https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

S2AS v1.0 and 2D polarity–volatility lumping framework v1.0: automated compound classification and scalable lumping for organic aerosol modelling

Dalrin Ampritta Amaladhasan[1], Dan Hassan-Barthaux[1], and Andreas Zuend[1]
[1]Department of Atmospheric Oceanic Sciences, McGill University, Montreal, Quebec, H3A 0B9, Canada
Correspondence: Andreas Zuend (andreas.zuend@mcgill.ca)

Abstract. Advancements in near-explicit chemical reaction mechanisms, such as the Master Chemical Mechanism (MCM) or the Generator of Explicit Chemistry and Kinetics of Organics in the Atmosphere (GECKO-A), have enabled highly detailed simulations of atmospheric chemistry. Such simulations offer a bottom-up approach to accompany and inform laboratory chamber experiments of organic aerosol formation or to model the complex chemistry of mixtures of volatile aerosol precursors for specific tropospheric conditions. These chemical reaction mechanisms, while comprehensive, generate hundreds to millions of organic components, creating computational challenges for subsequent applications in multiphase equilibrium gas–particle partitioning models to predict secondary organic aerosol (SOA) mass concentrations, phase compositions, and hygroscopicity. The wealth of simulated reactions and components also requires substantial simplifications for reduced-complexity representations in large-scale atmospheric models. This study introduces a suite of software tools to automate relevant pure-component property predictions as well as a 2-dimensional (2D) polarity–volatility lumping framework to systematically reduce the complexity of chemical mechanism outputs. We introduce a new polarity metric for use in the 2D framework, a ratio of a component's activity coefficients in water and an organic solvent (hexanediol). This ratio is computed using the Aerosol Inorganic–Organic Mixtures Functional groups Activity Coefficients (AIOMFAC) model. The 2D framework offers grid-based and cluster-based methods to select an adjustable number of surrogate species and offers flexibility in the choice of polarity axis. Our methods utilize the Simplified Molecular Input Line Entry System (SMILES) description of molecular structures. A new tool, SMILES to AIOMFAC subgroups (S2AS), is introduced to automatically generate AIOMFAC-model input files and to handle exception cases consistently. We demonstrate the application of our framework using systems of hundreds to thousands of components generated by near-explicit chemical mechanisms. The new framework enables tailored reduced-complexity representations of gas–particle systems.

1 Introduction

Secondary organic aerosol material (SOA) is formed through chemical processing and gas–particle partitioning of volatile organic precursors. SOA can consist of hundreds to millions of distinct kinds of molecules, stemming from biogenic and anthropogenic emission sources (Hallquist et al., 2009; Lannuque et al., 2021). In addition, atmospheric aerosols frequently contain primary organic aerosol material (POA), water and dissolved electrolytes, as well as insoluble species. The complexities

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

in sources, chemical and physical transformations and resulting gas- and particle-phase mixtures, introduces computational challenges when attempting to predict component properties and the partitioning behaviour of such organic–inorganic aerosol systems. Atmospheric chemical transport models often resort to the use of highly simplified volatility binning or approaches relying on surrogate components for gas–particle partitioning predictions of organics. Those models often also assume ideal condensed phase mixing behaviour, in part due to computational time considerations and in part due to a lack of efficient thermodynamic mixing models (Byun et al., 1999; Semeniuk and Dastoor, 2020).

Laboratory experiments, field studies and theory suggest that nonideal mixing in condensed particulate matter (PM) phases impacts the gas–particle partitioning process and influences the physicochemical properties of the condensed phase, often leading to liquid–liquid phase separation over a wide range of environmental conditions (Pankow, 2003; Erdakos and Pankow, 2004; Smith et al., 2011; Bertram et al., 2011; Zuend et al., 2010; Huang et al., 2021; Schervish and Shiraiwa, 2023). Models have been developed to predict SOA formation based on the thermodynamic equilibrium partitioning of semivolatile organic oxidation products, including versions for application in atmospheric large-scale models (Pun et al., 2002; Griffin et al., 2003; Chang and Pankow, 2006; Tulet et al., 2006; Pankow and Chang, 2008; Wang et al., 2022). The model by Griffin et al. (2003) allows for gas–particle equilibrium of organic and inorganic compounds considering an organic and an inorganic phase, but it does not allow organics and salts to partition between the two PM phases. A gas–particle partitioning approach using the activity coefficient model X-UNIFAC proposed by (Chang and Pankow, 2006), an extension of the UNIquac Functional group Activity Coefficients (UNIFAC) model (Fredenslund et al., 1975), enabled equilibrium of all species between all phases present yet restricted to single electrolyte components. Since then, improved thermodynamic multiphase modelling frameworks have been introduced (e.g. Zuend et al., 2010) for applications in box models, yet all such models reach computational limitations, such as excessive memory requirements, when applied to highly complex aerosol systems containing many hundreds to thousands of interacting components.

Based on theory, the equilibrium gas–particle partitioning of a certain semi-volatile (organic) compound is mainly governed by three key properties: (1) the pure-component saturation vapour pressure, (2) the effective activity coefficient in the absorbing aerosol phase and (3) the total mass concentration of all the material in the absorbing condensed phase (e.g., Pankow, 2003; Donahue et al., 2006; Zuend et al., 2010). Thus, the pure-component saturation vapour pressure is a critical input for equilibrium partitioning models, including for box models based on the Aerosol Inorganic–Organic Mixtures Functional group Activity Coefficients (AIOMFAC) model (Zuend et al., 2008, 2011), X-UNIFAC and other UNIFAC variants (Hansen et al., 1991; Pankow and Asher, 2008b; Chang and Pankow, 2010; Compernolle et al., 2011).

One way to generate the chemical composition of an air parcel is by predicting the gas- and/or particle-phase composition by means of integrating a chemical reaction scheme over time. In this study, we focus on the development and discussion of necessary tools for handling the output of detailed reaction mechanisms. The aim is to process such output for subsequent equilibrium gas–particle partitioning computations, which in turn predict the composition, PM mass concentration and other SOA properties.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

1.1 Near-explicit chemical mechanisms

The Master Chemical Mechanism (MCM, v3.3.1) is a near-explicit reaction scheme of the gas-phase chemistry, covering a substantial set of (volatile) aliphatic and aromatic hydrocarbon compounds in atmospheric chemistry models (Jenkin et al., 1997; Saunders et al., 2003). Oxidation products of volatile or intermediate-volatility compounds may be of sufficiently low volatility to contribute to condensed aerosol mass. In past work, molecular concentrations of a subset of oxidized compounds simulated by MCM have been used as input composition information in the gas–particle partitioning model by (Zuend and Seinfeld, 2012) to predict SOA mass concentrations at varying levels of relative humidity (RH). Another state-of-the-art, near-explicit model is the Generator for Explicit Chemistry and Kinetics of Organics in the Atmosphere (GECKO-A) by Aumont et al. (2005); Mouchel-Vallon et al. (2020). GECKO-A is a chemical mechanism generator, which automates the creation of thousands of reactions and thousands to millions of oxidation and fragmentation products from a single precursor or a mixture of precursors (depending on the structural complexity of the precursors). GECKO-A achieves this by algorithmically generating the likely chemical products and related kinetic rate constants for multiple generations of reactions of a precursor and its derivatives (Aumont et al., 2005; Mouchel-Vallon et al., 2020). A box model (as part of GECKO-A) can then be run under given conditions of temperature, RH, reaction time and oxidant concentrations to generate the molecular output concentrations at specified times of interest (Aumont et al., 2005; Mouchel-Vallon et al., 2020). The processing of the wealth of component information from such near-explicit methods and related box model simulations requires the use of automated compound classification tools – the motivation for this study.

1.2 Cheminformatics tools

Mapping molecular information from near-explicit chemical mechanisms onto a lower-dimensional parameter space enables representations of large data sets at customized resolution and allows for running equilibrium thermodynamic models within their computational feasibility range. Therefore, such mappings aid in achieving adjustable-resolution model–measurement comparisons of aerosol properties. One approach for achieving this dimensionality reduction involves representing molecular structures using methods that capture essential features in a compact format. Molecular structures can be represented using the Simplified Molecular Input Line Entry System (SMILES) (Weininger, 1988; Toropov et al., 2008), a linear ASCII text string convertible to 2D (or 3D) molecular drawings and related internal representations by cheminformatics packages such as OpenBabel (O'Boyle Jr et al., 2011), and RDKit (Landrum, 2013). Additionally, the SMiles ARbitrary Target Specification (SMARTS) language allows specifying substructure patterns in molecules using pattern matching relations. The SMARTS notation enables the development of customized algorithms to extract targeted molecular structure information of interest for a variety of applications, including to identify functional (sub)groups used to describe molecular structures within the AIOMFAC model.

Cheminformatics toolboxes such as the free Open Babel chemistry toolbox (O'Boyle Jr et al., 2011), the Chemistry Development Kit (Steinbeck et al., 2003), the OEChem (OpenEye Scientific) software (OEChem, 2012), and the open-source RDKit software (Landrum, 2013), are capable of converting and managing chemical molecular data and can be utilized to

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

apply existing or newly developed tools for substructure pattern matching (Allen et al., 2016; Ehrlich and Rarey, 2012). These tools can match the substructures of given functional groups by parsing molecular structures that are internally stored as assemblies of atoms and associated bonds using SMARTS strings for queries (Ruggeri et al., 2016). Since several AIOMFAC-based functional groups differ from the UNIFAC-based functional groups, an AIOMFAC-specific SMARTS-based pattern matching algorithm was developed (see Sect. 2.1.1) based on the open-source cheminformatics API from the epam Indigo toolkit (Pavlov et al., 2011), which builds on the Open Babel toolbox and offers an efficient and user-friendly option for customizing SMILES–SMARTS applications.

1.3 Need for reduced-complexity frameworks

Coupled liquid–liquid phase separation and gas–particle partitioning calculations, such as with the AIOMFAC-based model, are limited to systems containing less than $\sim$ 1,000 components for reasons of computational speed and limited random access memory – and in many practical applications to systems of less than $\sim$ 50 components. Therefore, output from near-explicit chemical mechanism simulations need to be drastically reduced in complexity. To address this at the system level, a two-dimensional (2D) structure–property space and related component lumping framework is introduced in this study. The main purpose of our framework is to effectively lump the hundreds to millions of system components into a manageable set of representative surrogate components while retaining an overall similar gas–particle partitioning behaviour. Furthermore, the method is designed to select surrogate components in an objective, automated manner, offering applications beyond the main use case discussed in this study.

Our scheme builds on related prior work by Pankow and Barsanti (2009), who introduced similar 2D carbon-number–polarity grid and the work by Jimenez et al. (2009); Donahue et al. (2011); Kroll et al. (2011) and Donahue et al. (2012), who introduced so-called volatility basis set spaces to characterize chemical compound evolution and/or thermodynamic mixing behaviour of organic aerosol systems. In contrast to a 1-dimensional (1D) volatility basis set (VBS) (e.g., Donahue et al., 2006; Sommers et al., 2022), a 2D scheme allows for a more nuanced representation of complex organic aerosol systems by considering both volatility and polarity (hygroscopicity) characteristics of individual components. The 2D space also offers a visual representation of the chemical diversity within organic aerosol systems, enabling researchers to identify patterns and time evolution trends in component behaviour. This approach facilitates a more intuitive understanding of complex aerosol systems and aids in the development of simplified models that retain essential physicochemical characteristics. Our approach introduces a new polarity metric and offers a flexible framework that can be adapted to various levels of detail required for different modelling scenarios. The restriction to two dimensions is both related to the theoretical basis of the dominant factors of volatility and polarity (and related nonideal mixing) on SOA gas–particle partitioning, as well as to account for the trade-off between computational cost and resolved details.

Section 2 describes our chain of tools developed for automatic characterization of the relevant pure-component properties as well as the use of different polarity axis choices and surrogate selection methods in our new 2D lumping framework. Section 3 shows applications to example systems generated by MCM or GECKO-A and discusses the performance of the new tools.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

2 Methods and Data

The equilibrium gas–particle partitioning model applied in this study has been introduced in previous work (Zuend et al., 2010; Zuend and Seinfeld, 2012). Briefly, this thermodynamic equilibrium model is built around the AIOMFAC thermodynamic model of nonideal mixing (Zuend et al., 2008, 2011). AIOMFAC predicts the mixing behaviour of organic–inorganic solutions by calculating the activity coefficients of electrolytes, water, and organics for any given (liquid or amorphous) mixture composition. The gas–particle partitioning calculations include the simultaneous consideration of liquid–liquid phase separation while treating the gas phase as an ideal gas mixture.

Figure 1 provides a schematic overview of the set of methods used to automate the processing of molecular-level data for gas–particle partitioning calculations and component property characterization. The different pure-component or system-level tools (blue and yellow boxes) will be discussed in separate subsections in the following. The gas-phase chemical mechanisms targeted in our work provide outputs at a selected time in the form of lists of components whose structures are expressed in (or converted to) SMILES format, alongside with the corresponding molecular amounts per unit volume of air, usually provided in units of $\mathrm{molec\,cm^{-3}}$ or $\mathrm{mol\,m^{-3}}$. These lists form the inputs to our multi-model toolchain (Fig. 1). In terms of aerosol applications, one key question concerns how much of the organic mass remains in the gas phase and how much of it contributes to the condensed PM mass concentration under equilibrium conditions. Additional questions concern the hygroscopicity, the potential multi-phase structure of the aerosol material, and related morphology and surface properties. The methods introduced here support answering such questions quantitatively and systematically, even for cases of highly complex chemical mechanism outputs. As indicated in Fig. 1, other temperature-dependent pure-component properties could be determined via existing or new SMILES- and SMARTS-based methods. Such examples include predictions of pure-component surface tension of interest for cloud droplet activation and liquid–liquid interfacial tension (Schmedding et al., 2025; Topping et al., 2007; Schmedding and Zuend, 2025), the solid- and liquid-state densities of organic compounds (Topping et al., 2016b; Girolami, 1994), and the glass transition temperature and related pure-component viscosity of interest for molecular diffusion and aerosol mixing timescale modelling (Galeazzo and Shiraiwa, 2022; Armeli et al., 2023; DeRieux et al., 2018).

As mentioned in Sect. 1.2, the SMILES format is a plain-text notation for describing a component's chemical structure in great detail. SMILES data can be processed by many existing chemical informatics tools. In this work, we make use of the tools based on the Open Babel project, the related Python bindings (via pybel) and/or the application processing interface (API) from the Indigo cheminformatics library (toolkit version 1.7.0) (Pavlov et al., 2011). These third-party libraries can be imported into Python programs. They offer a straightforward means to processing SMILES input, including molecule completeness verification, conversion of generic SMILES into unique SMILES, and the application of customized SMARTS pattern matching. Of note, the Indigo cheminformatics library, even when accessed via its Python library, is running performance-critical computations using an efficient, compiled version of the Open Babel code and Indigo features written in the C++ language.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

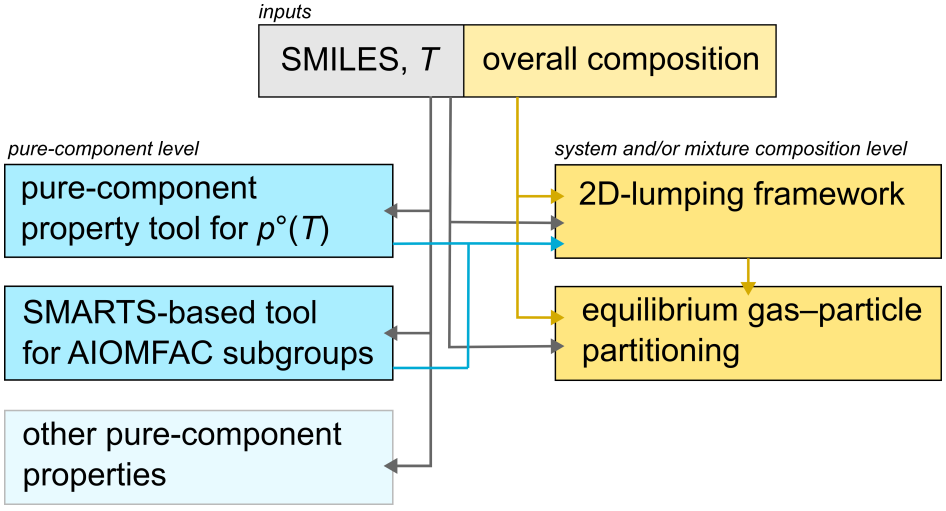


Figure 1. Schematic overview of the procedures for automatically handling detailed molecular-level data of the numerous components of an organic gas–aerosol system. Component inputs include SMILES for molecular structure, temperature T, and the system's overall (gas plus particle) composition in terms of mass or molar species concentrations. At the pure-component level (blue boxes), SMARTS libraries are employed for estimating pure-component vapour pressure at T and AIOMFAC (functional) subgroup characteristics of each molecule. At the system and/or mixture composition level (yellow boxes), 2D product lumping and surrogate selection is performed based on the volatility, polarity and mass concentrations of the system's organic compounds. Thermodynamic equilibrium partitioning computations are then performed using the selected surrogate system.

2.1 Tools for pure-component property predictions

2.1.1 SMILES to AIOMFAC subgroups (S2AS) tool

The SMILES to AIOMFAC subgroups (S2AS) tool is a new, automated algorithm written in Python. It is designed to identify and classify functional groups comprising organic aerosol components in the input format required by the AIOMFAC model. Details of AIOMFAC's representation of molecular structures by a so-called subgroup notation are provided elsewhere (Zuend et al., 2008, 2011); see also examples at https://aiomfac.lab.mcgill.ca/about.html. Briefly, the notation of aromatic and aliphatic organic compounds in AIOMFAC is based on that of UNIFAC; e.g. a molecule like ferulic acid (SMILES code: COc1cc(ccc1O)/C=C/C(=O)O), is comprised of the following AIOMFAC subgroups: 1 × (CH=CH), 3 × (ACH), 2 × (AC), 1 × (ACOH), 1 × (CH3O), 1 × (COOH), where AC denotes aromatic carbon (lower-case c in SMILES). Unlike in SMILES notation, the subgroup input format for AIOMFAC explicitly states the hydrogen atoms, yet the subgroup notation does not contain information about how the subgroups are connected to each other (because AIOMFAC does not require that information).

We note that the existing list of subgroups in AIOMFAC (about 60 subgroups supported for organic compounds plus special subgroups for inorganics) has limitations when it comes to representing rather exotic, highly functionalized compounds, which

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

may not allow for a perfect mapping by the S2AS tool. Consequently, we implemented a mechanism for detecting and handling exceptions, in most cases by introduction of additional SMARTS patterns to cover these cases. Encountering an exception typically means either that not all atoms can be uniquely associated with only one AIOMFAC subgroup, that a functionality needs to be approximated by an imperfect combination of existing subgroups, e.g. in the case of secondary ozonide functionalities, or that after parsing all existing SMARTS patterns, one or several unmatched atoms remain. Treating such exceptions by an algorithm allows for a consistent, user-independent approximation of the suboptimal mapping. Furthermore, encountered exception cases can be flagged to indicate the potential need for an additional SMARTS pattern to recognize a special case and/or to motivate future improvements by introducing new subgroups into AIOMFAC. To this end, based on our tests with tens of thousands of compounds generated by GECKO-A or MCM (see Sect. 3), unhandled exception cases are a rare occurrence.

The S2AS program carries out the following key steps: (i) parsing of each SMILES string from an input list to determine whether the SMILES input is valid and whether the component falls into the special category of being a pure alcohol or polyol according to the definition used by AIOMFAC; (ii) rendering of molecules for structure visualization as portable network graphics (.png) or scalable vector graphics (.svg) files (this step is optional); (iii) matching substructures to related AIOMFAC subgroups by iterating over a list of SMARTS as outlined in the flowchart of Fig. 2. During step (i), a character string filter is also applied to remove chirality information from input SMILES, which is unnecessary for AIOMFAC subgroups, and to replace radical atoms in a SMILES string by the corresponding non-radical atom (e.g., [O.] by O). The latter is done since AIOMFAC does not support radicals. The pure-component properties and thermodynamic mixing behaviour of such a compound is then approximated by that of a similar non-radical molecule.

Our implemented SMARTS pattern matching process follows a hierarchical, priority-based querying approach, with relatively large subgroups, such as peroxy acyl nitrate (SMARTS: [CH0;X3](=O)OO[NX3;0,+1](=O)-,=[O;0,-1]), assigned one of the highest matching priorities, while the SMARTS pattern for a single aliphatic carbon bonded to two non-hydrogen atoms and two hydrogens (SMARTS: [CH2]), is among the last patterns applied. The query list and order of SMARTS patterns is provided in Table 1. The company Daylight Chemical Information Systems, Inc., provides manuals, examples and tutorials for understanding and customizing the SMARTS and SMILES languages on their website (https://www.daylight.com, last access: 16 June 2025).

Figures 3 and 4 provide examples of the individual mappings of AIOMFAC subgroups by SMARTS. Alkyl groups, having the lowest matching priority, are matched after all other groups. Pure aliphatic alcohols and polyols are initially detected as such and treated in a separate code branch based on a distinct list of SMARTS, as demonstrated in the example of Fig. 4. Based on the polyol-specific subgroups and nomenclature introduced into a variant of UNIFAC by Marcolli and Peter (2005), which is also supported in AIOMFAC, the alcohols and polyols make use of a set of special alkyl subgroups for added specificity and better accuracy of AIOMFAC water uptake and liquid–liquid equilibrium predictions for this class of compounds. A key strength of the S2AS tool is its ability to systematically and efficiently account for the special alkyl groups in a wide variety of straight-chain, branched and cyclical aliphatic alcohols/polyols. For these compounds, the algorithm in S2AS follows a three-step procedure. In step (1) we determine and count the CH_n (n = 0, 1, 2, 3) directly bonded to OH groups as well as the associated OH groups, which are separate subgroups. In step (2), we match all alkyl groups belonging to hydrophobic tails by

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

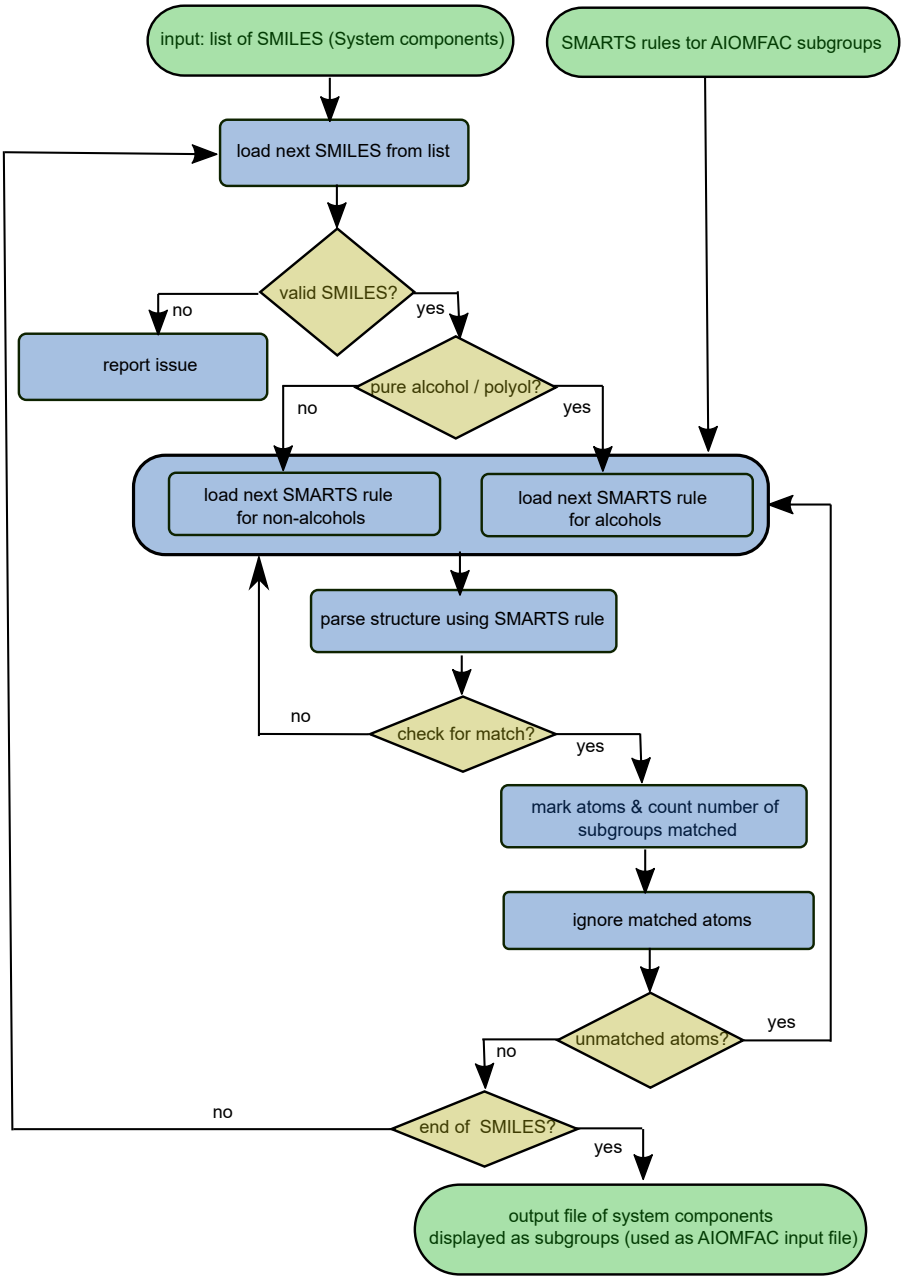


Figure 2. Flowchart illustrating the substructure pattern matching algorithm incorporated in the S2AS tool. SMARTS rules corresponding to all available AIOMFAC functional groups (subgroups) have been formulated in a priority-ordered list; see Table 1.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

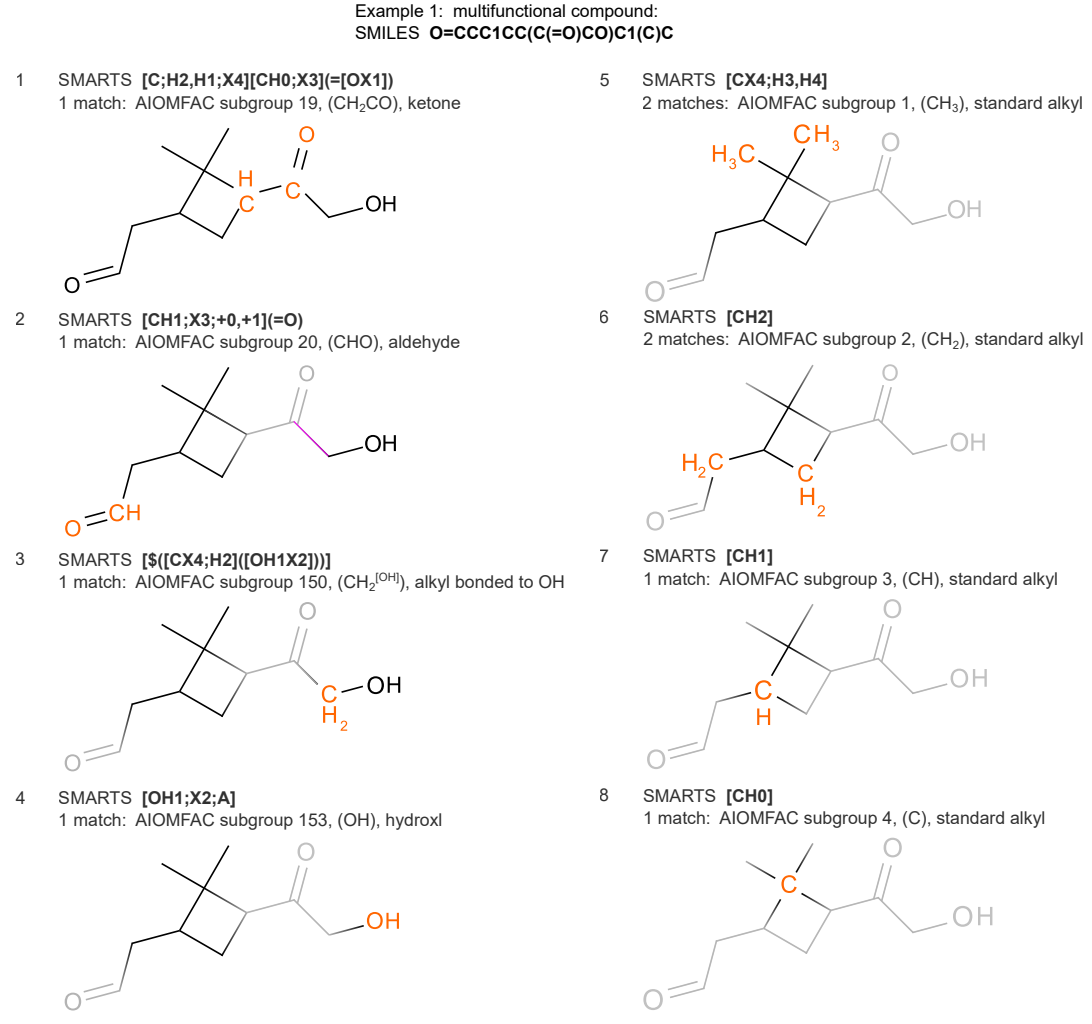


Figure 3. Example of AIOMFAC subgroup determination. Sequences 1–8 show the selection of SMARTS patterns for which at least one match was found, while the structure is probed sequentially by SMARTS from the list (ordered from high to low priority; see Table 1). Atoms matched by a stated SMARTS pattern are highlighted in orange, e.g. the ketone group in panel 1, and subsequently ignored. Light grey colouring denotes ignored atoms/bonds at a certain SMARTS parsing stage.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

first marking alkyl chains that terminate in $-CH_n-CH_3$ (where n = 0, 1, 2) as tails and then iteratively following along those alkyl chains one CH_n group at a time. An alkyl group is part of a hydrophobic tail if, and only if, it connects to at least one other alkyl group known to be part of a hydrophobic tail, while not being bonded to an OH group (those were already determined in step 1). In step (3) all thus far unassigned alkyl groups are detected and classified as being of "alkyl within alcohols" CH_n (n = 0, 1, 2, 3) subgroup type.

In general, the implementation of a hierarchical order and processing of the list of SMARTS is highly advantageous. It allows one to write the SMARTS codes for subgroups of lower priority in a far simpler notation than when each SMARTS code were required to work correctly regardless of the order of execution. For example, if [CH2] were applied as one of the first SMARTS pattern tested, it would likely result in several unwanted matches, such as matching the CH2 atoms associated with a ketone subgroup (e.g., CH2CO in AIOMFAC notation); subsequently, the $>C=O$ of the remaining atoms of the ketone group would not be detected as being part of a full ketone subgroup (clearly a mistake). To avoid this, the SMARTS pattern for only detecting intended CH2 groups would need to be much more complicated, such that it would avoid matching atoms that could be matched as part of a bigger substructure. When a SMARTS match is found for the molecule under consideration, the corresponding matched atoms are marked and excluded from further parsing if unmatched atoms remain in the component; the Indigo toolkit provides "ignore atom" and "highlight atom" functions that conveniently aid in avoiding any unwanted double-matching of atoms. A numerical counter corresponding to the key of the matched SMARTS rule and associated subgroup is incremented for each successful pattern match and added to the subgroup-array representation of the compound for later S2AS output. After all atoms have been matched or when the end of the list of SMARTS is reached, a check is performed to determine whether any unmatched atoms remain in a given molecule, potentially indicating an exception case (very rare). After, the next molecule from the SMILES list is processed. Once all SMILES have been processed, the S2AS program outputs a text file in the format of AIOMFAC-web input files (see examples at https://aiomfac.lab.mcgill.ca/about.html).

To validate the S2AS tool and associated SMARTS patterns, we used a comprehensive data set of aerosol-relevant organic compounds, including those produced by MCM v3.3.1 simulations for monoterpenes and alkanes, an excerpt is shown in Fig. 5. These tests serve as proof of concept of the tool's ability to extract and convert detailed molecular information from a diverse range of compounds commonly encountered in atmospheric chemistry simulations.

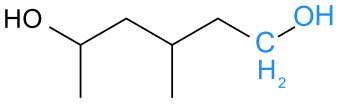
2.1.2 Pure-component vapour pressure estimation

The UManSysProp project developed by Topping et al. (2016a) is an open-source facility that employs cheminformatics tools for molecular and mixture property predictions. This facility allows users to input molecular information in the SMILES format, from which the relevant information for aerosol property calculations are extracted (Topping et al., 2016a). This tool utilizes the Open Babel and pybel cheminformatics libraries for the parsing of molecules using tool-specific SMARTS pattern matching, similar to the AIOMFAC-specific S2AS procedure described in Sect. 2.1.1. Of most relevance for this study, the pure-component, liquid-state saturation vapour pressure tool, available as part of the UManSysProp facility, provides several methods for estimating temperature-dependent vapour pressures. It uses method-specific SMARTS pattern matching to calculate the pure-component vapour pressures using predictive models, including the Estimation of VApour Pressure of ORganics,

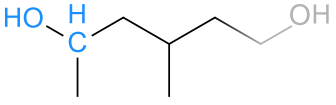
https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

Example 2: pure aliphatic alcohol/polyol compound:
SMILES **CC(O)CC(C)CCO**

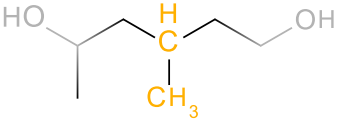
1 **Step 1:** detect hydroxyl groups and the CH_n groups directly bonded to them;
SMARTS **[CH2][OX2H1]**
1 match: AIOMFAC subgroup 150, ($CH_2^{[OH]}$), alkyl bonded to OH
and subgroup 153, (OH), hydroxyl



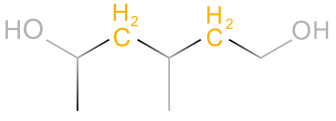
2 SMARTS **[CH1][OX2H1]**
1 match: AIOMFAC subgroup 151, ($CH^{[OH]}$), alkyl bonded to hydroxyl group
and subgroup 153, (OH), hydroxyl



3 **Step 2a:** determine end groups of hydrophobic tail chains terminating in $–CH_n–CH_3$;
SMARTS **[CH3][CH1]**
1 match: AIOMFAC subgroup 145, ($CH_3^{[alc\text{-}tail]}$), alkyl in hydrophobic tail
and subgroup 147, ($CH^{[alc\text{-}tail]}$), alkyl in hydrophobic tail



4 **Step 2b:** determine all $–CH_n–$ groups within hydrophobic tail chains;
SMARTS **[CX4]**, additional condition: must be neighbor of another hydrophobic tail group;
2 matches: AIOMFAC subgroup 146, ($CH_2^{[alc\text{-}tail]}$), alkyl in hydrophobic tail



5 **Step 3:** assign "alkyl within alcohols" type to all remaining –CHn– groups.
SMARTS **[CH3]**
1 match: AIOMFAC subgroup no. 141, ($CH_2^{[alc]}$), alkyl in alcohols

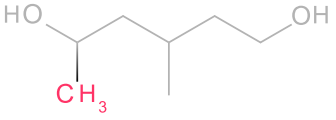


Figure 4. Example of the distinct AIOMFAC subgroup pattern matching applied to pure aliphatic alcohols or polyols, following the distinction of several alkyl group types introduced by (Marcolli and Peter, 2005). The figure illustrates our three-step algorithm to correctly identify alkyl groups in structures qualifying as pure aliphatic alcohols/polyols. Highlighted SMARTS matches follow the colour scheme of Marcolli and Peter (2005).

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

Table 1. Priority-ordered SMARTS query list for the parsing of non-polyol aliphatic and aromatic organic component SMILES and associated matching to the corresponding AIOMFAC subgroups. The key value indicates the corresponding AIOMFAC subgroup identifier a pattern will be mapped to – or, in exception cases (values > 300), an index for exception handling by the S2AS program.

Key	SMARTS	Description and remarks
172	[CH0;X3](=O)OO[NX3;0,+1](=O)-,=[O;0,-1]	peroxy acyl nitrate C(=O)OONO2 subgroup
155	[C;H2,H3][OH0X2][NX3;+0,+1](=O)-,=[O;0,-1]	organonitrate CH2ONO2 subgroup (also map CH3ONO2 as exception case)
156	[CH1][OH0X2][NX3;+0,+1](=O)-,=[O;0,-1]	organonitrate CHONO2 subgroup
157	[CH0][OH0X2][NX3;+0,+1](=O)-,=[O;0,-1]	organonitrate CONO2 subgroup
157	[OH0;X2][OH0X2][NX3;+0,+1](=O)-,=[O;0,-1]	exception: peroxide organonitrate -O-ONO2 subgroup mapped as CONO2 group due to lack of a specific subgroup;
1550	[OH0X2][NX3;+0,+1](=O)-,=[O;0,-1]	exception: special organonitrate -ONO2 subgroup without the CH2 group (remove one CHn later if possible);
54	[CH3;X4][NX3;+0,+1](=O)-,=[O;0,-1]	CH3NO2 nitro group
55	[CH2;X4][NX3;+0,+1](=O)-,=[O;0,-1]	CH2NO2 nitro group
56	[C;H1,H0][NX3;+0,+1](=O)-,=[O;0,-1]	CHNO2 nitro group (or as exception: CNO2)
57	[cH0][NX3;+0,+1](=O)-,=[O;0,-1]	aromatic nitro group ACNO2
560	[NX3;+0,+1](=O)-,=[O;0,-1]	pure nitro group; exception for one carbon having two such groups
28	[CH3;X4]-[NH2;X3]	primary amine CH3NH2 subgroup
29	[CH2;X4]-[NH2;X3]	primary amine CH2NH2 subgroup
30	[C;H1,H0;X4]-[NH2;X3]	primary amine CHNH2 subgroup; as exception also for C-NH2
31	[CH3;X4]-[NH1;X3]	secondary amine CH3NH subgroup
32	[CH2;X4]-[NH1;X3]	secondary amine CH2NH subgroup
33	[C;H1,H0;X4]-[NH1;X3]	secondary amine CHNH subgroup; as exception also for CH-NH
1027	[c]1:[c;H0;X3]:[o]:[c]:[c]1	furfural variant, mapped as 2 AC + 1 ACH + 1 ether subgroup
1026	[c;H0;X3]1:[c]:[o]:[c]:[c]1	furfural, mapped as 2 AC + 1 ACH + 1 ether subgroup
926	[c]1:[c]:[o]:[c]:[c]1	furan, mapped as 3 ACH subgroups + 1 ether subgroup
161	[CX3](=[OH0])[OH0;X2][OH1;X2]	peroxy acid C(=O)OOH subgroup
158	[C;H2,H3][OH0;X2][OH1]	hydroperoxide CH2OOH subgroup; (as exception also CH3OOH)
159	[CH1][OH0;X2][OH1]	hydroperoxide CHOOH subgroup
160	[CH0,cH0][OH0;X2][OH1]	hydroperoxide COOH subgroup (or as exception also aromatic hydroperoxide cOOH)
1580	[$([OH0;X2]-[C])][OH1]	exception: CHn-ignored hydroperoxide -O-OH subgroup, mapped as -CH2-O-OH subgroup minus 1 alkyl subgroup (CH2, if possible)
43	[CH1;X3](=O)[OH1;X2]	formic acid HC(=O)OH subgroup/molecule
137	[CH0;X3](=O)[OH1;X2]	carboxylic acid C(=O)OH subgroup
2224	[CX4][CH0;X3](=O)[OH0;X2]-[OH0;X2;A]	exception: perester group CHnC(=O)-O-O- as exception to ester group; mapped as CH3COO + CH3O subgroups minus CH3 subgroup to account for correct number of C and O atoms
2220	[CX3;H0,H1](=O)[CH0;X3](=O)[OH0;X2]-[OH0;X2;A]	exception: perester + carbonyl group CHn(=O)C(=O)-O-O- mapped in part by using aldehyde group and deducting CHn group
2022	[CX4]-[CH0;X3](=O)[CH0;X3](=O)[OH0;X2]	exception: ester + aldehyde group for CH0COO + CHn=O combination of subgroups

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

Table 1. Continued.

Key	SMARTS	Description and remarks
2022	[CH1;X3](=O)[CH0;X3](=O)[OH0;X2]-[C]	exception: ester + aldehyde group for CH1COO + CHn=O variant combination of subgroups
21	[CX4;H3][CH0;X3](=O)[OH0;X2]	ester CH3COO subgroup
22	[CX4;H2][CH0;X3](=O)[OH0;X2]	ester CH2COO subgroup
22	[C;H1,H0][CH0;X3](=O)[OH0;X2]	exception: ester CH2COO subgroup also used for CH1COO and CH0COO
18	[CH3;X4][CH0;X3](=[OX1])	ketone CH3C(=O) subgroup
19	[C;H2,H1;X4][CH0;X3](=[OX1])	ketone CH2C(=O) subgroup (also for CH1C(=O) as exception case)
20	[CH1;X3;+0,+1](=O)	aldehyde -CH(=O) subgroup
20	[CH0;X3](=O)	exception: aldehyde subgroup if ketone group cannot be used for a carbonyl in multifunctional structure
20	[CH2]=O	formaldehyde; special CH2=O subgroup mapped to subgroup 20
162	[CH3][OH0;X2][OH0;X2][CH3]	peroxide CH3OOCH3 subgroup
163	[CH3][OH0;X2][OH0;X2][CH2]	peroxide CH3OOCH2 subgroup
164	[CH3][OH0;X2][OH0;X2][CH1]	peroxide CH3OOCH subgroup
165	[CH3][OH0;X2][OH0;X2][CH0]	peroxide CH3OOC subgroup
166	[CH2][OH0;X2][OH0;X2][CH2]	peroxide CH2OOCH2 subgroup
167	[CH2][OH0;X2][OH0;X2][CH1]	peroxide CH2OOCH subgroup
168	[CH2][OH0;X2][OH0;X2][CH0]	peroxide CH2OOC subgroup
169	[CH1][OH0;X2][OH0;X2][CH1]	peroxide CHOOCH subgroup
170	[CH1][OH0;X2][OH0;X2][CH0]	peroxide CHOOC subgroup
171	[CH0][OH0;X2][OH0;X2][CH0]	peroxide COOC subgroup
1700	[C;H0,H1,H2][OH0;X2][OH0;X2]	exception: special peroxide CHn-O-O- subgroup when second carbon atom at end is already matched to another group; in this case we map CHnOO group to the CHOOC subgroup and deduct one alkyl (CHn) group detected later (if possible)
1710	[OH0;X2;A]-[OH0;X2;A]	exception: peroxide -O-O- subgroup when both carbon atoms at ends are already matched to other groups; in this case we map -O-O- to the COOC subgroup and deduct two alkyl (CHn) groups detected later (if possible)
1120	[CX4;H2]([OX2;H1])[OX2;H1]	geminal diol case (CH2 aliphatic)
1121	[CX4;H1]([OX2;H1])[OX2;H1]	geminal diol case (CH1 aliphatic)
1122	[CX4;H0]([OX2;H1])[OX2;H1]	geminal diol case (C aliphatic)
154	[CH2;X4;R0][OH0;X2;R0][CH2;X4;R0;$([CH2][OH0][CH2][CH2][OH0][CH2])]	oxyethylene group (-CH2-O-CH2-) in oligomers like Poly(ethylene glycol), PEG; has to have at least two of these groups in succession
27	[CX4;H2;r5;R1;$([CH2;R1][CH2;R1][CH2;R1])][OH0;r5;R1]	tetrahydrofuran (oxolane), special ether group: THF[CH2O]
24	[CX4;H3][OH0;X2]	ether CH3O-
25	[$([CX4;H2]([A!O]))][OH0;X2]	ether -CH2O-; first prefer a matching option where the C is not bonded to an OH group
25	[CX4;H2][OH0;X2]	ether -CH2O-

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

Table 1. Continued.

Key	SMARTS	Description and remarks
26	[$([CX4;H1]([A!O])[A!O])][OH0;X2]	ether CHO-; first prefer a matching option where the C is not bonded to an OH group
26	[CX4;H1,H0][OH0;X2]	ether CHO-, as exception also for ether C-O- (with zero H)
5	[CH2]=[CH1]	alkene CH2=CH subgroup
6	[CH1]=[CH1]	alkene CH=CH subgroup
7	[CH2]=[CH0]	alkene CH2=C subgroup
8	[CH1]=[CH0]	alkene CH=C subgroup
70	[CH0]=[CH0]	alkene C=C subgroup
149	[$([CX4;H3]([OH1X2]))]	CH3 alkyl attached to OH (while not counting OH)
150	[$([CX4;H2]([OH1X2]))]	CH2 alkyl attached to OH (while not counting OH)
151	[$([CX4;H1]([OH1X2]))]	CH1 alkyl attached to OH (while not counting OH)
152	[$([CX4;H0]([OH1X2]))]	C alkyl attached to OH (while not counting OH)
17	[cX3][OH1;X2]	aromatic carbon alcohol ACH subgroup (phenol)
153	[OH1;X2;A]	OH group; must always be attached to aliphatic carbon
9	[cH1]	aromatic hydrocarbon ACH subgroup
10	[cH0]	aromatic hydrocarbon AC subgroup
1	[CX4;H3,H4]	CH3 standard alkyl; as exception also for CH4
2	[CH2]	CH2 standard alkyl
3	[CH1]	CH1 standard alkyl
4	[CH0]	C standard alkyl
2502	[OH0;X2]	exception: map ether oxygen without carbon as ether -CH2O- minus CHn
2501	[oH0;X2]	exception: map aromatic oxygen without carbon as ether -CH2O- minus CHn
2002	[OX1;H0]	exception: map carbonyl =O as aldehyde group minus the CHn group

Accounting for Temperature, Intramolecular, and Non-additivity effects (EVAPORATION) model (Compernolle et al., 2011), the model by Nannoolal et al. (2008), hereafter called Nanoolal method, and the SIMPOL method (Pankow and Asher, 2008a). The source code for the original and subsequent releases of UManSysProp is available from an online repository (see Sect. 4. We note that UManSysProp also includes a few tools for aerosol mixture predictions, including a version of the AIOMFAC model and related SMARTS patterns for generating AIOMFAC input data. However, the list of SMARTS patterns for AIOMFAC in UManSysProp differs in several ways from the more extensive SMARTS list and related S2AS program introduced in this study. Specifically, Topping et al. (2016a) do not follow the same SMARTS priority order, use a different, less comprehensive approach for handling matching exceptions, and the way pure alcohol compounds are detected and mapped may differ as well for more complex polyols. Hence, SMARTS codes from their study are not directly transferrable for use in the S2AS tool.

Several studies have compared liquid-state pure-component vapour pressure prediction methods suitable for SOA systems alongside with critical evaluations of existing experimental data for the particularly relevant semi-volatile and low-volatility compounds. It has been shown that, when applicable, the EVAPORATION method and the Nanoolal method are among the

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

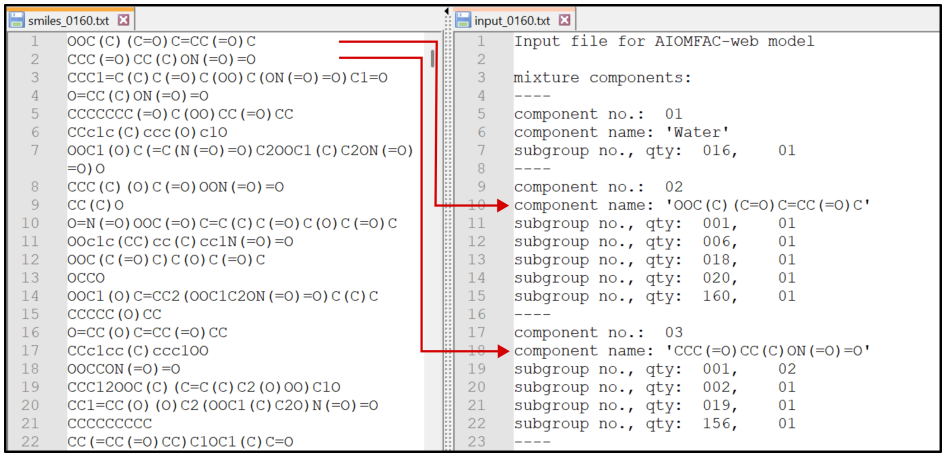


Figure 5. A sample subset of input SMILES strings on the left hand side and their corresponding output functional subgroups generated by the S2AS tool on the right hand side. The generated text file can serve as input file for the AIOMFAC model alongside with mixture composition information.

best performing options in those volatility ranges of particular importance for the gas–particle partitioning of SOA (e.g., Barley and McFiggans, 2010; O'Meara et al., 2014; Bilde et al., 2015). The Nanoolal method is more versatile in terms of the variety of functional groups and chemical elements covered, while the EVAPORATION method is constrained to compounds containing the elements $\mathrm{C, H, O, N}$. However, these are also the main elements supported in AIOMFAC and we opted to use the EVAPORATION method as our first choice for the lumping framework and the gas–particle partitioning calculations.

The parameterization of the temperature-dependence of pure-component vapour pressures used in our work is identical to the two-parameter relation introduced for the EVAPORATION model (Compernolle et al., 2011):

$$\log_{10}\left[\frac{p_j^\circ}{p^{\mathrm{ref}}}\right] = A_j + \frac{B_j}{T^\kappa}\,. \tag{1}$$

Here, p_j° denotes the pure-component, liquid-state (saturation) vapour pressure of component j in units of atmospheres (atm) and $p^{\mathrm{ref}} = 1\ \mathrm{atm}$ is the unit reference pressure. T is the temperature (K), and A_j and B_j are two component-specific parameters. A common value of $\kappa = 1.5$ was adopted based on optimization tests by Compernolle et al. (2011). It was shown to be appropriate for estimating vapour pressure of hydrocarbons with or without heteroatoms across a wide temperature range. The A_j and B_j values are directly predicted by the EVAPORATION model. Alternatively, they can be determined by running any pure-component vapour pressure prediction method at two sufficiently distinct temperatures (typically $\Delta T > 30\,\mathrm{K}$), including the temperature interval of interest, followed by solving the system of two linear equations for A_j and B_j. The use of Eq. (1) enables the flexibility of p_j° estimation at any reasonable temperature for the given set of input molecules, serving as a key input to the gas–particle partitioning model. Computing the A_j, B_j parameters for all organic aerosol system components once eliminates the need for calling the EVAPORATION model repeatedly when the temperature changes, thus improving computational efficiency.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

This approach streamlines the application of distinct SMILES-based tools for (any) pure-component properties and enhances the flexibility of the gas–particle partitioning model in terms of its readiness for computations being carried out over a range of temperatures. In summary, the UManSysProp pure-component property models (written in Python) are run for a given list of SMILES characterizing a system. The A_j and B_j parameters and corresponding SMILES of all organic system components are then written to a text file for read-access by other tools. In particular, the list of parameters is read by the Fortran code of the AIOMFAC equilibrium model, which makes subsequent use of Eq. (1) to obtain the p_j° at a temperature of interest. The pure-component vapour pressure files are also used in the 2D lumping framework described next.

2.2 2D lumping framework

We implemented a new variant of a two-dimensional product lumping framework with the aim to (1) categorize and visualize a representation of all oxidation and fragmentation products from an organic aerosol system at a given point in time and (2) to enable an objective, yet adjustable, selection of surrogate species for a reduced-complexity representation of the system. We constructed a 2D space for mapping the entire aerosol component system using the logarithm of the pure-component vapour pressure, $\log_{10}\left[p_j^\circ/(1\,\mathrm{Pa})\right]$, as the volatility dimension (x-axis). This choice is similar to that of 1D VBS models and several 2D VBS variants, which typically either use the pure-component saturation vapour concentration (C_j° in units of $\mu\mathrm{g\,m^{-3}}$) or the effective saturation concentration ($C_j^\star$ in units of $\mu\mathrm{g\,m^{-3}}$) on a logarithmic scale as volatility dimension (Donahue et al., 2006, 2011). Assuming the ideal gas law to apply, pure-component vapour pressures and pure-component saturation concentrations can be inter-converted via (Zuend and Seinfeld, 2012)

$$C_j^\circ = p_j^\circ \frac{M_j}{RT} \cdot 10^9 \left[\mu\mathrm{g\,kg^{-1}}\right]. \tag{2}$$

Here, M_j is the molar mass ($\mathrm{kg\,mol^{-1}}$) and R the universal gas constant ($\mathrm{J\,mol^{-1}\,K^{-1}}$). Several metrics exist for representations of an organic molecule's effective polarity, including the elemental O:C ratio and the average oxidation state of carbon , denoted by $\overline{\mathrm{OS}}_\mathrm{C}$ (Kroll et al., 2011). In the case of atmospheric organics consisting of the elements carbon, hydrogen, nitrogen and oxygen, typically the approximation $\overline{\mathrm{OS}}_\mathrm{C} \approx 2\cdot(\mathrm{O:C}) - \mathrm{H:C} - 5\cdot(\mathrm{N:C})$ applies (Kroll et al., 2011). Our framework offers those metrics as optional choices, yet our preferred choice for the polarity axis (y-axis) of the lumping framework is to use a metric based on the logarithm of an activity coefficient ratio (ACR) (as detailed in Sect. 2.2.1 and defined by Eq. 5).

In the context of gas–particle partitioning in aerosol systems as a main application of the 2D lumping framework, one way to guide appropriate choices for the two dimensions of the framework is to consider the main factors governing absorptive gas–particle partitioning (Pankow, 2003). Zuend et al. (2010) derived that for equilibrium gas–particle partitioning involving an ideal gas phase and a single (liquid) condensed phase, the following relationship must hold for $K_j^{\mathrm{PM},(n)}$, the equilibrium partitioning coefficient on a molar basis:

$$K_j^{\mathrm{PM},(n)} = \frac{x_j^{\mathrm{PM}}}{p_j} RT = \frac{1}{\gamma_j^{(x)} p_j^\circ} RT. \tag{3}$$

Here, x_j^{PM} is the mole fraction of component j in the liquid particulate matter (PM) phase, p_j is the partial pressure in the gas phase (ideal gas assumption), and $\gamma_j^{(x)}$ is the activity coefficient in the liquid phase (superscript (x) denotes mole-fraction-

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

based quantities). While p_j° indicates the importance of a component's saturation vapour pressure, $\gamma_j^{(x)}$ indicates the influence of nonideal mixing in the liquid phase. The degree of nonideal mixing depends both on the molecular properties of j as well as its interactions with all other molecules in solution. In this context, a polar organic compound present in an aqueous phase will exhibit a lower activity coefficient than a nonpolar compound. Hence, activity coefficients offer a way to express a component's affinity for less or more polar liquid media. This supports the choice of proxies for polarity as the second dimension of our 2D framework. We note that for aerosol systems undergoing liquid–liquid phase separation (LLPS), Eq. (3) holds when appropriately modified, such as by introducing a phase-abundance-weighted effective activity coefficient (Zhang et al., 2024).

2.2.1 Activity coefficient ratio as polarity metric

The idea of using an activity coefficient ratio (ACR) as a polarity metric is inspired by liquid–liquid equilibrium (LLE) thermodynamics. In a macroscopic LLE state, the chemical potential of a component present in both coexisting phases must be identical. Consequently, the way a component j partitions between two liquid phases, α and β, can be described using an equilibrium partitioning constant $K_j^{(x)}$ (Zuend and Seinfeld, 2013; Topping and Bane, 2022):

$$K_j^{(x)} = \frac{x_j^\alpha}{x_j^\beta} \overset{\text{LLE}}{=} \frac{\gamma_j^{\beta,(x)}}{\gamma_j^{\alpha,(x)}}. \tag{4}$$

Here, x_j^α and x_j^β are mole fractions in phases α and β; $\gamma_j^{\alpha,(x)}$ and $\gamma_j^{\beta,(x)}$ are the activity coefficients in those phases under LLE conditions. Equation (4) indicates that knowledge of the $\gamma_j^{\beta,(x)}/\gamma_j^{\alpha,(x)}$ ratio provides an accurate representation of the thermodynamic phase preference of organic components by quantifying the relative enrichment or depletion of a component via the equivalent mole fraction ratio, with a value greater than 1 indicating enrichment in phase α. Since activity coefficients depend on a phase's mixture composition established by all components, the values are sensitive to various forms of molecular interactions (e.g., dipol–dipol, dispersion) and reflect the chemical affinity for a phase. That is, activity coefficients are influenced by molecular structure properties, such as whether an oxygen-bearing functional group is more polar (e.g., hydroxyl, carboxyl) or less polar (e.g., ether, ester) and how it interacts with other organic and inorganic components present. Thus, the ACR encompasses more detailed functional group characteristics than simpler metrics like the O:C ratio.

In this work, the ACR of a component j is based on a prediction of the component's activity coefficient when present as a dilute solute in a weakly polar organic reference solvent, here we use 1,2-hexanediol, relative to that of being a dilute solute in a strongly polar reference solvent, here water. Figure S1 illustrates the binary solution setup used to calculate the activity coefficient ratio (ACR) as a polarity metric. The figure depicts two scenarios: an organic solute dissolved in water (left) and the same solute dissolved in 1,2-hexanediol (right), representing polar and less polar solvents, respectively. This metric is therefore similar to the well-established octanol–water partitioning coefficient (e.g., Kamlet et al., 1988; Wienke et al., 1998), but our choice of organic reference solvent differs. We elected to use a slightly more polar organic than octanol, with 1,2-hexanediol serving as a more typical representation of an organic-rich phase medium in aerosols.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

While inspired by the LLE isoactivity condition (Eq. 4), the procedure of obtaining our ACR metric differs from that of solving a ternary solute–solvent-1–solvent-2 LLE problem. This is because in the ternary LLE system the present solvents may be partially miscible, while we choose to compute the activity coefficients of each component independently based on evaluating two binary solute–solvent mixtures. Using binary systems as a gauge is computationally simpler and substantially faster. Specifically, our polarity metric ACR is defined by the following dimensionless quantity:

$$\log_{10}\left[\frac{\gamma_{j,\mathrm{hex}}^{(x)}}{\gamma_{j,\mathrm{w}}^{(x)}}\right]. \tag{5}$$

Here, $\gamma_{j,\mathrm{hex}}^{(x)}$ is the (predicted) mole-fraction-based activity coefficient of solute j in a binary mixture with the solvent 1,2-hexanediol, in which the mass fraction of j is $w_j = 0.01$ (therefore, $w_{\mathrm{hex}} = 0.99$). Analogously, for $\gamma_{j,\mathrm{w}}^{(x)}$, except for the solvent being water. These activity coefficients are typically evaluated at a reference temperature of 298.15 K. Given this definition, one can think of these two separate activity coefficients, as well as their ratio, as pure-component properties. In principle, those values would only need to be computed once for each component (each SMILES code) and could then be saved and retrieved from a look-up table.

The selection of the two reference solvents provides a robust basis for characterizing the behaviour of organic components across a wide range of polarities spanning several orders of magnitude (see Sect. 3). Water represents a highly polar solution environment, typical of that in aqueous aerosol phases, while 1,2-hexanediol (O:C ratio of 0.333) serves as a proxy for moderately oxidized organic aerosol components. The latter would preferentially partition into an organic-rich phase in the presence of LLPS. In our framework, the AIOMFAC model is employed to compute activity coefficients for Eq. (5). Since AIOMFAC is run only for binary, single-phase systems, the calculations are fast; they are comparable or faster than the pure-component vapour pressure predictions with UManSysProp. Of note, for a system consisting of tens of thousands of organic components, computing activity coefficients of using the related multicomponent mixtures (containing as many components) with AIOMFAC, is prohibitively slow due to the many functional groups present, all the possible group–group and molecule–molecule interactions that need to be summed over and the associated computer memory requirements. In contrast, computing the ACR values based on the binary solute–solvent mixtures for all those components is fast and small in memory footprint.

2.2.2 Methods for surrogate selection

We implemented four distinct methods to analyze organic aerosol data in the 2D lumping framework and to objectively select a set of surrogate components. These methods are designed to reduce the complexity of the gas–aerosol system while preserving important physicochemical aspects, such as the conservation of total mass concentration and the consideration of the system's diversity in terms of volatility and polarity ranges. The four methods are: (a) the grid cell midpoint method, (b) the grid cell medoid method, (c) the grid cell mass-weighted medoid method, and (d) the k-means-based medoids method. The latter is a grid-independent 2D clustering method.

The pure-component property prediction tools described in Sect. 2.1 provide high-fidelity estimates of the necessary values for positioning components in the 2D space. This 2D space is then subdivided into a number of grid cells (or clusters) based

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

on the desired level of system resolution and the related targeted reduction in complexity, allowing for flexible adaptation to different research questions and computational resources. The 2D space division is accomplished by specifying the targeted number of rows and columns, followed by identification of the component coordinates that set the upper and lower coordinate limits within the range that should be gridded (see use of a volatility threshold below). The grid resolution is adjustable, but typical choices range from 2×1 to 20×10 in terms of setting the number of grid subdivisions and associated $n_x \times m_y$ grid cells.

A crucial aspect of our methodology is the primary focus of the gridded domain on the compound volatility ranges of interest for substantial partitioning of organics to the particle phase. These volatility ranges includes semi-volatile (SVOC), low-volatility (LVOC) and extremely low-volatility (ELVOC) organic compounds, while intermediate volatility (IVOC) and volatile organic compounds (VOC) are too volatile under typical tropospheric aerosol mass loading conditions (Donahue et al., 2012; Zuend and Seinfeld, 2012). As such, an adjustable high-volatility vapour pressure threshold (p_{high}) is introduced, typically set to $\log_{10}\left(p_{\text{high}}/[1\,\text{Pa}]\right) > -1$. Regardless of the polarity range, all compounds with $p_j^\circ > p_{\text{high}}$ are identified and lumped into a single high-volatility surrogate component using the mass-weighted medoid method defined below. Examples for this are shown in Sect. 3, e.g., Figs. 7 and 8. Since the p_{high} threshold value is an input parameter, this special VOC lumping step can also be avoided entirely by setting a very high threshold value. Similarly, a low-volatility threshold and related lumping of LVOC and ELVOC compounds into a single (or a few), quasi-nonvolatile surrogate could be introduced to focus the majority of surrogates on representing the SVOC range at comparably higher resolution. However, in this work we have not included a low-volatility threshold since we are interested in a diverse surrogate-based representation of all PM-relevant volatility ranges to better resolve potential trends in polarity with decreasing volatility accross the SVOC, LVOC, and ELVOC ranges.

Our lumping process adheres to the principle of mass conservation for the entire system. This is achieved by aggregating the mass concentrations of components within each grid cell or cluster into the selected surrogate component. This approach differs from methods that choose to conserve the number of carbon atoms during lumping. The surrogate component selection process for each of the four methods is visually depicted in Fig. 6. The shown examples illustrate how one surrogate component is chosen within a single grid cell (or cluster in case of k-means).

2.2.3 Grid cell midpoint method

The midpoint method operates on the simple principle that the component closest to a grid cell's midpoint coordinates should be a reasonable choice representing all other components within the same cell. This method involves determining the normalized distance of each grid cell component's location to the grid cell midpoint. The squared Euclidean distance, $D^2_{\text{mid},i}$, of each component i within a grid cell is calculated by

$$D^2_{\text{mid},i} = \left[\frac{x_{\text{mid}} - x_i}{x_{\text{range}}} \cdot \chi\right]^2 + \left[\frac{y_{\text{mid}} - y_i}{y_{\text{range}}}\right]^2. \tag{6}$$

Here, x_{mid} and y_{mid} denote the midpoint coordinates of the grid cell, x_i and y_i are the coordinates of component i, and x_{range} and y_{range} denote the magnitudes of the x and y axes ranges of the 2D space over which the lumping grid was placed. They normalize the distances expressed along the two axes. The x-to-y aspect ratio χ is introduced to scale the normalized length

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

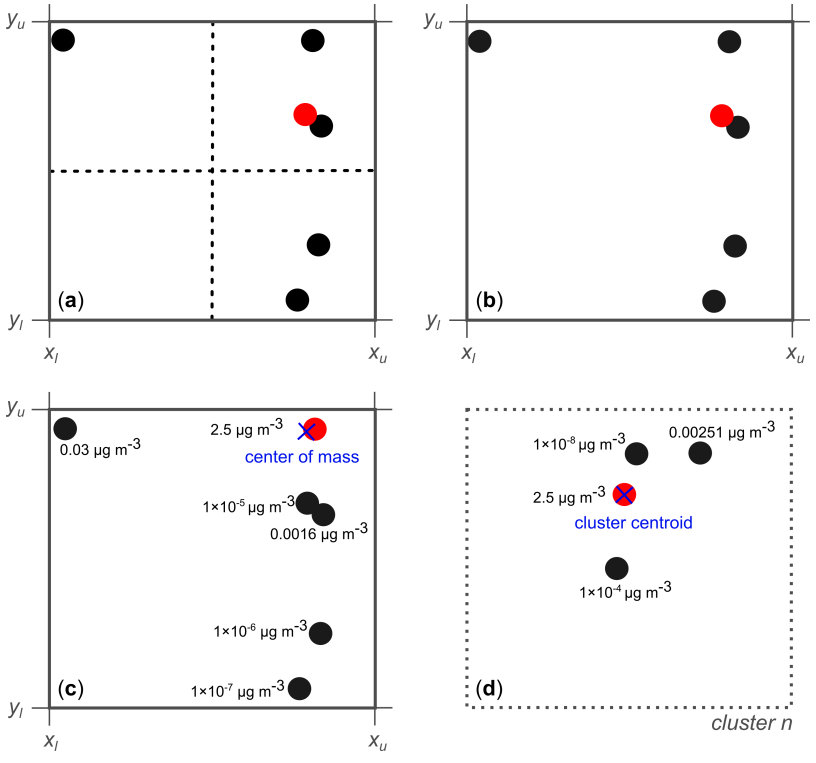


Figure 6. Illustration of the four introduced methods for selecting a single surrogate component (marked in red) within a specific grid cell or cluster of the 2D space. x_l and x_u indicate the lower and upper x-axis values of the grid cell; analogously for y_l and y_u. (a) The midpoint method selects a surrogate based on the component closest to the geometric centre of the grid cell, denoted by the dotted lines. (b) The medoid method selects the component with the smallest cumulative Euclidean distance from all other components. (c) The mass-weighted medoid method prioritizes components with higher mass concentrations. It determines the most representative surrogate as the one closest to the centre of mass established by the components in the cell. (d) The k-means-based medoid method is independent of a grid. It iteratively assigns components to clusters, then identifies a cluster's optimal surrogate as the nonzero-mass component closest to the cluster centre (indicated by the blue $\times$ symbol). Individual clusters are not confined to predetermined grid cells (in (d) denoted by dotted outline); here, the four points are assumed to form a specific cluster n.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

scales of the grid space. That is, χ is the (prescribed) multiplying factor of the normalized, dimensionless value range along the x-axis that is regarded as equivalent to the normalized, dimensionless range along the y-axis. To express the importance of volatility in gas–particle partitioning, we elected to set a 2:1 aspect ratio as the default choice. If the number of grid lines along the x-dimension is already set as twice that of the y-dimension (e.g., 8 by 4), then χ will account for this ($\chi = 1$ in that case). In this study, the x-coordinate refers to the logarithm of the pure-component vapour pressure ($\log_{10}(p_i^\circ/[1\,\mathrm{Pa}])$), while the y-coordinate corresponds to one of the proxies for polarity (ACR, O:C ratio, $\overline{\mathrm{OS}}_\mathrm{C}$). The value ranges of the components along the x and y coordinates may differ substantially, possibly by several orders of magnitude. Therefore, we chose to normalize and scale the magnitudes of the x and y axes ranges relative to each other, as shown by Eq. (6). The x_range ($= |x_\mathrm{max} - x_\mathrm{min}|$) is determined by the maximum and minimum x-coordinates from the set of all system components that contribute nonzero mass concentration and belong to the regular lumping space, i.e., those not lumped to the single high-volatility surrogate compound. The y_range is determined analogously. Finally, within each grid cell containing at least one component, the component of minimum D^2_mid is selected as surrogate and the mass concentrations of all other grid cell members are lumped (additively) into the surrogate. Except for the method-specific expressions for computing a component's distance metric within a grid cell, the variable definitions and steps outlined for this method also apply to the other methods described in the following sub-sections.

The grid cell midpoint method is straightforward to understand and implement. It has the advantage of providing unbiased coverage of the 2D space by ensuring that surrogate components are nearly evenly distributed across the grid domain. The approach demonstrates scalability, as it can be easily applied to systems with a large number of components, making it suitable for complex mixture analyses. Additionally, the method offers computational efficiency since the calculation of midpoints and distances is relatively fast, even for large datasets. However, this method may not always select the most representative component, especially when a low-resolution grid is applied in which the few cells may exhibit unevenly distributed components (e.g., all components clustered around the lower left corner of a cell). This method is visualized in Fig. 6a.

2.2.4 Grid cell medoid method

The grid cell medoid method operates by selecting the medoid component of each grid cell as the surrogate. The medoid member is the component in closest cumulative proximity to all other components of the same cell. Therefore, the main step involves calculating the cumulative squared Euclidean distance of each component from all other components of the grid cell, as follows:

$$\Sigma D^2_{\mathrm{med},i} = \sum_{j=1}^{m} \left[\frac{x_j - x_i}{x_\mathrm{range}} \cdot \chi \right]^2 + \left[\frac{y_j - y_i}{y_\mathrm{range}} \right]^2 . \tag{7}$$

Here, index j covers the $1, \ldots, m$ components of the grid cell, with other variables as defined for Eq. (6).

The medoid method offers potential advantages over the midpoint method. First, it may provide better representation by selecting a surrogate that is located close to most other components, which is especially of importance in the case of a grid space with only a few large cells filled with geometrically uneven component distribution. Second, the medoid method demonstrates robustness to outliers, being less influenced by extreme values or uneven clustering of points within a cell compared to the midpoint. In comparison to the midpoint method, the medoid method is computationally more costly, especially for cells that

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

contain a large number of components. However, in practice this is rarely a concern. Figure 6b illustrates the grid cell medoid method.

2.2.5 Grid cell mass-weighted medoid method

The mass-weighted medoid method prioritizes surrogate selection based on a combination of a component's importance, as measured by its mass concentration, and the cumulative distance from other components of a grid cell, as in the unweighted medoid method. Using such a mass-weighted approach is particularly useful in the case of systems consisting of many components, yet with the majority of the mass contributed by a small minority of molecules. The mass-weighting ensures that the important components are more likely to be selected as surrogates, so that their particular molecular properties are then also more appropriately considered in subsequent equilibrium partitioning computations. The core of this method lies in the calculation of the squared distance between a component and the coordinates of the centre of mass of the cell, $D^2_{\text{wmed},i}$, as follows:

$$D^2_{\text{wmed},i} = \left[\frac{x_{\text{mc}} - x_i}{x_{\text{range}}} \cdot \chi\right]^2 + \left[\frac{y_{\text{mc}} - y_i}{y_{\text{range}}}\right]^2, \tag{8}$$

with

$$x_{\text{mc}} = \sum_{j=1}^{m} w_j \cdot x_j, \quad y_{\text{mc}} = \sum_{j=1}^{m} w_j \cdot y_j, \quad \text{and} \quad w_i = \frac{C_i}{\sum_{j=1}^{m} C_j}. \tag{9}$$

Here, x_{mc} and y_{mc} are the coordinates of the centre of mass of the grid cell under consideration. These coordinates are determined by computing the mass-weighted average coordinates of the $j = 1, \ldots, m$ grid cell components, as shown by Eq. (9), only requiring information about the cell's components of nonzero mass concentration. w_j and C_j represent a component's grid-cell mass fraction and mass concentration, respectively. Components of large mass concentration relative to others in a grid cell have the effect of "pulling" the centre of mass toward their location, thereby benefiting from a smaller $D^2_{\text{wmed},i}$. Among the grid cell components of nonzero mass concentration, the component of minimum $D^2_{\text{wmed},i}$ value is identified as the surrogate. Components of zero mass concentration are excluded from the calculation and from being selected as surrogate.

This method strongly favours the selected surrogate components to be from the subset of most abundant species in the gas–aerosol system. The mass-weighted medoid method offers several advantages. The method provides improved representation of dominant components. The resulting lumped system is likely to reflect bulk aerosol properties more accurately, especially when only a low-resolution grid is applied. To aid in understanding the interplay between mass concentration and spatial distribution in the surrogate selection process, Fig. 6c exemplifies how the component of highest mass concentration, which is also closest to the centre of mass within the grid cell, is selected as the surrogate component.

2.2.6 *k*-means-based clustering method

The mass-weighted, k-means-based medoid clustering method, for brevity hereafter referred to as the k-means method, is employed to generate a predefined number of centroids (k clusters) in the 2D space. The clustering process is implemented

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

using a variant of the weighted k-means clustering algorithm; specifically, subroutine kmeans_w_01, from the Fortran 90 implementation provided by Burkardt (2008). The Fortran code is based on the theory, algorithm and existing code from the works by Sparks (1973) and Hartigan and Wong (1979). This algorithm iteratively reassigns points (here chemical components) to clusters, minimizing the total energy of the system. Refer to Sparks (1973) and Hartigan and Wong (1979) for a detailed description of the k-means method and its variants.

In our approach for surrogate selection with the k-means method, we combine the mass-weighted clustering with a final, distance-based surrogate component selection process. Figure 6d visualizes an example cluster's surrogate selection. Specifically, k-means returns the coordinates of a predefined number of cluster centres, which usually do not coincide with any actual component's coordinates. Among each cluster's population, we then select the component of nonzero mass concentration that is located closest to the cluster centre coordinates as the cluster's surrogate. Similar to the mass-weighted medoid approach, the distance of each cluster component to the centre is determined by the squared Euclidean distance based on normalized coordinates. In outcome, this is therefore akin to (but algorithmically not identical with) the method of k-medoids clustering.

By selecting surrogate species that are typically centrally located within their clusters and of significant mass concentration, this approach ensures that the set of surrogate components represents the overall physicochemical characteristics of the system well. This approach is particularly valuable in atmospheric chemistry, where reducing the complexity of chemical mechanisms while retaining properties of the most abundant components is crucial, since mass or number concentration of components is a critical factor in understanding subsequent partitioning or chemical reaction behaviour.

In our implementation, the clustering process begins by setting the desired number of k clusters. The initial cluster centre coordinates are then set based on the grid cell midpoints (since in our code the grid-based methods are run prior to running k-means). For comparison with the grid-based methods, we typically chose the cluster numbers as being equal to the number of populated grid cells. However, the k-means clustering method can also be run independently from the gridded approaches. If a higher number of clusters is set as target than the number of populated grid cells, additional initial cluster centre coordinates are generated using pseudo-random coordinates within the scaled 2D space. This initialization step ensures a broad distribution of initial cluster centres across the normalized 2D space. Components associated with the special high-volatility surrogate are filtered out, since they are marked as being part of a special cluster; these are not considered during the k-means clustering.

An innovation of the weighted k-means variant is the incorporation of, in our case, mass-concentration-based weighting during the algorithm's assignment of components to clusters. When k-means computes each cluster's "energy" and iteratively assigns components to a certain cluster, the weights are factored in. As discussed in Sect. 3, actual example cases indicate that the mass concentrations of components may range over several orders of magnitude. Therefore, the weighting aids in prioritizing components with relatively high mass concentrations as potential cluster surrogates, ensuring that the clustering process is not solely based on spatial proximity. The algorithm considers the entire dataset simultaneously and can effectively handle cases in which data points are not uniformly distributed in a 2D space. While we favour weighted k-means clustering, if desired, one can assign each component the same weight, thereby returning to the non-weighted k-means method.

Overall, the k-means method offers several advantages. The method can identify natural groupings in the data, independent of the arbitrary limits of an imposed grid, potentially leading to more meaningful, unbiased surrogate selection. However, this

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

method is computationally the most costly of the four outlined in this study. Users can adjust the targeted number of k-means clusters to strike a balance between model simplicity and resolution of the original data. We note that the computational cost of this method is relatively insensitive to the number of clusters targeted.

3 Results and Discussion

The primary goals of this study were to develop efficient and practical tools for calculating pure-component properties, activity coefficients and the gas–particle partitioning for complex systems containing a large number of organic species. To this end, we implemented and evaluated a 2D lumping framework to reduce system complexity while maintaining an adjustable level of accuracy. Additionally, we sought to compare the effectiveness of different lumping methods. These objectives were pursued to advance our understanding of organic species behaviour and improve computational efficiency in geoscientific modelling. The S2AS pure-component property prediction tool was implemented in Python. The 2D lumping framework was implemented in modern Fortran. The generated computer programs are a main product of this work. Related code and data are provided via code repositories and related archives, see Sect. 4.

3.1 Example systems: α-pinene and toluene oxidation products

To demonstrate the application of the new chain of software tools, we introduce two example systems showcasing a multitude of chemical components and related properties derived from simulations of (1) an α-pinene ozonolysis system simulated using the MCM (v3.3.1) model and (2) a toluene photo-oxidation system simulated based on a near-explicit GECKO-A mechanism of its gas phase chemistry. The input parameters used for these simulations are listed in Tables S1 and S2. The system component structures were either directly output or subsequently converted to the SMILES format and the component concentrations retrieved from a selected output time of the simulations.

The GECKO-A toluene photolysis system comprises an extensive array of $\sim 68,000$ distinct organic components. Figure 7 provides a visual representation of the obtained system of oxidation and fragmentation products in the 2D polarity versus volatility space. The complete set of components is shown in Fig. 7a (ordered by mass concentration to show the most abundant components on top), while panels (b)–(d) show the application of a 10×5 volatility $\times$ polarity (here using ACR) grid to select surrogate components by the three grid-based methods. Figure 8 shows the corresponding data clustered by the k-means method and the pertaining selection of surrogate components. The comparison of the full system and the various surrogate representations demonstrate visually the framework's ability to simplify complex chemical systems. Despite the massive reduction in the number of species, our lumping framework successfully preserves key physicochemical characteristics of the system of relevance for SOA formation predictions.

The 2D representation of the α-pinene-derived components is shown in Fig. 9, with panel (a) showing the full set of MCM-derived system components. Panels (b)–(d) of Fig. 9 show the surrogate selections based on a 10×5 grid. Since several grid cells are empty in this example, only 36 surrogates are determined at the chosen grid resolution. Figure 10 demonstrates the application of the k-means method to this system when using ACR as the polarity axis.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

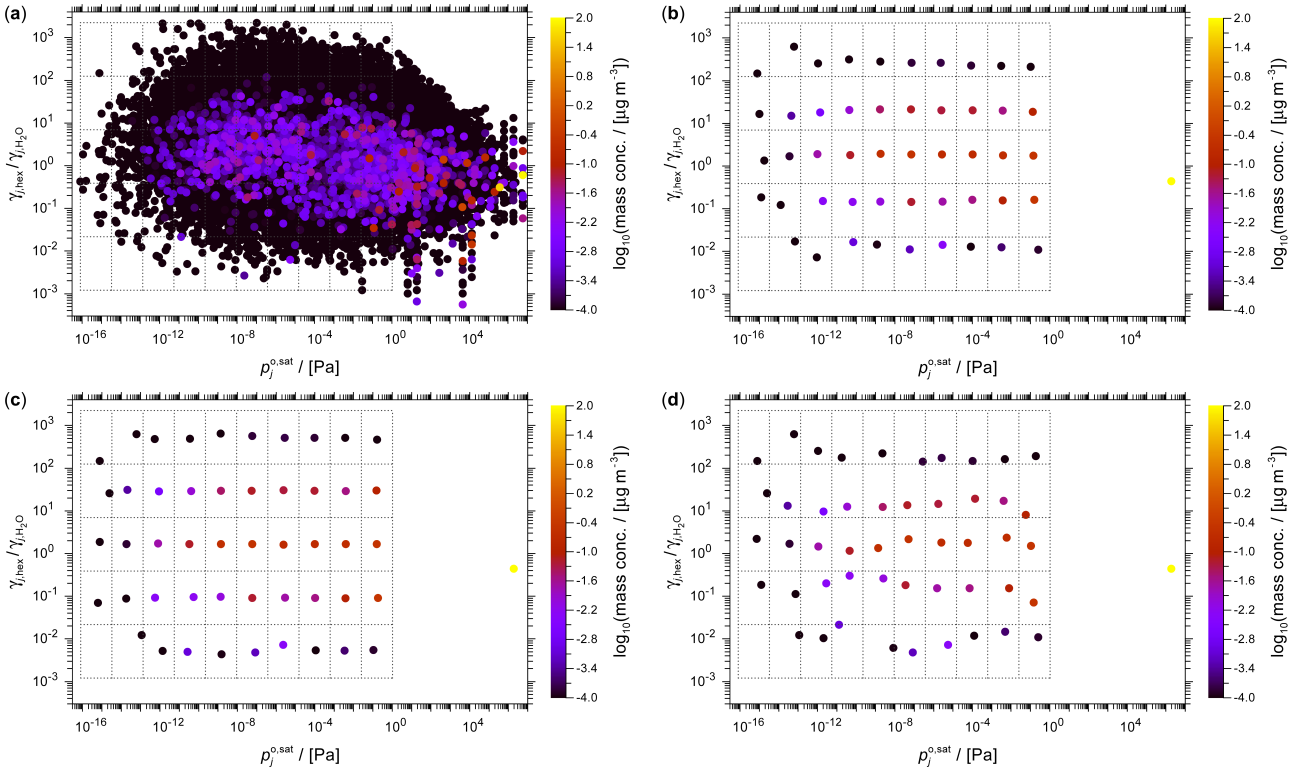


Figure 7. The toluene SOA system shown in the 2D space of activity coefficient ratio versus vapour pressure at 298 K. (a) The full set of ∼68,000 components derived from a simulation based on the GECKO-A mechanism with an overlaid 10×5 grid (dotted) which is used by the different selection methods to determine grid cell surrogates. (b–d) Surrogates selected by (b) the medoid method, (c) the midpoint method and (d) the weighted medoid method.

3.2 Performance of Automated Property Prediction Tools

3.2.1 Functional Group Identification (S2AS Tool)

The S2AS tool demonstrated high accuracy in identifying AIOMFAC functional groups for the α-pinene and toluene oxidation products. The 174 α-pinene-derived products were mapped to AIOMFAC subgroups without needing any exception treatments, while an average of 0.5 exceptions (special SMARTS mappings) per molecule were encountered in case of the ∼68,000 toluene-derived products. The latter exhibited a higher level of nitrogen-containing functional groups, some of which were responsible for most of the exception cases triggered. Given the systematic handling of exception cases, the S2AS tool was successful in mapping all carbon, oxygen and nitrogen atoms to relevant subgroups. This is particularly noteworthy given the structural complexity of many of the oxidation products, which often contain multiple functional groups, branched functionalized chains and ring structures. The S2AS tool is also more reliable than manual classification in cases involving complex

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

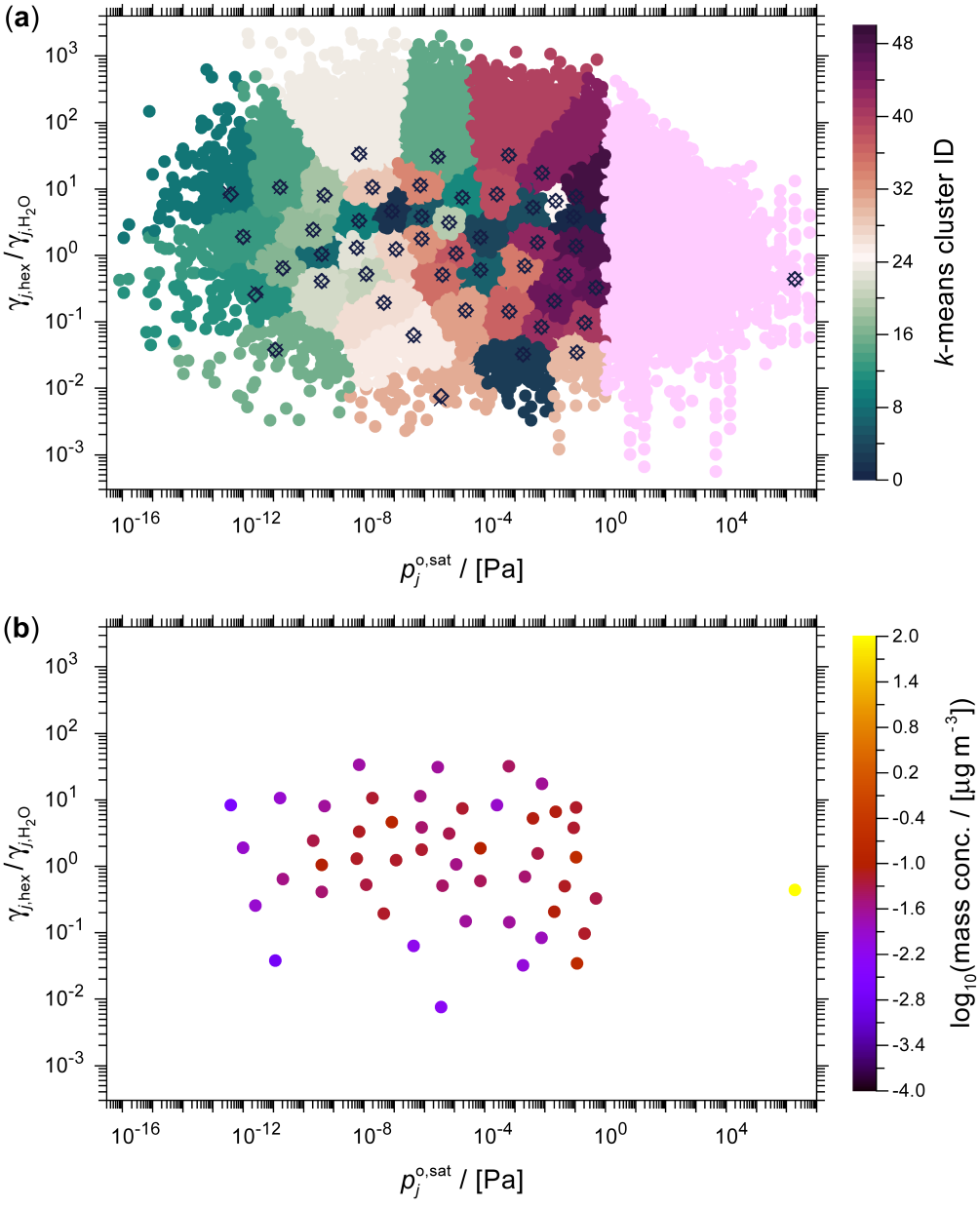


Figure 8. 2D space representations of the toluene-derived SOA system at 298 K using the k-means clustering and surrogate selection method. (a) 49 clusters plus 1 high-volatility cluster (pink data points) are shown with individual cluster members identified by the same colour. The cluster centres are denoted by $\times$ and the corresponding selected cluster surrogates by the $\diamond$ symbols. (b) The lumped mass concentrations of the surrogate components selected by the mass-weighted k-means method.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

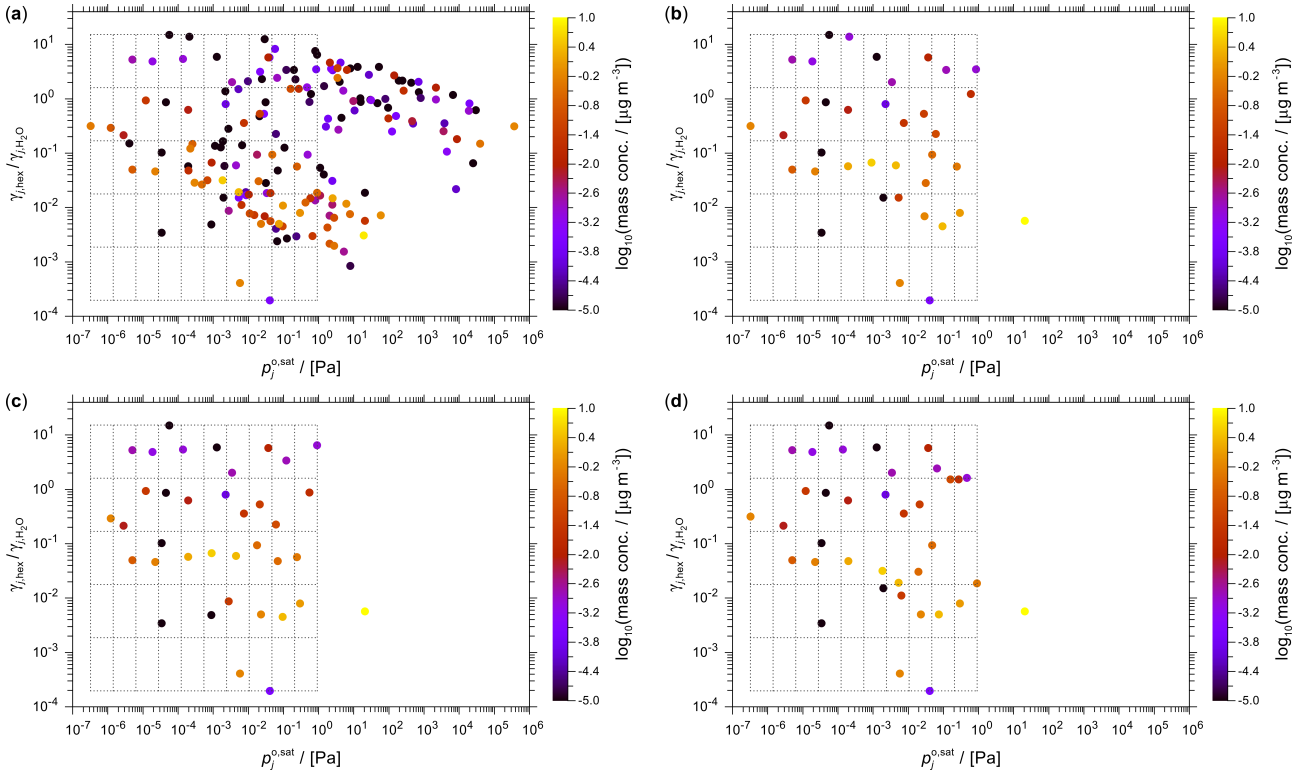


Figure 9. The α-pinene SOA system components shown in the 2D space of activity coefficient ratio versus vapour pressure at 298 K. (a) The full MCM-derived 174 components with an overlaid 10×5 grid (dotted) that is used by the different selection methods to determine grid cell surrogates. (b–d) Surrogates selected by (b) the medoid method, (c) the midpoint method and (d) the weighted medoid method.

molecular structures, for which human experts may occasionally overlook less prominent functional groups or perform inconsistent exception treatments when an imperfect mapping to the available set of AIOMFAC subgroups is present.

The S2AS tool showed high computational efficiency in processing large datasets. It successfully processed the list of 174 α-pinene oxidation product SMILES in less than 0.8 seconds, and the ∼68,000 SMILES from the toluene system in ∼ 5 minutes – a processing rate of ∼ 13,800 SMILES per minute on a single laptop processor core (1 thread, Intel Core i7-10710U CPU). This automatic classification rate represents a tremendous advance over manually assigning AIOMFAC subgroups for each component, which would be an infeasible task for systems containing thousands of components. The S2AS tool's processing speed scales linearly with dataset size. Further speedup via parallelization on multi-core computers is possible, but such enhancements have not been attempted in the current version of the source code.

3.2.2 Vapour pressure estimation

The UManSysProp pure-component vapour pressure estimation tool demonstrated robust performance across the wide range of molecular structures present in the α-pinene and toluene oxidation systems. It successfully parsed the list of SMILES and

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

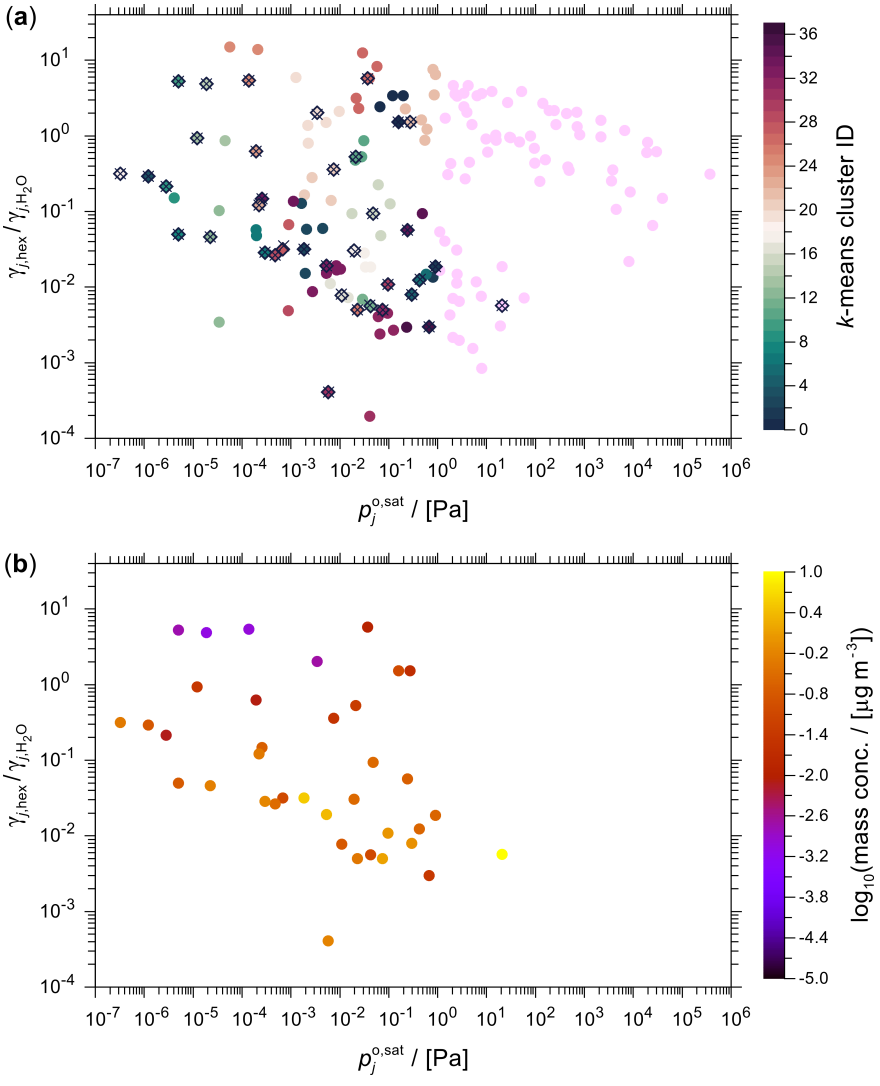


Figure 10. 2D space representations of the α-pinene SOA system at 298 K using the k-means clustering and surrogate selection method. (a) 36 clusters are shown with individual cluster members identified by the same colour. The cluster centres are denoted by $\times$ and the corresponding selected cluster surrogates by the $\diamond$ symbols. Data points coloured pink are members of the special high-volatility cluster. (b) The lumped mass concentrations of the surrogate components selected by k-means.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

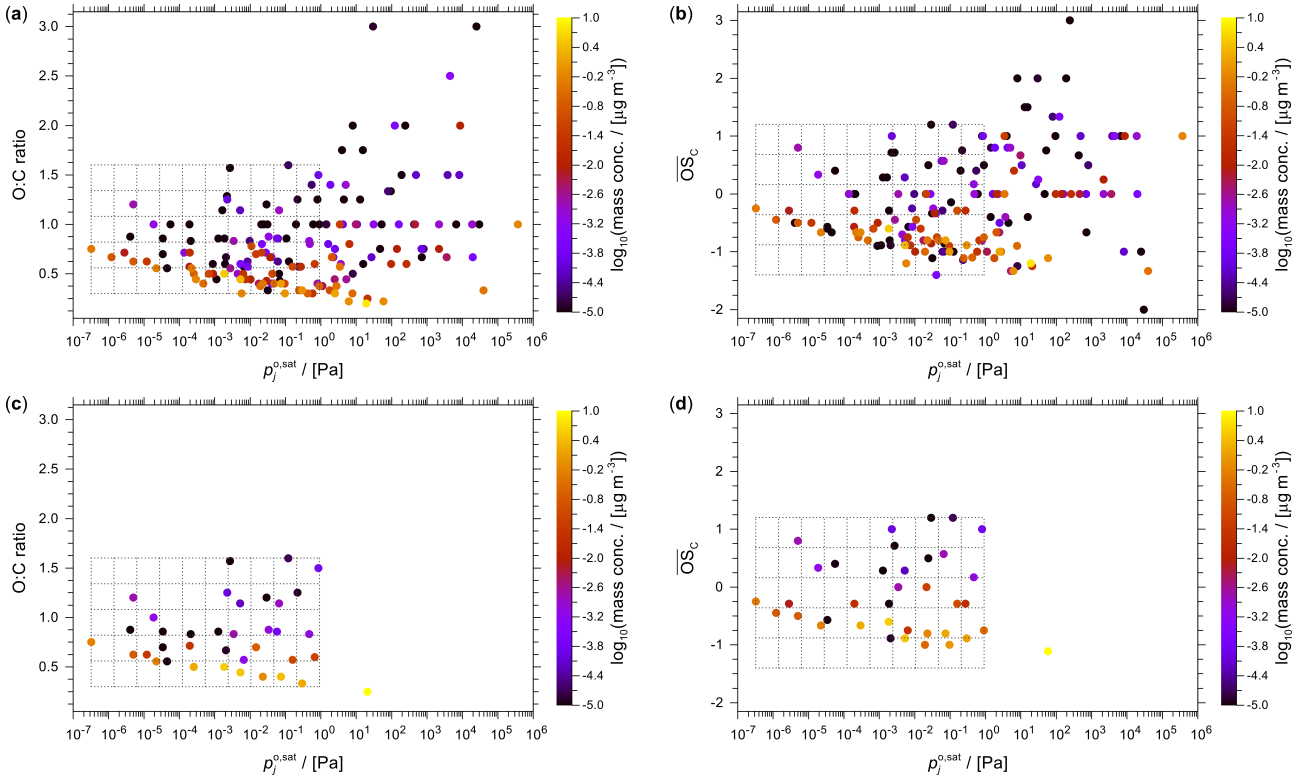


Figure 11. 2D space representations of the α-pinene SOA system at 298 K using (a, c) the O:C ratio or (b, d) the mean carbon oxidation state as polarity axis. A 10×5 grid is shown in case of the full SOA system (panels a, b) as well as its use for the surrogate selection by the weighted medoid method (panels c and d).

predicted the vapour pressures for all molecules. For the examples shown in this study, we used the output from the EVAPORATION method, yet the seven other pure-component vapour pressure prediction methods available from UManSysProp were also completed successfully.

The version of UManSysProp we employed in this work includes a new module we developed to determine a two-parameter temperature dependence parameterization of pure-component vapour pressures using the form of Eq. (1) for each of the vapour pressure methods included (not just EVAPORATION). In a first step, a component's vapour pressure is predicted by each method at seven temperatures equally spaced between 260 and 320 K. In a second step, the two parameters of Eq. (1) are fitted to these data for each method. In the case of EVAPORATION, no fitting is required since one can solve for parameters A_j and B_j using the output from the end points of the temperature range (two equations, two unknowns). If speed is of the essence, the parameter fitting step can also be bypassed for all other methods by solving for the parameters in the same way, usually yielding similar parameter values to those obtained from a more elaborate fit. The single-thread processing of the $\sim$68,000 toluene-derived product SMILES took 482 s on a Intel Core i7-10710U CPU (one thread) for the vapour pressure data creation plus 83 s for determining the parameters A_j and B_j pertaining to each method when bypassing the parameter fitting step.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

This amounts to a SMILES processing rate of $\sim 120\,\mathrm{s}^{-1}$ or ~ 7200 SMILES per minute (including the parameter fitting step reduces the rate to ~ 1650 SMILES per minute). As in the case of the S2AS tool, further speed-up is possible by introducing parallel processing (with good scaling potential) in the case of extensive lists of SMILES.

Together with the introduced S2AS, these two pure-component property prediction tools enable automated and efficient processing of SMILES data for atmospheric and environmental chemistry applications.

3.3 Evaluation metrics for the 2D lumping framework

To facilitate an evaluation of the 2D lumping framework in terms of impacts of polarity axis choices and grid resolutions, we performed 2D lumping at several grid resolutions, each followed by AIOMFAC-based equilibrium gas–particle partitioning computations to generate aerosol properties of interest for a quantitative comparison. Specifically, we calculated the mean absolute percentage error (MAPE) and mean percentage error (MPE) for the resulting SOA mass concentrations as well as the aerosol hygroscopicity parameter κ. Aerosol mass concentration and hygroscopicity are two (among several) insightful characteristics of the (gas–)aerosol partitioning behavior and water uptake potential. The MAPE and MPE are relative deviation metrics defined as follows:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{F_i - A_i}{A_i}\right| \times 100\%, \tag{10}$$

$$\mathrm{MPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{F_i - A_i}{A_i}\right) \times 100\%. \tag{11}$$

Here, A_i represents the actual (observed or reference) value, F_i the predicted value, and n is the number of observations. MAPE is scale-independent, making it useful for comparing overall prediction precision across different datasets (especially datasets free of extreme outliers). MPE measures the average bias in the predictions relative to the reference data. By considering the direction of errors, MPE complements MAPE by highlighting any systematic high or low biases in the model's predictions, which can be crucial for understanding the limitations and potential improvements of each method.

MAPE and MPE were calculated for a selection of grid resolutions (4×2, 6×3, 8×4, 10×5) and polarity axis metrics for both α-pinene-derived SOA and toluene-derived SOA. In order to calculate the evaluation metrics for the toluene SOA system, for which a full-system calculation was not feasible with the AIOMFAC equilibrium partitioning model, the mass-weighted k-means method with the ACR polarity axis and a higher 25×10 grid resolution was used as the reference (benchmark case) for all relative deviation evaluations. A validation check was also carried out using the k-means method for a 40×20 grid resolution (also with ACR as polarity axis) to verify whether the MAPE and MPE values across different (higher) grid resolutions and associated numbers of k-means clusters, were consistent with the reference case. The predicted reference mass concentration at the 25×10 grid resolution (219 surrogates) agrees with that from the 40×20 (612 surrogates) case within a MAPE of $0.1\,\%$, confirming it to be an appropriate reference.

Table 2 compares the relative deviations of different surrogate selection methods for predicting SOA mass concentrations using the selected surrogate species from the α-pinene SOA system. The table lists the MAPE and MPE for the four surrogate

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

selection methods (midpoint, medoid, mass-weighted medoid, mass-weighted k-means) combined with three polarity axis options across different grid resolutions. The corresponding absolute SOA mass concentrations and κ values predicted for this system are listed in Table S3. The weighted k-means method generally performs best, with lowest errors across most grid sizes. Table 3 presents similar MAPE and MPE data but for the predictions of the hygroscopicity parameter κ for the same α-pinene SOA system. Again, the weighted k-means method tends to show the smallest errors overall compared to other surrogate selection approaches. A more detailed discussion of different axis choices and impacts follows in Sect. 3.3.1.

Table 4 provides MAPE and MPE values for SOA predictions from the toluene oxidation system, comparing the same surrogate selection methods and polarity metrics across five grid resolutions. Table 5 shows the corresponding evaluation of the hygroscopicity parameter predictions for the toluene-derived SOA. The corresponding absolute SOA mass concentrations and κ values predicted for this system are listed in Table S4. Deviations from the reference case at lower grid resolutions are generally higher than for SOA mass predictions, but the mass-weighted k-means method still performs relatively well at most resolutions. The larger variability in κ predictions stems in part also from the relatively low absolute values (Table 7). Additionally, the relatively small modelled SOA mass concentrations of ~ 1.2 µgm^{-3} from AIOMFAC for the toluene SOA and of ~ 1.9 µgm^{-3} for the α-pinene SOA contribute to the observed metric fluctuations, as minor absolute differences can result in larger relative errors.

Table 6 compares the predicted aerosol mass concentrations for the toluene and α-pinene SOA systems at different water activities, using the weighted k-means method or, in case of α-pinene SOA the full system, for the model calculations. In both cases, the predicted SOA mass concentrations are relatively low (< 2 or < 4 µg m^{-3}), yet in a realistic range for relatively clean air quality conditions. These absolute SOA mass concentrations may contribute to the observed fluctuations in MAPE and MPE for different resolutions and polarity axis choices, since minor absolute differences can result in larger relative deviations.

Table 7 compares predicted κ values for toluene and α-pinene systems at two water activity levels often used in the estimation of diameter growth factors and hygroscopicity parameters from field observations. The α-pinene SOA exhibits higher κ values, indicating greater hygroscopicity than toluene SOA, yet both SOA types are relatively low in hygroscopicity with κ values of less than 0.1, a value often assumed as representative of the organic aerosol fraction in aged tropospheric particles (e.g., Rastak et al., 2017). As an example, Fig. S2 shows the speciated SOA mass concentrations predicted for the toluene SOA system at different water activities when using surrogate components derived from the mass-weighted medoid method for a 10×5 grid resolution. The water activity (or equilibrium RH) has only a weak influence on the predicted total SOA mass concentration since the hygroscopicity of the SOA is relatively low, leading to a weak feedback on the partitioning of semivolatile organics due to water uptake.

3.3.1 Analysis of the polarity axes and surrogate selection methods

We compare three choices for the polarity axis of our 2D lumping framework: ACR, O:C ratio, and $\overline{\mathrm{OS}}_{\mathrm{C}}$. These polarity metrics show a surprisingly large degree of variation in representing the polarity-related molecular properties, as demonstrated in Fig. 12 for the components from the toluene SOA system. While O:C ratio and $\overline{\mathrm{OS}}_{\mathrm{C}}$ serve as well-established polarity proxies, the ACR captures additional functional-group-level information, resulting in a wide spread of ACR values for a given

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

Table 2. Comparison of relative deviations in predicted SOA mass concentrations at 298 K for the α-pinene-derived SOA system. The MAPE and MPE values are listed for the four different surrogate selection methods combined with three choices for the polarity axis (y-axis) and for several grid/cluster resolutions. The reference values are from the partitioning computation based on the full system (174 organic components). For a comparison of related absolute quantities, see Table S3.

Surrogate selection	y-axis	4×2		6×3		8×4		10×5	
		MAPE	MPE	MAPE	MPE	MAPE	MPE	MAPE	MPE
Midpoint	$\log_{10}\left[\gamma_{j,\mathrm{hex}}^{(x)}/\gamma_{j,\mathrm{w}}^{(x)}\right]$	13.6%	12.6%	25.9%	-25.9%	2.1%	-2.6%	20.1%	20.1%
Medoid		4.1%	3.7%	24.8%	-24.8%	24.2%	-24.2%	17.6%	17.6%
Weighted medoid		20.4%	-20.4%	2.2%	2.2%	3.2%	3.2%	5.0%	5.0%
Weighted k-means		2.0%	2.0%	1.9%	0.2%	0.4%	-0.4%	0.1%	-0.1%
Midpoint	O:C	51.2%	51.2%	6.6%	2.5%	17.4%	17.4%	24.3%	24.3%
Medoid		16.0%	-8.8%	8.0%	8.0%	6.9%	4.4%	39.8%	39.8%
Weighted medoid		6.1%	0.9%	18.0%	18.0%	7.1%	-0.8%	6.6%	6.6%
Weighted k-means		6.7%	3.9%	2.9%	2.9%	0.9%	-0.9%	0.2%	-0.2%
Midpoint	$\overline{\mathrm{OS}}_{\mathrm{C}}$	71.4%	71.4%	9.2%	8.4%	22.9%	22.9%	1.5%	1.1%
Medoid		16.4%	-1.8%	25.6%	-25.6%	25.6%	-25.6%	20.3%	19.0%
Weighted medoid		23.8%	-23.8%	2.2%	-0.1%	1.6%	-0.8%	3.0%	3.0%
Weighted k-means		1.7%	1.0%	1.6%	0.7%	1.2%	1.0%	0.9%	0.6%

O:C ratio or $\overline{\mathrm{OS}}_{\mathrm{C}}$, e.g. compare the spread at $\overline{\mathrm{OS}}_{\mathrm{C}} = 1$ in Fig. 12. This demonstrates that compounds with similar $\overline{\mathrm{OS}}_{\mathrm{C}}$ can have distinctly different relative affinities for water, as expressed by their predicted ACR.

The accuracy of the grid-based and clustering approaches, as measured by MPE and MAPE, varies notably depending on the surrogate selection method and grid resolution chosen; refer to Tables 2–5. This variability underscores the importance of selecting appropriate methods for specific modelling objectives. At high resolutions, e.g. 10×5, the polarity axis choice has a modest impact, while it can be substantial at lower resolutions, especially in terms of predicted κ values when using the O:C polarity axis paired with one of the non-weighted gridded surrogate selection methods, in which case MAPE $> 50\,\%$ can occur.

In the case of the α-pinene system (Tables 2 and 3), the grid-based methods generally exhibit higher MAPE compared to the clustering-based k-means approach, especially at lower resolutions. This trend is consistent across the evaluated metrics and resolutions, highlighting the potential limitations of the simpler grid-based techniques in accurately reducing and representing complex chemical systems. Higher resolutions typically yield lower MAPE, especially in case of the two mass-weighted surrogate selection methods, indicating improved accuracy and preference in applications.

Overall, the activity coefficient ratio as polarity metric choice slightly outperformed the O:C ratio and the $\overline{\mathrm{OS}}_{\mathrm{C}}$ in case of the systems studied, demonstrating its ability in representing polarity in the context of subsequent gas–particle partitioning

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

Table 3. Similar to Table 2 but for the MAPE and MPE of the predicted hygroscopicity parameter κ (evaluated at water activities of 85 % and 90 %) relative to the full α-pinene SOA system prediction used as benchmark.

Surrogate selection	y-axis	4×2		6×3		8×4		10×5	
		MAPE	MPE	MAPE	MPE	MAPE	MPE	MAPE	MPE
Midpoint	$\log_{10}\left[\gamma_{j,\mathrm{hex}}^{(x)}/\gamma_{j,\mathrm{w}}^{(x)}\right]$	3.3%	-3.3%	16.0%	16.0%	3.5%	3.5%	9.3%	-9.3%
Medoid		16.1%	16.1%	11.5%	11.5%	8.2%	8.2%	6.5%	-6.5%
Weighted medoid		3.6%	3.6%	3.0%	-3.0%	3.0%	-3.0%	1.8%	-1.8%
Weighted k-means		3.5%	-3.5%	3.6%	-3.6%	2.1%	-2.1%	0.0%	-0.0%
Midpoint	O:C	73.3%	73.3%	13.9%	13.9%	19.2%	19.2%	1.5%	1.5%
Medoid		89.1%	89.1%	39.8%	39.8%	36.4%	36.4%	14.6%	14.6%
Weighted medoid		11.7%	11.7%	1.3%	0.5%	14.4%	14.4%	15.9%	15.9%
Weighted k-means		14.9%	14.9%	5.2%	-5.2%	2.1%	-2.1%	0.3%	-0.3%
Midpoint	$\overline{\mathrm{OS}}_{\mathrm{C}}$	35.8%	35.8%	26.2%	26.2%	5.9%	-5.9%	5.9%	-5.9%
Medoid		32.8%	32.8%	1.6%	1.6%	21.4%	21.4%	7.5%	-7.5%
Weighted medoid		12.9%	12.9%	5.0%	-5.0%	5.8%	-5.8%	7.8%	-7.8%
Weighted k-means		8.2%	-8.2%	7.1%	-7.1%	4.5%	-4.5%	2.5%	-2.5%

impacts on aerosol mass and hygroscopicity. In the case of the α-pinene system, the mass-weighted k-means approach using ACR achieved impressively low MAPE values, ranging from 2.0 % (4×2 resolution) to 0.1% (10×5 resolution). The toluene system also showed good performance of the mass-weighted k-means approach with ACR, with MAPE values ranging from 6.0 % (4×2 resolution) to 4.3 % (10×5 resolution).

In a comparison of the different surrogate selection methods, the mass-weighted k-means method consistently scored as the most consistent and accurate across both systems. This is evident in case of the α-pinene SOA system when using the O:C ratio for polarity. The mass-weighted k-means method demonstrated superior performance with low MAPE values (0.2 % to 6.7 %) for predicted SOA mass concentrations across various system resolutions. In contrast, the midpoint method showed significant variability, with the MAPE ranging from 6.6 % to 51.2 %. The medoid and weighted medoid methods also displayed lower consistency with the MAPE reaching up to 39.8 %. In terms of the predicted hygroscopicity parameter of the α-pinene SOA, MAPE values spanned from 0.0 % (high resolution, ACR, k-means method) to 89.1 % (low resolution, O:C ratio, medoid method).

In the case of the toluene SOA system using the O:C ratio, k-means performed again well and consistently accros different resolutions in terms of predicted SOA mass concentrations, with MAPE values ranging from 1.3 % to 4.9 %. The other surrogate selection methods showed higher variability, with MAPE values of up to 13 % in case of the midpoint method at 4×2 grid resolution. Although, we note that a MAPE of 13 % could still be considered a good performance compared to uncertainties in field measurements of mass concentrations and aerosol composition. The predicted κ values for the toluene system showed a

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

Table 4. Comparison of relative deviations in predicted SOA mass concentrations at 298 K for the toluene-derived SOA system. The MAPE and MPE values are listed for the four different surrogate selection methods combined with three choices for the polarity axis and for several grid/cluster resolutions. The computation with the mass-weighted k-means method using the ACR polarity proxy at 25×10 resolution (219 surrogate components) is used as reference for MAPE and MPE calculations. For related absolute quantities, see Table S4.

Surrogate selection	y-axis	4×2		6×3		8×4		10×5		25×10	
		MAPE	MPE	MAPE	MPE	MAPE	MPE	MAPE	MPE	MAPE	MPE
Midpoint	$\log_{10}\left[\gamma_{j,\mathrm{hex}}^{(x)}/\gamma_{j,\mathrm{w}}^{(x)}\right]$	5.1%	5.1%	2.7%	-2.7%	6.8%	6.8%	1.3%	1.3%	3.4%	3.4%
Medoid		8.2%	8.2%	5.0%	-5.0%	1.9%	1.5%	1.2%	0.6%	2.9%	2.9%
Weighted medoid		12%	12%	6.4%	6.4%	2.0%	2.0%	2.4%	2.4%	3.6%	3.6%
Weighted k-means		6.0%	6.0%	2.4%	2.4%	3.9%	3.9%	4.3%	4.3%	0.0%	0.0%
Midpoint	O:C	13%	13%	10%	-10%	1.4%	1.4%	2.1%	-2.1%	2.5%	-2.5%
Medoid		6.7%	6.7%	1.2%	-0.3%	8.3%	8.3%	4.7%	-4.7%	2.2%	-2.2%
Weighted medoid		8.9%	8.9%	2.0%	2.0%	4.5%	4.5%	4.4%	4.4%	0.7%	-0.7%
Weighted k-means		2.4%	2.4%	1.3%	0.3%	4.9%	-4.9%	4.1%	-4.1%	1.8%	1.8%
Midpoint	$\overline{\mathrm{OS}}_\mathrm{C}$	10%	10%	6.7%	-6.7%	7.8%	7.8%	0.8%	-0.5%	5.1%	5.1%
Medoid		1.8%	-1.8%	4.2%	-4.2%	7.6%	7.6%	0.9%	0.4%	5.1%	5.1%
Weighted medoid		11%	11%	2.0%	-0.9%	1.0%	-0.9%	1.8%	-1.8%	6.3%	6.3%
Weighted k-means		4.7%	3.4%	2.1%	2.1%	0.8%	-0.1%	2.9%	2.9%	2.1%	2.1%

broader range of MAPE values, in the case of the O:C ratio axis ranging from 1.0 % (weighted medoid, highest resolution) to 143 % (midpoint method, lowest resolution) and substantial MPE variability (-143 % to +131 %), indicating significant over- and under-predictions by different methods at the lowest grid resolution studied (4×2).

When using the $\overline{\mathrm{OS}}_\mathrm{C}$ metric, the mass-weighted k-means method again outperformed alternative approaches at most resolutions. In case of the α-pinene SOA system, k-means resulted in MAPE values ranging from 0.9 % to 1.7 % for predicted SOA mass concentrations across various grid resolutions, approximately as accurate as with the ACR polarity axis. The other surrogate selection methods exhibited higher variability, with the midpoint method showing MAPE values as high as 71.4 % for the 4×2 grid resolution. For the predicted κ values, MAPE ranges from 2.5 % (k-means) to 36 % (midpoint method).

In the case of the toluene SOA system, using $\overline{\mathrm{OS}}_\mathrm{C}$, the k-means method continued to perform well and consistently, with a MAPE between 0.8 % and 4.7% for predicted SOA mass concentrations. The other methods showed relatively good performance as well, with a maximum MAPE of 11 % (weighted medoid method, 4×2 grid resolution). In terms of the predicted SOA κ, MAPE values ranged from 8.5 % to 90 %, and MPE values spanned from -49 % (k-means, 6×3 resolution) to 90 % (medoid, 8×4 resolution), again showing higher variability in case of this metric, even when using the mass-weighted k-means method (MAPE from 10 % to 70 %). In comparison, when using ACR as polarity axis, the k-means method achieves a worst MAPE for κ of 31 % for the 4×2 resolution – approximately equivalent to the MAPE of 30 % for an 8×4 resolution when

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

Table 5. Similar to Table 4 but for the MAPE and MPE of the predicted hygroscopicity parameter κ of the predicted toluene-derived SOA (evaluated at water activities of 85 % and 90 %). The mass-concentration-weighted k-means method at 25×10 resolution is used as reference case.

Surrogate selection	y-axis	4×2		6×3		8×4		10×5		25×10	
		MAPE	MPE	MAPE	MPE	MAPE	MPE	MAPE	MPE	MAPE	MPE
Midpoint	$\log_{10}\left[\gamma_{j,\mathrm{hex}}^{(x)}/\gamma_{j,\mathrm{w}}^{(x)}\right]$	1.4%	0.9%	26%	-26%	27%	27%	21%	-21%	2.9%	-2.9%
Medoid		9.4%	-9.4%	47%	47%	27%	27%	6.3%	-6.3%	5.0%	5.0%
Weighted medoid		10%	10%	53%	53%	6.5%	6.5%	6.3%	-6.3%	9.8%	-9.8%
Weighted k-means		31%	-31%	8.1%	-8.1%	9.0%	-9.0%	13%	-13%	0.0%	0.0%
Midpoint	O:C	143%	-143%	38%	38%	25%	-25%	7.5%	7.5%	45%	-45%
Medoid		131%	131%	38%	38%	35%	-35%	40%	40%	18%	-18%
Weighted medoid		46%	46%	8.5%	8.5%	51%	51%	42%	-42%	1.0%	1.0%
Weighted k-means		18%	-18%	45%	45%	26%	-26%	11%	-11%	10%	-10%
Midpoint	$\overline{\mathrm{OS}}_{\mathrm{C}}$	34%	-34%	49%	49%	30%	-30%	25%	25%	12%	12%
Medoid		24%	-24%	23%	-23%	90%	90%	42%	-42%	30%	30%
Weighted medoid		70%	70%	8.5%	-8.5%	44%	-44%	42%	-42%	14%	14%
Weighted k-means		70%	70%	49%	-49%	30%	-30%	25%	25%	10%	10%

Table 6. Comparison of predicted SOA mass concentrations at selected water activities (equilibrium RH) for the toluene and α-pinene SOA systems. For the toluene SOA system, the data are based on ACR as the polarity axis, 25×10 resolution and k-means surrogate selection. The full set of MCM-derived system components is used for the α-pinene SOA case.

Water activity (–)	Toluene SOA ($\mu g\,m^{-3}$)	α-Pinene SOA ($\mu g\,m^{-3}$)
0.95	1.981	3.505
0.85	1.786	3.232
0.75	1.640	2.997
0.65	1.523	2.791
0.50	1.382	2.524
0.40	1.305	2.368
0.30	1.237	2.225

using $\overline{\mathrm{OS}}_{\mathrm{C}}$ for polarity. This is also consistent with ACR leading to better k-means performance compared to $\overline{\mathrm{OS}}_{\mathrm{C}}$ in case of the α-pinene SOA system.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

Table 7. Comparison of predicted SOA hygroscopicity parameters for the toluene SOA system (κ_{Tol}) and the α-pinene SOA system ($\kappa_{\alpha\mathrm{P}}$) at two water activity levels. Surrogate selection and resolutions are as for Table 6.

Water activity	κ_{Tol}	$\kappa_{\alpha\mathrm{P}}$
0.90	0.038	0.061
0.85	0.049	0.072

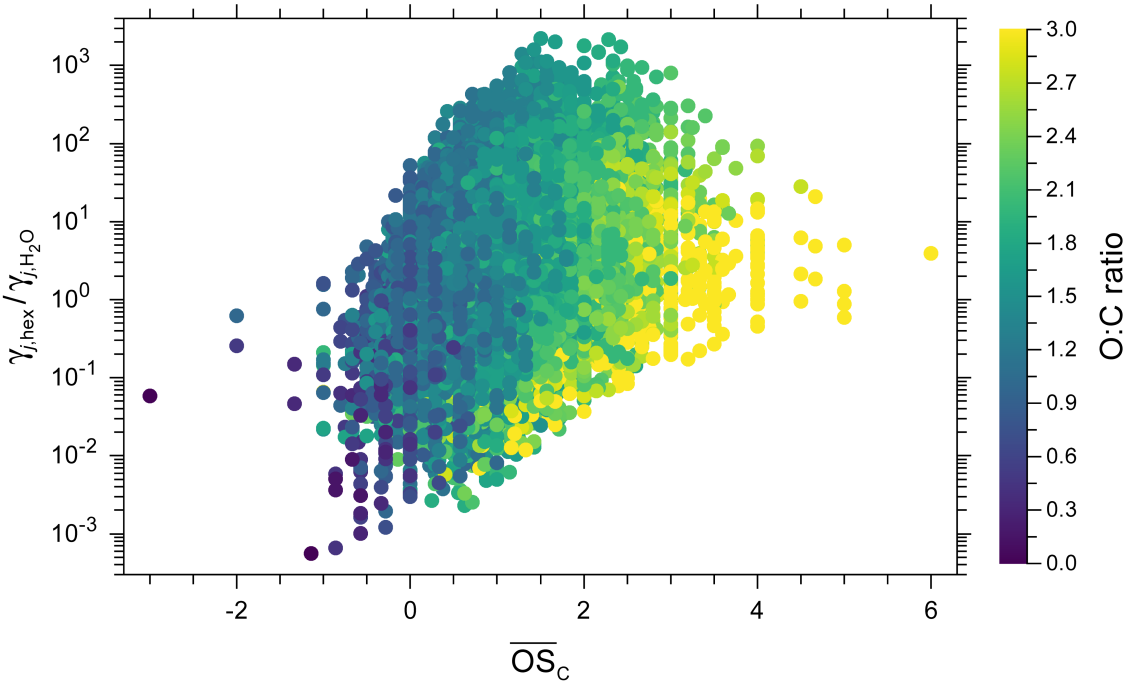


Figure 12. Comparison of three polarity axis metrics: $\overline{\mathrm{OS}}_{\mathrm{C}}$, ACR and O:C ratio for all components of the toluene SOA system. Both $\overline{\mathrm{OS}}_{\mathrm{C}}$ and ACR account for the impact of nitrogen-containing functionalities in a molecule on the respective metric, while the O:C ratio does not.

4 Conclusions

This study introduces a novel chain of computational tools to advance the prediction of organic aerosol formation, which can be applied to chemical product evolution predictions from near-explicit gas phase mechanisms. This will allow the community to approach atmospheric chemistry modelling of organics and related aerosol composition and mass concentrations at a highly detailed level. A level that is only accessible with the aid of automatic product classification methods. By integrating structure- and system-level tools, our work paves the way for gas–particle partitioning modelling involving large numbers of aerosol components, a challenge that has long hindered highly detailed system studies on SOA formation and composition.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

The developed suite of automated tools, including the SMILES to AIOMFAC Subgroups tool (S2AS), enhances processing capabilities for complex chemical systems, as demonstrated for compounds resulting from α-pinene ozonolysis or toluene oxidation. The introduction of a scalable 2D framework with grid-based and cluster-based surrogate selection methods, represents a major advancement in simplifying complex chemical systems while maintaining controllable accuracy in SOA partitioning predictions and retaining structural information (of the surrogates).

The evaluation of the framework's applications to two SOA systems provided information about the effects of polarity axis choices and system resolution. We examined the impact of choosing relatively low to moderate grid resolutions (4×2, 6×3, 8×4, 10×5) in surrogate representations of gas–particle systems. As expected, higher grid resolutions (or higher number of k-means clusters) generally yield better accuracy at the cost of increased computational cost for the gas–particle partitioning computation step. By reducing the number of species from hundreds or thousands to just $\sim$ 32 surrogate components, substantial computational efficiency gains were achieved without compromising accuracy in SOA mass predictions substantially. Based on the two example systems discussed, we recommend grid resolutions of at least 8×4 to maintain expected prediction errors within a 10 % threshold. More generally, for reference offline gas–particle partitioning computations, in which computational costs are secondary concern, we recommend choosing the highest grid or cluster resolution feasible for the use case, up to about 200 surrogate components. Beyond 200 surrogate components, the cost versus accuracy trade-off will likely result in diminishing returns in terms of improvements in predicted (aerosol) system properties.

Our quantitative analysis highlights the effectiveness of the mass-weighted k-means method. When it is applied with the ACR polarity representation, good performance is achieved even at relatively coarse resolutions of an organic aerosol system. The choice of polarity axis metric impacted the accuracy and consistency of SOA system representation. Compared to the O:C ratio and $\overline{\mathrm{OS}}_\mathrm{C}$ proxies for polarity, the ACR metric is particularly effective since it encodes more information about a components structure and mixing properties in aqueous SOA. The k-means based medoid approach paired with the ACR metric emerged as the most effective 2D lumping and surrogate selection method among those evaluated. While it may be possible to improve the grid-based surrogate selection methods, e.g. by introducing variable-resolution grids, with higher resolution in the semivolatile range, the strong performance of the k-means method renders such improvements unnecessary.

The new computational framework will aid in bridging the wide gap between detailed, molecular-level reaction and simulation mechanisms and the computationally constrained, much simpler aerosol schemes of regional and large-scale atmospheric models. It provides automated tools and lumping techniques for generating reduced-complexity representations of aerosol properties and gas–particle partitioning inputs in a systematic and objective manner. Furthermore, the scalability of the 2D framework allows researchers to adjust the level of detail based on specific research needs while being mindful of computational constraints. Alongside with detailed chemical mechanisms, such as GECKO-A, future work could make use of the introduced tools in simulations of specific environmental chamber studies on aerosol formation from known precursors – and to represent related gas–particle systems by a tuneable selection of structure-resolved surrogates.

Moreover, this work opens new avenues for interdisciplinary research, combining atmospheric chemistry with data science and environmental engineering. As the scientific community adopts and further develops such frameworks, it has the potential

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

to drive progress in our understanding of atmospheric processes and our ability to study complex multiphase gas–particle systems via simulation.

Code and data availability. The Python code of the S2AS model (v1.0) and related documentation are available via an online code repository (https://github.com/andizuend/S2AS__SMILES_to_AIOMFAC). The Fortran code of the 2D polarity–volatility framework (v1.0) as well as an associated plotting program and documentation are available via an online repository (https://github.com/andizuend/2D_Polarity_Volatility_lumping). The UManSysProp code (v1.0) by Topping et al. (2016a) is available via an online code repository (https://github.com/loftytopping/UManSysProp_public; last access: 18 September 2025). The specific version of the used UManSysProp code including the adaptations for temperature-dependent pure-component vapour pressure parameterizations used in this work, has been archived as a Zenodo repository at https://doi.org/10.5281/zenodo.17172675 (Zuend et al., 2025). The Master Chemical Mechanism (v3.3.1) and the related AtChem online box model are available online via https://mcm.york.ac.uk/MCM/ (last access: 18 September 2025). Predicted SOA mass concentrations and hygroscopicity parameters for various surrogate methods and polarity metrics are summarized in the electronic Supplement. The data underlying the shown figures and tables, as well as related output from the property prediction tools and the 2D lumping framework, have been archived as a Zenodo repository at https://doi.org/10.5281/zenodo.17088391 (Amaladhasan et al., 2025).

Author contributions. DAA and AZ conceptualized the project. DAA and AZ developed the code of the S2AS property prediction tool and the lumping framework. DAA carried out the MCM simulations. DHB carried out the GECKO-A simulations. DHB modified the vapour pressure estimation tool. DAA carried out the lumping framework data analysis with input by AZ. DAA and AZ co-wrote the manuscript with contributions by DHB.

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (grants no. RGPIN/04315-2014 and RGPIN-2021-02688). We extend our thanks to Dr. Bernard Aumont for providing output from a GECKO-A simulation, supporting initial tests of our software tools.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

References

Allen, F., Pon, A., Greiner, R., and Wishart, D.: Computational prediction of electron ionization mass spectra to assist in GC/MS compound identification, Analytical chemistry, 88, 7689–7697, 2016.

Amaladhasan, D. A., Zuend, A., and Hassan-Barthaux, D.: Alpha-pinene and Toluene SOA System data used in Amaladhasan et al for 2D lumping, https://doi.org/10.5281/zenodo.17088391, data set, 2025.

Armeli, G., Peters, J.-H., and Koop, T.: Machine-Learning-Based Prediction of the Glass Transition Temperature of Organic Compounds Using Experimental Data, ACS Omega, 8, 12 298–12 309, https://doi.org/10.1021/acsomega.2c08146, doi: 10.1021/acsomega.2c08146, 2023.

Aumont, B., Szopa, S., and Madronich, S.: Modelling the evolution of organic carbon during its gas-phase tropospheric oxidation: development of an explicit model based on a self generating approach, Atmos. Chem. Phys., 5, 2497–2517, http://www.atmos-chem-phys.net/5/2497/2005/, 2005.

Barley, M. H. and McFiggans, G.: The critical assessment of vapour pressure estimation methods for use in modelling the formation of atmospheric organic aerosol, Atmos. Chem. Phys., 10, 749–767, https://doi.org/10.5194/acp-10-749-2010, aCP, 2010.

Bertram, A. K., Martin, S. T., Hanna, S. J., Smith, M. L., Bodsworth, A., Chen, Q., Kuwata, M., Liu, A., You, Y., and Zorn, S. R.: Predicting the relative humidities of liquid-liquid phase separation, efflorescence, and deliquescence of mixed particles of ammonium sulfate, organic material, and water using the organic-to-sulfate mass ratio of the particle and the oxygen-to-carbon elemental ratio of the organic component, Atmos. Chem. Phys., 11, 10 995–11 006, https://doi.org/10.5194/acp-11-10995-2011, 2011.

Bilde, M., Barsanti, K., Booth, M., Cappa, C. D., Donahue, N. M., Emanuelsson, E. U., McFiggans, G., Krieger, U. K., Marcolli, C., Topping, D., Ziemann, P., Barley, M., Clegg, S., Dennis-Smither, B., Hallquist, M., Hallquist, A. M., Khlystov, A., Kulmala, M., Mogensen, D., Percival, C. J., Pope, F., Reid, J. P., Ribeiro da Silva, M. A. V., Rosenoern, T., Salo, K., Soonsin, V. P., Yli-Juuti, T., Prisle, N. L., Pagels, J., Rarey, J., Zardini, A. A., and Riipinen, I.: Saturation Vapor Pressures and Transition Enthalpies of Low-Volatility Organic Molecules of Atmospheric Relevance: From Dicarboxylic Acids to Complex Mixtures, Chemical Reviews, 115, 4115–4156, https://doi.org/10.1021/cr5005502, 2015.

Burkardt, J.: ASA058 the K-Means Problem, https://people.math.sc.edu/Burkardt/f_src/asa058/asa058.html, (last accessed: August 8, 2025), 2008.

Byun, D. W., Young, J., and Odman, M. T.: Governing Equations and Computational Structure of the Community Multiscale Air Quality (CMAQ) Chemical Transport Model, Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System, pp. 6–1–6–41, https://www.cmascenter.org/cmaq/science_documentation/pdf/ch06.pdf, ePA/600/R-99/030, Chapter 6, 1999.

Chang, E. I. and Pankow, J. F.: Prediction of activity coefficients in liquid aerosol particles containing organic compounds, dissolved inorganic salts, and water - Part 2: Consideration of phase separation effects by an X-UNIFAC model, Atmos. Environ., 40, 6422–6436, https://doi.org/10.1016/j.atmosenv.2006.04.031, 2006.

Chang, E. I. and Pankow, J. F.: Organic particulate matter formation at varying relative humidity using surrogate secondary and primary organic compounds with activity corrections in the condensed phase obtained using a method based on the Wilson equation, Atmospheric Chemistry and Physics, 10, 5475–5490, https://doi.org/10.5194/acp-10-5475-2010, 2010.

Compernolle, S., Ceulemans, K., and Müller, J.-F.: EVAPORATION: a new vapour pressure estimation methodfor organic molecules including non-additivity and intramolecular interactions, Atmos. Chem. Phys., 11, 9431–9450, https://doi.org/10.5194/acp-11-9431-2011, 2011.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

DeRieux, W. S. W., Li, Y., Lin, P., Laskin, J., Laskin, A., Bertram, A. K., Nizkorodov, S. A., and Shiraiwa, M.: Predicting the glass transition temperature and viscosity of secondary organic material using molecular composition, Atmos. Chem. Phys., 18, 6331–6351, https://doi.org/10.5194/acp-18-6331-2018, aCP, 2018.

Donahue, N. M., Robinson, A. L., Stanier, C. O., and Pandis, S. N.: Coupled Partitioning, Dilution, and Chemical Aging of Semivolatile Organics, Environ. Sci. Technol., 40, 02 635–2643, https://doi.org/10.1021/es052297c, 2006.

Donahue, N. M., Epstein, S. A., Pandis, S. N., and Robinson, A. L.: A two-dimensional volatility basis set: 1. organic-aerosol mixing thermodynamics, Atmos. Chem. Phys., 11, 3303–3318, https://doi.org/10.5194/acp-11-3303-2011, 2011.

Donahue, N. M., Kroll, J. H., Pandis, S. N., and Robinson, A. L.: A two-dimensional volatility basis set – Part 2: Diagnostics of organic-aerosol evolution, Atmos. Chem. Phys., 12, 615–634, https://doi.org/10.5194/acp-12-615-2012, 2012.

Ehrlich, H.-C. and Rarey, M.: Systematic benchmark of substructure search in molecular graphs-From Ullmann to VF2, Journal of cheminformatics, 4, 1–17, 2012.

Erdakos, G. B. and Pankow, J. F.: Gas/particle partitioning of neutral and ionizing compounds to single- and multi-phase aerosol particles. 2. Phase separation in liquid particulate matter containing both polar and low-polarity organic compounds, Atmos. Environ., 38, 1005–1013, https://doi.org/10.1016/j.atmosenv.2003.10.038, 2004.

Fredenslund, A., Jones, R. L., and Prausnitz, J. M.: Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures, AIChE J., 21, 1086–1099, 1975.

Galeazzo, T. and Shiraiwa, M.: Predicting glass transition temperature and melting point of organic compounds via machine learning and molecular embeddings, Environmental Science: Atmospheres, 2, 362–374, https://doi.org/10.1039/D1EA00090J, 2022.

Girolami, G. S.: A Simple "Back of the Envelope" Method for Estimating the Densities and Molecular Volumes of Liquids and Solids, Journal of Chemical Education, 71, 962, https://doi.org/10.1021/ed071p962, 1994.

Griffin, R. J., Nguyen, K., Dabdub, D., and Seinfeld, J. H.: A Coupled Hydrophobic-Hydrophilic Model for Predicting Secondary Organic Aerosol Formation, J. Atmos. Chem., 44, 171–190, 2003.

Hallquist, M., Wenger, J. C., Baltensperger, U., Rudich, Y., Simpson, D., Claeys, M., Dommen, J., Donahue, N. M., George, C., Goldstein, A. H., Hamilton, J. F., Herrmann, H., Hoffmann, T., Iinuma, Y., Jang, M., Jenkin, M. E., Jimenez, J. L., Kiendler-Scharr, A., Maenhaut, W., McFiggans, G., Mentel, T. F., Monod, A., Prevot, A. S. H., Seinfeld, J. H., Surratt, J. D., Szmigielski, R., and Wildt, J.: The formation, properties and impact of secondary organic aerosol: current and emerging issues, Atmos. Chem. Phys., 9, 5155–5236, http://www.atmos-chem-phys.net/9/5155/2009/, 2009.

Hansen, H. K., Rasmussen, P., Fredenslund, A., Schiller, M., and Gmehling, J.: Vapor–liquid equilibria by UNIFAC group contribution. 5. Revision and extension, Industrial & Engineering Chemistry Research, 30, 2352–2355, https://doi.org/10.1021/ie00058a017, 1991.

Hartigan, J. A. and Wong, M. A.: Algorithm AS 136: A K-Means Clustering Algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics), 28, 100–108, https://doi.org/10.2307/2346830, 1979.

Huang, Y., Mahrt, F., Xu, S., Shiraiwa, M., Zuend, A., and Bertram, A. K.: Coexistence of three liquid phases in individual atmospheric aerosol particles, Proceedings of the National Academy of Sciences, 118, e2102512 118, https://doi.org/10.1073/pnas.2102512118, 2021.

Jenkin, M. E., Saunders, S. M., and Pilling, M. J.: The tropospheric degradation of volatile organic compounds: A protocol for mechanism development, Atmos. Environ., 31, 81–104, https://doi.org/10.1016/S1352-2310(96)00105-7, 1997.

Jimenez, J. L., Canagaratna, M. R., Donahue, N. M., Prevot, A. S. H., Zhang, Q., Kroll, J. H., DeCarlo, P. F., Allan, J. D., Coe, H., Ng, N. L., Aiken, A. C., Docherty, K. S., Ulbrich, I. M., Grieshop, A. P., Robinson, A. L., Duplissy, J., Smith, J. D., Wilson, K. R., Lanz, V. A., Hueglin, C., Sun, Y. L., Tian, J., Laaksonen, A., Raatikainen, T., Rautiainen, J., Vaattovaara, P., Ehn, M., Kulmala, M., Tomlinson,

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

J. M., Collins, D. R., Cubison, M. J., Dunlea, E. J., Huffman, J. A., Onasch, T. B., Alfarra, M. R., Williams, P. I., Bower, K., Kondo, Y., Schneider, J., Drewnick, F., Borrmann, S., Weimer, S., Demerjian, K., Salcedo, D., Cottrell, L., Griffin, R., Takami, A., Miyoshi, T., Hatakeyama, S., Shimono, A., Sun, J. Y., Zhang, Y. M., Dzepina, K., Kimmel, J. R., Sueper, D., Jayne, J. T., Herndon, S. C., Trimborn, A. M., Williams, L. R., Wood, E. C., Middlebrook, A. M., Kolb, C. E., Baltensperger, U., and Worsnop, D. R.: Evolution of Organic Aerosols in the Atmosphere, Science, 326, 1525–1529, https://doi.org/10.1126/science.1180353, 2009.

Kamlet, M. J., Doherty, R. M., Abraham, M. H., Marcus, Y., and Taft, R. W.: Linear solvation energy relationship. 46. An improved equation for correlation and prediction of octanol/water partition coefficients of organic nonelectrolytes (including strong hydrogen bond donor solutes), J. Phys. Chem., 92, 5244–5255, https://doi.org/10.1021/j100329a035, 1988.

Kroll, J. H., Donahue, N. M., Jimenez, J. L., Kessler, S. H., Canagaratna, M. R., Wilson, K. R., Altieri, K. E., Mazzoleni, L. R., Wozniak, A. S., Bluhm, H., Mysak, E. R., Smith, J. D., Kolb, C. E., and Worsnop, D. R.: Carbon oxidation state as a metric for describing the chemistry of atmospheric organic aerosol, Nature Chemistry, 3, 133–139, https://doi.org/10.1038/nchem.948, 2011.

Landrum, G.: RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling, 2013.

Lannuque, V., D'Anna, B., Couvidat, F., Valorso, R., and Sartelet, K.: Improvement in Modeling of OH and HO2 Radical Concentrations during Toluene and Xylene Oxidation with RACM2 Using MCM/GECKO-A, Atmosphere, 12, https://doi.org/10.3390/atmos12060732, 2021.

Marcolli, C. and Peter, T.: Water activity in polyol/water systems: new UNIFAC parameterization, Atmos. Chem. Phys., 5, 1545–1555, 2005.

Mouchel-Vallon, C., Lee-Taylor, J., Hodzic, A., Artaxo, P., Aumont, B., Camredon, M., Gurarie, D., Jimenez, J.-L., Lenschow, D. H., Martin, S. T., et al.: Exploration of oxidative chemistry and secondary organic aerosol formation in the Amazon during the wet season: explicit modeling of the Manaus urban plume with GECKO-A, Atmospheric Chemistry and Physics, 20, 5995–6014, 2020.

Nannoolal, Y., Rarey, J., and Ramjugernath, D.: Estimation of pure component properties: Part 3. Estimation of the vapor pressure of non-electrolyte organic compounds via group contributions and group interactions, Fluid Phase Equilibria, 269, 117–133, 2008.

O'Boyle Jr, E. H., Humphrey, R. H., Pollack, J. M., Hawver, T. H., and Story, P. A.: The relation between emotional intelligence and job performance: A meta-analysis, Journal of Organizational Behavior, 32, 788–818, 2011.

OEChem, T.: OpenEye Scientific Software, Inc., Santa Fe, NM, USA, 2012.

O'Meara, S., Booth, A. M., Barley, M. H., Topping, D., and McFiggans, G.: An assessment of vapour pressure estimation methods, Physical Chemistry Chemical Physics, 16, 19 453–19 469, https://doi.org/10.1039/C4CP00857J, 2014.

Pankow, J. F.: Gas/particle partitioning of neutral and ionizing compounds to single and multi-phase aerosol particles. 1.Unified modeling framework, Atmos. Environ., 37, 3323–3333, https://doi.org/10.1016/S1352-2310(03)00346-7, 2003.

Pankow, J. F. and Asher, W. E.: SIMPOL.1: a simple group contribution method for predicting vapor pressures and enthalpies of vaporization of multifunctional organic compounds, Atmos. Chem. Phys., 8, http://dx.doi.org/10.5194/acp-8-2773-2008, 2008a.

Pankow, J. F. and Asher, W. E.: SIMPOL. 1: a simple group contribution method for predicting vapor pressures and enthalpies of vaporization of multifunctional organic compounds, Atmospheric Chemistry and Physics, 8, 2773–2796, 2008b.

Pankow, J. F. and Barsanti, C. K.: The carbon number-polarity grid: A means to manage the complexity of the mix of organic compounds when modeling atmospheric organic particulate matter, Atmospheric Environment, 43, 2829 – 2835, https://doi.org/10.1016/j.atmosenv.2008.12.050, 2009.

Pankow, J. F. and Chang, E. I.: Variation in the sensitivity of predicted levels of atmospheric organic particulate matter (OPM), Environmental science & technology, 42, 7321–7329, 2008.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

Pavlov, D., Rybalkin, M., Karulin, B., Kozhevnikov, M., Savelyev, A., and Churinov, A.: Indigo: universal cheminformatics API, Journal of cheminformatics, 3, P4, 2011.

Pun, B. K. L., Griffin, R. J., Seigneur, C., and Seinfeld, J. H.: Secondary organic aerosol - 2. Thermodynamic model for gas/particle partitioning of molecular constituents, J. Geophys. Res. Atmos., 107, https://doi.org/10.1029/2001JD000542, 2002.

Rastak, N., Pajunoja, A., Acosta Navarro, J. C., Ma, J., Song, M., Partridge, D. G., Kirkevåg, A., Leong, Y., Hu, W. W., Taylor, N. F., Lambe, A., Cerully, K., Bougiatioti, A., Liu, P., Krejci, R., Petäjä, T., Percival, C., Davidovits, P., Worsnop, D. R., Ekman, A. M. L., Nenes, A., Martin, S., Jimenez, J. L., Collins, D. R., Topping, D. O., Bertram, A. K., Zuend, A., Virtanen, A., and Riipinen, I.: Microphysical explanation of the RH-dependent water affinity of biogenic organic aerosol and its importance for climate, Geophys. Res. Lett., 44, 5167–5177, https://doi.org/10.1002/2017GL073056, 2017.

Ruggeri, G., Bernhard, F. A., Henderson, B. H., and Takahama, S.: Model–measurement comparison of functional group abundance in α-pinene and 1, 3, 5-trimethylbenzene secondary organic aerosol formation, Atmospheric Chemistry and Physics, 16, 8729–8747, 2016.

Saunders, S. M., Jenkin, M. E., Derwent, R. G., and Pilling, M. J.: Protocol for the development of the Master Chemical Mechanism, MCM v3 (Part A): tropospheric degradation of non-aromatic volatile organic compounds, Atmos. Chem. Phys., 3, 161–180, https://doi.org/10.5194/acp-3-161-2003, 2003.

Schervish, M. and Shiraiwa, M.: Impact of phase state and non-ideal mixing on equilibration timescales of secondary organic aerosol partitioning, Atmospheric Chemistry and Physics, 23, 221–233, https://doi.org/10.5194/acp-23-221-2023, 2023.

Schmedding, R. and Zuend, A.: The role of interfacial tension in the size-dependent phase separation of atmospheric aerosol particles, Atmos. Chem. Phys., 25, 327–346, https://doi.org/10.5194/acp-25-327-2025, aCP, 2025.

Schmedding, R., Franssen, M., and Zuend, A.: A Machine Learning Approach for Predicting the Pure-Component Surface Tension of Atmospherically Relevant Organic Compounds, ACS ES&T Air, https://doi.org/10.1021/acsestair.4c00291, doi: 10.1021/acsestair.4c00291, 2025.

Semeniuk, K. and Dastoor, A.: Current State of Atmospheric Aerosol Thermodynamics and Mass Transfer Modeling: A Review, Atmosphere, 11, https://doi.org/10.3390/atmos11020156, 2020.

Smith, M. L., Kuwata, M., and Martin, S. T.: Secondary Organic Material Produced by the Dark Ozonolysis of α-Pinene Minimally Affects the Deliquescence and Efflorescence of Ammonium Sulfate, Aerosol Sci. Technol., 45, 244–261, https://doi.org/10.1080/02786826.2010.532178, 2011.

Sommers, J. M., Stroud, C. A., Adam, M. G., O'Brien, J., Brook, J. R., Hayden, K., Lee, A. K. Y., Li, K., Liggio, J., Mihele, C., Mittermeier, R. L., Stevens, R. G., Wolde, M., Zuend, A., and Hayes, P. L.: Evaluating SOA formation from different sources of semi- and intermediate-volatility organic compounds from the Athabasca oil sands, Environmental Science: Atmospheres, 2, 469–490, https://doi.org/10.1039/D1EA00053E, 2022.

Sparks, D. N.: Algorithm AS 58: Euclidean Cluster Analysis, Journal of the Royal Statistical Society. Series C (Applied Statistics), 22, 126–130, https://doi.org/10.2307/2346321, 1973.

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., and Willighagen, E.: The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics, Journal of chemical information and computer sciences, 43, 493–500, 2003.

Topping, D., Barley, M., Bane, M., Higham, N. J., Aumont, B., Dingle, N., and McFiggans, G.: UManSysProp v1. 0: an online and open-source facility for molecular property prediction and atmospheric aerosol calculations, Geoscientific Model Development, 9, 899–914, 2016a.

https://doi.org/10.5194/egusphere-2025-4673
Preprint. Discussion started: 30 January 2026
© Author(s) 2026. CC BY 4.0 License.

Topping, D., Barley, M., Bane, M. K., Higham, N., Aumont, B., Dingle, N., and McFiggans, G.: UManSysProp v1. 0: an online and open-source facility for molecular property prediction and atmospheric aerosol calculations, Geoscientific Model Development, 9, 899–914, 2016b.

Topping, D. L. and Bane, M., eds.: Introduction to aerosol modelling: From theory to code, John Wiley & Sons, ISBN 1119625718, 2022.

Topping, D. O., McFiggans, G. B., Kiss, G., Varga, Z., Facchini, M. C., Decesari, S., and Mircea, M.: Surface tensions of multi-component mixed inorganic/organic aqueous systems of atmospheric significance: measurements, model predictions and importance for cloud activation predictions, Atmos. Chem. Phys., 7, 2371–2398, https://doi.org/10.5194/acp-7-2371-2007, aCP http://www.atmos-chem-phys.net/7/2371/2007/acp-7-2371-2007.pdf, 2007.

Toropov, A. A., Rasulev, B. F., Leszczynska, D., and Leszczynski, J.: Multiplicative SMILES-based optimal descriptors: QSPR modeling of fullerene C60 solubility in organic solvents, Chemical Physics Letters, 457, 332–336, 2008.

Tulet, P., Grini, A., Griffin, R. J., and Petitcol, S.: ORILAM-SOA: A computationally efficient model for predicting secondary organic aerosols in three-dimensional atmospheric models, Journal of Geophysical Research, 111, https://api.semanticscholar.org/CorpusID:129103245, 2006.

Wang, Z., Couvidat, F., and Sartelet, K.: GENerator of reduced Organic Aerosol mechanism (GENOA v1.0): an automatic generation tool of semi-explicit mechanisms, Geoscientific Model Development, 15, 8957–8982, https://doi.org/10.5194/gmd-15-8957-2022, 2022.

Weininger, D.: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, Journal of Chemical Information and Computer Sciences, 28, 31–36, https://doi.org/10.1021/ci00057a005, doi: 10.1021/ci00057a005, 1988.

Wienke, G., , and Gmehling, J.: Prediction of octanol–water partition coefficients, Henry coefficients and water solubilities using UNIFAC, Toxicological & Environmental Chemistry, 65, 57–86, https://doi.org/10.1080/02772249809358557, 1998.

Zhang, J., Zuend, A., Top, J., Surdu, M., Ei Haddad, I., Slowik, J. G., Prevot, A. S. H., and Bell, D. M.: Estimation of the Volatility and Apparent Activity Coefficient of Levoglucosan in Wood-Burning Organic Aerosols, Environ. Sci. Technol. Lett., 11, 1214–1219, https://doi.org/10.1021/acs.estlett.4c00608, doi: 10.1021/acs.estlett.4c00608, 2024.

Zuend, A. and Seinfeld, J. H.: Modeling the gas-particle partitioning of secondary organic aerosol: the importance of liquid-liquid phase separation, Atmos. Chem.and Phys., 12, 3857–3882, https://doi.org/10.5194/acp-12-3857-2012, 2012.

Zuend, A. and Seinfeld, J. H.: A practical method for the calculation of liquid-liquid equilibria in multicomponent organic-water-electrolyte systems using physicochemical constraints, Fluid Phase Equilib., 337, 201–213, https://doi.org/10.1016/j.fluid.2012.09.034, 2013.

Zuend, A., Marcolli, C., Luo, B. P., and Peter, T.: A thermodynamic model of mixed organic-inorganic aerosols to predict activity coefficients, Atmos. Chem. Phys., 8, 4559–4593, https://doi.org/10.5194/acp-8-4559-2008, 2008.

Zuend, A., Marcolli, C., Peter, T., and Seinfeld, J. H.: Computation of liquid-liquid equilibria and phase stabilities: implications for RH-dependent gas/particle partitioning of organic-inorganic aerosols, Atmos. Chem. Phys., 10, 7795–7820, https://doi.org/10.5194/acp-10-7795-2010, 2010.

Zuend, A., Marcolli, C., Booth, A. M., Lienhard, D. M., Soonsin, V., Krieger, U. K., Topping, D. O., McFiggans, G., Peter, T., and Seinfeld, J. H.: New and extended parameterization of the thermodynamic model AIOMFAC: calculation of activity coefficients for organic-inorganic mixtures containing carboxyl, hydroxyl, carbonyl, ether, ester, alkenyl, alkyl, and aromatic functional groups, Atmos. Chem. Phys., 11, 9155–9206, https://doi.org/10.5194/acp-11-9155-2011, 2011.

Zuend, A., Hassan-Barthaux, D., and Amaladhasan, D. A.: SMILES_to_sat_vapour_pressure, https://doi.org/10.5281/zenodo.17172675, 2025.