limited by the fact that our decision boundary cuts through the p_{Sat} distribution (cf. Section 2.2) and molecules close to the threshold have a high likelihood of being misclassified. Secondly, ELVOCs are underrepresented in the Gecko data and it will thus be harder to predict their p_{Sat} accurately with regression models.

The cost of GS and GE is determined by the total number of molecules that need to be labeled. The base cost for both methods are the number of unique molecules in the training sets, *i.e.* 24,000 for GS and 31,637 for GE. In addition, all predicted ELVOC were checked by computing their p_{Sat}, which added further 3165 molecules for GS and 3939 for GE. Dividing the number of correctly identified ELVOCs by the cost gives us the following cost ratios of 0.077 found ELVOCs per cost unit for GS and 0.069 for GE. The higher ELVOC identification success rate of GE is therefore offset by its higher overall cost.

3.1.2. Active learning

With AL, we identified 1606 ELVOCs correctly, which is similar, although slightly lower, than with GS and GE. AL was the only method that was applied to the whole raw Gecko dataset (minus the 500 molecules of the initial training set). With an expected 8,027 ELVOCs we arrive at an identification accuracy of 20% after 5 AL iterations. Continuing the iterations did not produce appreciably more ELVOCs (see below). While the identification accuracy is lower than for the other methods, the total cost of only 3000 labels is significantly lower than for any other method in this work. The resulting cost ratio of 0.54 implies that we find one ELVOC for any two labeled molecules, which is by far the highest of any of the tested methods.

Fig. 4 and Table 3 illustrate the evolution of the AL strategy. Fig. 4 shows the p_{Sat} distributions of each AL batch, both sequentially (a) and cumulatively (b) and Table 3 lists the corresponding statistics. Batch 0 still resembles the GeckoQ distribution (with 5.1% ELVOC), because it is randomly picked. In contrast, batch 1 already homes in on the ELVOC region and contains 58% ELVOC. For batches 2 and 3 the percentage rises to 66%. For comparison, the GS method achieved the same percentage (2088 ELVOCs out 3165 potential ELVOCs) albeit with a lot more training data.

The batch distributions in Fig. 4 show that our exploitation-only acquisition strategy of adding only molecules with low p_{Sat} to the AL training set immediately finds the decision boundary. Due to the difficulty of classifying molecules close to the decision boundary accurately, each batch also includes molecules with p_{Sat} higher than the decision threshold. The peak of the distribution shifts only slightly to lower p_{Sat} with increasing iteration. The fact that the distribution does not shift more appreciably is an indication of the ELVOC scarcity in our dataset. We are searching for the low p_{Sat} tail of the raw Gecko distribution and our AL algorithm does this efficiently.

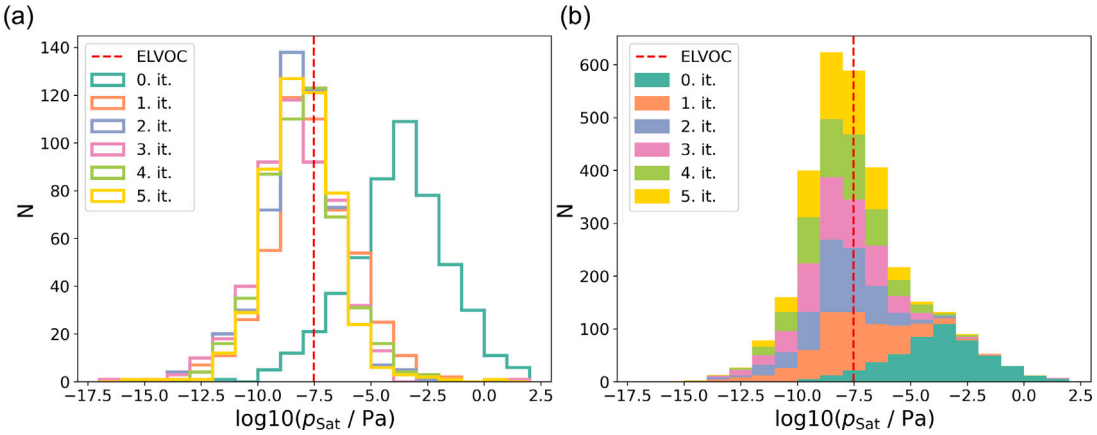

Fig. 4. Evolution of ELVOC content in chosen batch, (a) sequentially and (b) cumulatively. Histogram of the p_{Sat} distributions by batch with a bin size of one log unit for all active learning iterations, where each batch contains 500 molecules. The counts for each bin are listed in the SI, Table. S1. Four molecules with a p_{Sat} below 10^{-17} Pa are not depicted, because the corresponding histogram bar would have been too low to be visible.

Table 3 further shows that the number of potential ELVOCs predicted by AL decreases after batch 2. This suggests that AL has identified most of the ELVOC in the raw Gecko data that are similar to its training set. The characteristics of molecules found by AL and GS/GE are contrasted and analyzed further in the SI. The total number of correctly identified ELVOCs could be increased further by restarting the AL method with a new initial molecule distribution. Considering that the initial distributions are randomly drawn from the raw Gecko data, the likelihood that we end up with similar ELVOCs is high. It is therefore plausible that a certain percentage of ELVOCs are structurally and chemically so different from the rest of the Gecko data that our regression models cannot capture them accurately enough. Finding such ultra-rare molecules would require different strategies, *e.g.*, coupling AL with generative models that learn the structure and chemical identity of ELVOCs. However, such approaches are beyond the scope of this work.

3.1.3. SIMPOL and RULE

By applying SIMPOL we correctly identified 39%, *i.e.* 1951, of all ELVOC in the pool of 42k molecules. This identification accuracy is the second highest identification accuracy, narrowly higher than that of GE. Nevertheless, the SIMPOL identification cost is very difficult to quantify. While the immediate cost to the user is zero, SIMPOL was fitted to experimental measurements of 272 compounds at different temperatures. Such measurements carry a disproportionately higher cost than a computation, due to the cost of lab equipment and invested time of human labor. The ELVOC identification cost can therefore not be expressed easily in terms of computational units as we did for GS, GE and AL.

The parametrization of SIMPOL comes with some caveats that require consideration in the analysis: it is limited to molecules containing the FGs SIMPOL was parameterized with. Also, species that actually drive pure organic particle formation mostly include C_{15} – C_{20} accretion products, (Dada et al., 2023) which are chemically even further from the SIMPOL fitting data than the sequential oxidation products contained in GeckoQ. (This is also a reason why there is such a low percentage of ELVOC in the GeckoQ data.) Finally, SIMPOL serves the atmospheric science community and does not generalize to other application domains as the ML methods do (once they have been retrained on the new data).

With RULE we correctly identified 2136 out of 42k molecules giving us an identification accuracy of 42%. This comparatively high percentage illustrates that once we have sufficient data, we can use it to derive chemical and physical rules from it. The price to pay is the dataset generation, which in this case amounts to 42k labeled molecules. The resulting cost ratio of 0.05 is then comparable to that of GS and GE. We nevertheless marked the cost ratio as unspecified in Table 2, because we did not test how much or rather how little labeled data would be needed to derive the rules for RULE reliably. Furthermore, it is not clear if the rules derived for the Gecko dataset in this work will be transferable to other atmospheric science datasets.

3.2. Classification accuracy

The classification accuracy indicates the percentage of actual ELVOC in a set of molecules predicted to be ELVOC. As such it is obtained by dividing the true positives (“TP”) by the sum of all positives. Fig. 5 depicts the confusion matrices for GS, GE, SIMPOL and RULE, including the percentage of each class. The classification accuracy is the percentage indicated at “TP”. Because AL is constructed iteratively, no single confusion matrix can be created. Nevertheless, we can determine the overall AL classification accuracy, which is the number of correctly identified ELVOC (1606) divided by the labeled molecules (3000). Then the AL classification accuracy is 54%.

The ML-based methods have the highest classification accuracy: 66% (GS), 62% (GE) and 54% (AL), in contrast to SIMPOL and RULE with 44% and 52%, respectively. The reason for this is that the classification accuracy reflects how well a method understands

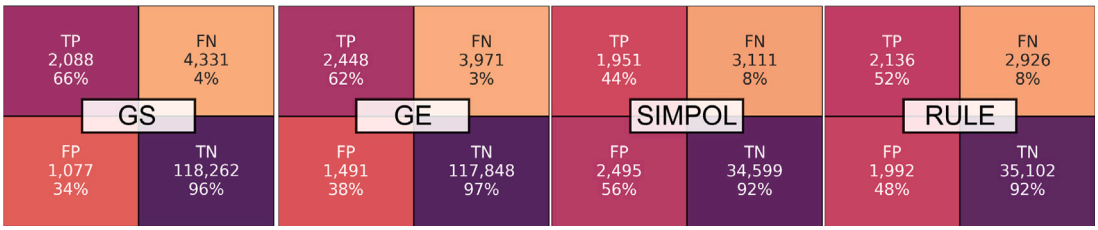

Fig. 5. The classification matrices for GS, GE, SIMPOL, and RULE. In each square the first line is the classification outcome (**T**rue **P**ositive, **F**alse **N**egative, **F**alse **P**ositive and **T**rue **N**egative), the second line is the absolute number of instances for the outcome and the third line indicates the percentage of instances that where classified correctly and incorrectly (= TP/(TP+FP)). The color corresponds to the third line. We refer with the *classification accuracy*, to the percentage of actual ELVOC within all predicted ELVOC.

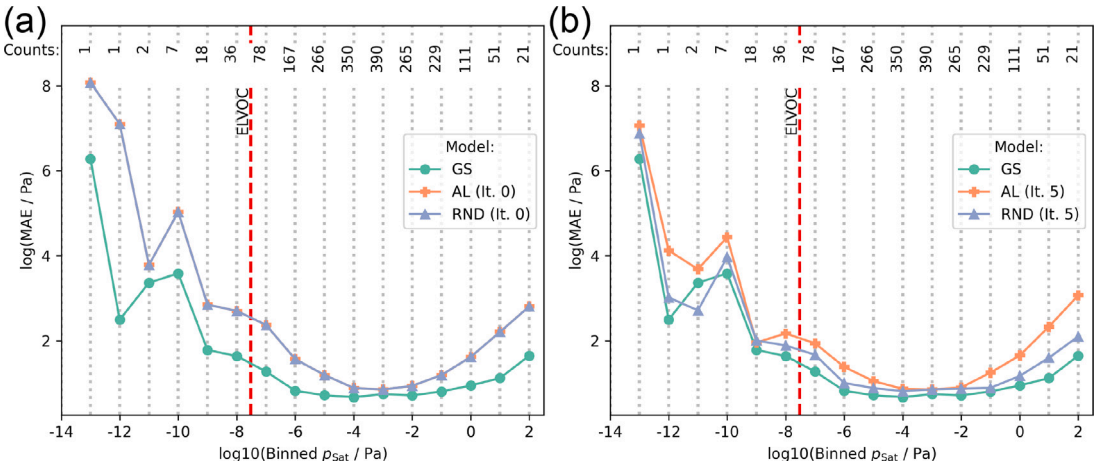

Fig. 6. Binned MAE on the general test set (test set size 2000) of AL, RND, GS and SIMPOL. The iterative models AL, and RND evolve with iterations. Iteration zero (a) and five (b) are depicted. "Counts" indicates the number of molecules in each bin. The corresponding MAE values are listed in the SI, Table S2.

the data overall, including if it knows what does **not** constitute an ELVOC. ML methods have a fairly good understanding of the full p_{Sat} range (cf. Section 3.3), and thus, yield much fewer false positives than the less selective SIMPOL and RULE methods. SIMPOLs accuracy for higher p_{Sat} molecules is so poor that it even identifies more false positives than true positives. The AL classification is comparably low because of the random initial batch, and if this batch was omitted, the AL classification would be $1579/2500 = 63\%$

3.3. ML prediction accuracy

Finally, we want to evaluate if the applied ML methods are capable of producing a reasonably accurate p_{Sat} predictor, and to this end we examine the prediction accuracy. Because GS and GE MAEs are practically identical, only GS, together with AL, and the RND reference are considered. Fig. 6(a) and (b) display the test set MAE for AL training set sizes 500 (iteration 0) and 3000 (iteration 5), respectively, as a function of p_{Sat}. For p_{Sat} lower than 10^{-10} Pa, we have only 11 molecules in the test set and therefore we excluded this region from our analysis.

At zeroth iteration, AL and RND are identical. Their p_{Sat} predictions are most accurate for molecules around the peak of the p_{Sat} distribution of the GeckoQ dataset (10^{-4}–10^{-3} Pa (MAE = 0.85 log units)). Predictions degrade for lower and higher p_{Sat}. Both, the AL and RND, models are worse than GS. At iteration 5, the performance of AL and RND has improved. Their predictions are now level with those of GS, although at much lower training set sizes. AL performs poorer for high p_{Sat}, which is expected, since its training data predominately comprises molecules with low p_{Sat}. For p_{Sat} below 10^{-6} Pa, the error of all methods rises and the error begins to vary more. This indicates that we are approaching the low p_{Sat} tail of the distribution. Notably, AL predictions are also less accurate than RND predictions for low p_{Sat}. The SI contains a detailed analysis of the molecules acquired by AL. We observed that the ELVOC can be grouped into three structurally distinct clusters. AL is only able to identify ELVOCs reliably in two of these three clusters.

Overall, GS (and GE) provide the highest predictive accuracy for the whole p_{Sat} range. Considering the much higher cost of the global models, the performance of the much cheaper AL method is satisfying.

3.4. Final assessment

Having analyzed our performance measures, we now reconnect with our research objectives. Our main objective was to identify new ELVOC cost-efficiently. Our secondary objective was to compare different ELVOC identification methods and assess their pros and cons. We discuss the performance of the different methods for two different scenarios:

1. **Labeled data is available**:
   If a sufficient (uniform) subsample of the molecular data is labeled or can be computed, we recommend training ML models on as much data as possible. The larger GS and GE models in our work had a better identification and classification accuracy and identified more ELVOCs than AL. These larger models could then be used as starting point for further AL searches to identify more ELVOC.
2. **No labeled data is available**:
   With a cost ratio of 0.5, AL is by far the most cost efficient method to identify new ELVOCs. Its identification accuracy, however, is not the highest and it depends on the application, if it is good enough. If the ELVOC search is performed in a large or potentially infinite space, the demonstrated identification accuracy will be sufficient. For smaller datasets, the accuracy should be improved, which could be done by running AL ensembles or adding exploration criteria to the acquisition function. Both improvement strategies, however, come at the expense of increased labeling cost.
   Pre-parameterized methods such as SIMPOL or rule-based methods such as RULE derived in this work, can also be applied, if no labeled data is available. SIMPOL and RULE identified slightly more ELVOCs than AL, but at the price of many labeled false positives. However, both methods have been derived specifically for p_{Sat} predictions and will not generalize to identification tasks for other properties, unlike AL. Shortcomings of SIMPOL or RULE will become even more apparent for new molecular datasets that emerge from auto-oxidation and accretion reactions that are currently being added to the GECKO-A reaction mechanism generator (RMG) (Franzon, Camredon, Valorso, Aumont, & Kurtén, 2024). The resulting molecules are more complex than the data for which SIMPOL and RULE were developed for and their accuracy is not hitherto known.

The more atmospheric chemistry GECKO-A or other RMGs include, the more molecular data they will produce, which makes it essential to have reliable tools to identify molecules of interest for aerosol formation, such as ELVOCs. As a specific example of a use case, RMGs combined with mass spectrometric measurements can be used to connect measured mass peaks of complex oxidation products with likely molecular structures (Sandström, Rissanen, Rousu, & Rinke, 2024). The machine learning methods developed in this work can then be used to determine which of these structures are the most likely to participate for example in new particle formation.

4. Conclusions

We have compared active learning, traditional machine learning, the group contribution method SIMPOL, and a rule-based approach (RULE) for their ability to identify sparsely represented EVLOC in a pool of OOM data. We found that active learning is particularly data efficient, while larger, traditional machine learning models, SIMPOL and RULE exhibit better identification accuracy. Altogether, we identified 3459 new ELVOC in this work with our different search strategies, which can now be further investigated. The machine learning methods investigated in this work are more generally applicable than SIMPOL or RULE and will be beneficial in atmospheric science as more data becomes available.

CRediT authorship contribution statement

Vitus Besel: Writing – original draft, Software, Data curation, Conceptualization. **Milica Todorović:** Writing – review & editing, Conceptualization. **Theo Kurtén:** Writing – review & editing, Methodology. **Hanna Vehkamäki:** Writing – review & editing, Supervision. **Patrick Rinke:** Writing – original draft, Supervision.

Declaration of competing interest

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

Data availability

All data utilized in and generated through this work is publicly available through the referenced f.a.i.r data repository: https://doi.org/10.23729/dd0396b3-9017-40f2-ae4b-6876bf33dd08.

Acknowledgments

We thank Bernard Aumont, who shared the original raw GECKO-A data with us. This work was supported by the CSC - IT Center for Science who provided access to the Mahti computer cluster, as well as EuroHPC for facilitating our work on the LUMI platform. This study received financial support from the Academy of Finland through its flagship program, the Atmosphere and Climate Competence Center (Grant No. 337549), and the Centers of Excellence Program (CoE VILMA, Grant Nos. 346369, 346368, and 346377).

References

Arias, P., Bellouin, N., Coppola, E., Jones, R., Krinner, G., Marotzke, J., et al. (2021). Technical summary. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, B. Zhou (Eds.), *Climate change 2021: the physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change* (pp. 33–144). Cambridge University Press, http://dx.doi.org/10.1017/9781009157896.002.

Aumont, B., Szopa, S., & Madronich, S. (2005-09-22). Modelling the evolution of organic carbon during its gas-phase tropospheric oxidation: development of an explicit model based on a self generating approach. *Atmospheric Chemistry and Physics*, *5*(9), 2497–2517. http://dx.doi.org/10.5194/acp-5-2497-2005.

Balasubramani, S. G., Chen, G. P., Coriani, S., Diedenhofen, M., Frank, M. S., Franzke, Y. J., et al. (2020). TURBOMOLE: Modular program suite for ab initio quantum-chemical and condensed-matter simulations. *Journal of Chemical Physics*, *152*(18).

Becke, A. D. (1988). Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, *38*, 3098–3100.

Besel, V., Todorović, M., Kurtén, T., Rinke, P., & Vehkamäki, H. (2023). Atomic structures, conformers and thermodynamic properties of 32k atmospheric molecules. *Scientific Data*, *10*(1), 450. http://dx.doi.org/10.1038/s41597-023-02366-x.

Bianchi, F., Kurtén, T., Riva, M., Mohr, C., Rissanen, M. P., Roldin, P., 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. http://dx.doi.org/10.1021/acs.chemrev.8b00395.

Dada, L., Stolzenburg, D., Simon, M., Fischer, L., Heinritzi, M., Wang, M., et al. (2023). Role of sesquiterpenes in biogenic new particle formation. *Science Advances*, *9*(36), eadi5297. http://dx.doi.org/10.1126/sciadv.adi5297.

Dassault Systèmes (2022). BIOVIA COSMOconf. URL http://www.3ds.com.

Donahue, N. M., Kroll, J. H., Pandis, S. N., & Robinson, A. L. (2012). A two-dimensional volatility basis set – Part 2: Diagnostics of organic-aerosol evolution. *Atmospheric Chemistry and Physics*, *12*(2), 615–634. http://dx.doi.org/10.5194/acp-12-615-2012.

Eckert, F., & Klamt, A. (2002). Fast solvent screening via quantum chemistry: COSMO-RS approach. *AIChE Journal*, *48*(2), 369–385.

Franzon, L., Camredon, M., Valorso, R., Aumont, B., & Kurtén, T. (2024). Ether and ester formation from peroxy radical recombination: A qualitative reaction channel analysis. personal communication/manuscript in preparation.

Himanen, L., Jäger, M. O. J., Morooka, E. V., Canova, F. F., Ranawat, Y. S., Gao, D. Z., et al. (2020). DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, *247*, Article 106949.

Hyttinen, N., Pullinen, I., Nissinen, A., Schobesberger, S., Virtanen, A., & Yli-Juuti, T. (2022). Comparison of saturation vapor pressures of α-pinene + o_3 oxidation products derived from COSMO-RS computations and thermal desorption experiments. *Atmospheric Chemistry and Physics*, *22*(2), 1195–1208. http://dx.doi.org/10.5194/acp-22-1195-2022.

Hyttinen, N., Wolf, M., Rissanen, M. P., Ehn, M., Perakyla, O., Kurtén, T., et al. (2021). Gas-to-particle partitioning of cyclohexene-and α-pinene-derived highly oxygenated dimers evaluated using COSMO therm. *The Journal of Physical Chemistry A*, *125*(17), 3726–3738.

Isaacman-VanWertz, G., & Aumont, B. (2021). Impact of organic molecular structure on the estimation of atmospherically relevant physicochemical parameters. *Atmospheric Chemistry and Physics*, *21*(8), 6541–6563. http://dx.doi.org/10.5194/acp-21-6541-2021.

James, C., & Weininger, D. (1995). *Daylight theory manual*. Aliso Viejo, CA, USA: Daylight Chemical Information Systems Inc..

Kerminen, V.-M., Chen, X., Vakkari, V., Petäjä, T., Kulmala, M., & Bianchi, F. (2018). Atmospheric new particle formation and growth: review of field observations. *Environmental Research Letters*, *13*(10), Article 103003. http://dx.doi.org/10.1088/1748-9326/aadf3c.

Klamt, A., Jonas, V., Bürger, T., & Lohrenz, J. C. W. (1998). Refinement and parametrization of COSMO-RS. *The Journal of Physical Chemistry A*, *102*(26), 5074–5085. http://dx.doi.org/10.1021/jp980017s.

Klamt, A., & Schüürmann, G. (1993). COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *Journal of the Chemical Society, Perkin Transactions 2*, 799–805. http://dx.doi.org/10.1039/P29930000799.

Kupc, A., Williamson, C. J., Hodshire, A. L., Kazil, J., Ray, E., Bui, T. P., et al. (2020). The potential role of organics in new particle formation and initial growth in the remote tropical upper troposphere. *Atmospheric Chemistry and Physics*, *20*(23), 15037–15060.

Kurtén, T., Hyttinen, N., D'Ambro, E. L., Thornton, J., & Prisle, N. L. (2018). Estimating the saturation vapor pressures of isoprene oxidation products $C_5H_{12}O_6$ and $C_5H_{10}O_6$ using COSMO-RS. *Atmospheric Chemistry and Physics*, *18*(23), 17589–17600. http://dx.doi.org/10.5194/acp-18-17589-2018.

Landrum, G., Tosco, P., Kelley, B., Ric, Cosgrove, D., Sriniker, et al. (2023). rdkit/rdkit: 2023_03_2 (Q1 2023) release. http://dx.doi.org/10.5281/zenodo.8053810, Zenodo.

Langer, M. F., Goeßmann, A., & Rupp, M. (2022). Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *npj Computational Materials*, *8*(1), 41. http://dx.doi.org/10.1038/s41524-022-00721-x.

Lumiaro, E., Todorović, M., Kurten, T., Vehkamäki, H., & Rinke, P. (2021-09-06). Predicting gas–particle partitioning coefficients of atmospheric molecules with machine learning. *Atmospheric Chemistry and Physics*, *21*(17), 13227–13246. http://dx.doi.org/10.5194/acp-21-13227-2021.

Metzger, A., Verheggen, B., Dommen, J., Duplissy, J., Prevot, A. S. H., Weingartner, E., et al. (2010-04-13). Evidence for the role of organics in aerosol particle formation under atmospheric conditions. *Proceedings of the National Academy of Sciences*, *107*(15), 6646–6651.

Pankow, J., & Asher, W. (2008-05-19). SIMPOL.1: A simple group contribution method for predicting vapor pressures and enthalpies of vaporization of multifunctional organic compounds. *Atmospheric Chemistry and Physics*, *8*, http://dx.doi.org/10.5194/acp-8-2773-2008.

Perdew, J. P. (1986). Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Physical Review B*, *33*, 8822–8824.

Ruggeri, G., & Takahama, S. (2016). Technical Note: Development of chemoinformatic tools to enumerate functional groups in molecules for organic aerosol characterization. *Atmospheric Chemistry and Physics*, *16*(7), 4401–4422. http://dx.doi.org/10.5194/acp-16-4401-2016.

Sandström, H., Rissanen, M., Rousu, J., & Rinke, P. (2024). Data-driven compound identification in atmospheric mass spectrometry. *Advanced Science*, *11*(8), Article 2306235. http://dx.doi.org/10.1002/advs.202306235.

Schervish, M., & Donahue, N. M. (2020). Peroxy radical chemistry and the volatility basis set. *Atmospheric Chemistry and Physics*, *20*(2), 1183–1199. http://dx.doi.org/10.5194/acp-20-1183-2020.

Surdu, M., Lamkaddam, H., Wang, D. S., Bell, D. M., Xiao, M., Lee, C. P., et al. (2023). Molecular understanding of the enhancement in organic aerosol mass at high relative humidity. *Environmental Science and Technology*, *57*(6), 2297–2309. http://dx.doi.org/10.1021/acs.est.2c04587.

Tabor, D. P., Gómez-Bombarelli, R., Tong, L., Gordon, R. G., Aziz, M. J., & Aspuru-Guzik, A. (2019). Mapping the frontiers of quinone stability in aqueous media: implications for organic aqueous redox flow batteries. *Journal of Materials Chemistry A*, *7*, 12833–12841. http://dx.doi.org/10.1039/C9TA03219C.

The MathWorks Inc. (2022). *MATLAB version: 9.13.0 (R2022b)*. Natick, Massachusetts, United States: The MathWorks Inc., URL https://www.mathworks.com.

Tröstl, J., Chuang, W. K., Gordon, H., Heinritzi, M., Yan, C., Molteni, U., et al. (2016). The role of low-volatility organic compounds in initial particle growth in the atmosphere. *Nature*, *533*(7604), 527–531. http://dx.doi.org/10.1038/nature18271.

Wang, C., Yuan, T., Wood, S. A., Goss, K.-U., Li, J., Ying, Q., et al. (2017). Uncertain Henry's law constants compromise equilibrium partitioning calculations of atmospheric oxidation products. *Atmospheric Chemistry and Physics*, *17*(12), 7529–7540. http://dx.doi.org/10.5194/acp-17-7529-2017.

Westermayr, J., Gilkes, J., Barrett, R., & Maurer, R. J. (2023). High-throughput property-driven generative design of functional organic molecules. *Nature Computational Science*, *3*(2), 139–148. http://dx.doi.org/10.1038/s43588-022-00391-1.

Yan, C., Nie, W., Äijälä, M., Rissanen, M. P., Canagaratna, M. R., Massoli, P., et al. (2016). Source characterization of highly oxidized multifunctional compounds in a boreal forest environment using positive matrix factorization. *Atmospheric Chemistry and Physics*, *16*(19), 12715–12731.

Zhang, R., Suh, I., Zhao, J., Zhang, D., Fortner, E. C., Tie, X., et al. (2004). Atmospheric new particle formation enhanced by organic acids. *Science*, *304*(5676), 1487–1490.

Zheng, P., Chen, Y., Wang, Z., Liu, Y., Pu, W., Yu, C., et al. (2023). Molecular characterization of oxygenated organic molecules and their dominating roles in particle growth in Hong Kong. *Environmental Science and Technology*, *57*(20), 7764–7776. http://dx.doi.org/10.1021/acs.est.2c09252, PMID: 37155674.