Manuscript prepared for Geosci. Model Dev.
with version 2015/04/24 7.83 Copernicus papers of the LaTeX class copernicus.cls.
Date: 3 February 2016

UManSysProp [*]v1.0, an online [*]and open-source facility for molecular property prediction and atmospheric aerosol calculations

David Topping[1,2], Mark Barley[2], Michael Bane[3], Nicholas Higham[4], Bernard Aumont[5], Nicholas Dingle[6], and Gordon McFiggans[2]

[1]National Centre for Atmospheric Science, UK
[2]Centre for Atmospheric Science, University of Manchester, Manchester, M13 9PL, UK
[3]High End Compute, Manchester, M13 9PL, UK
[4]School of Mathematics, University of Manchester, Manchester, M13 9PL, UK
[5]LISA, UMR CNRS 7583, Universite Paris Est Creteil et Universite Paris Diderot, Creteil, France
[6]Numerical Algorithms Group (NAG), Ltd Peter House Oxford Street Manchester, M1 5AN, UK

Correspondence to: David Topping
(david.topping@manchester.ac.uk)

Abstract. In this paper we describe the development and application of a new web based facility, UManSysProp (http://umansysprop.seaes.manchester.ac.uk), for automating predictions of molecular and atmospheric aerosol properties. Current facilities include: pure component vapour pressures, critical properties and sub-cooled densities of organic molecules; activity coefficient predictions for mixed inorganic-organic liquid systems; hygroscopic growth factors and CCN activation potential of mixed inorganic/organic aerosol particles; absorptive partitioning calculations with/without a treatment of non-ideality. The aim of this new facility is to provide a single point of reference for all properties relevant to atmospheric aerosol that have been checked for applicability to atmospheric compounds where possible. The group contribution approach allows users to upload molecular information in the form of SMILES strings and UManSysProp will automatically extract the relevant information for calculations. Built using open source chemical informatics, and hosted at the University of Manchester, the facilities are provided via a browser and device-friendly web-interface, or can be accessed using the user's own code via a JSON API. [c3]We also provide the source code for all predictive techniques provided on the site, covered by the GNU GPL license to encourage development of a user community. We have released this via a Github repository (DOI 10.5281/zenodo.45143). In this paper we demonstrate its use with specific examples that can be simulated using the web-browser interface.

[*] *Text added.*
[*] *Text added.*
[c3] *Text added.*

1 Introduction

The many thousands of individual aerosol components ensure that explicit manual calculation of properties that influence their environmental impacts is laborious and time-consuming. The emergence of explicit automatic mechanism generation techniques (Aumont et al. (2005), Jenkin et al. (2012)), including up to many millions of individual gas phase products as aerosol precursors, renders manual calculations impossible and automation is necessary. For example, both inorganic and organic material can transfer between the gas and particle phase. Inorganic electrolytes are restricted to a few well-understood compounds. However, organic material can comprise many thousands of compounds with potentially a vast range of properties (Hallquist et al., 2009). Predicting the evolution of aerosol requires calculating the distribution of all components between the gas and aerosol phase according to equilibrium partitioning or disequilibrium mass transfer. Either treatment requires knowledge of all component vapour pressures and other thermodynamic properties. In the moist atmosphere, the most abundant material that can readily interact with aerosol particles is water vapour. The formation of atmospheric liquid water has a profound influence on the aerosol life cycle and climate. Predicting the hygroscopic response of complex inorganic-organic mixtures requires treatment of solution non-ideality, for example.

It can be difficult to establish what factors are responsible for the outcome of a model prediction. This is particularly true when the number of components might be high in, for example, SOA mass partitioning simulations. It then becomes difficult for others in the community to assess the results presented. This might be complicated by the need to include pure component vapour pressures or activity coefficient predictions for a wide range of highly multifunctional compounds. For example, predictions of aerosol hygroscopicity have either been based on simplified Kohler theory at one extreme (Kreidenweis et al., 2005) or thermodynamic equilibrium models at the other (Topping et al., 2005). [c4] [c5]It is not clear to what extent replication of results is ever achieved for a range of aerosol simulations. Whilst this might also be an issue with results from instrumentation, the development of community driven software at least enables modellers to tackle this problem directly. [c6]There are a number of property predictions facilities that are available online. For example, the US EPA host predictive models and tools for assessing chemicals under the Toxic Substances Control Act (TSCA) (http://www.epa.gov/tsca-screening-tools). From this site one can access the simulation program EPi Suite (http://www.epa.gov/tsca-screening-tools/download-epi-suitetm-estimation-program-interface-v411). This provides a number of facilities including estimates of physical / chemical properties (melting point, water solubility, etc.) and environmental fate properties (breakdown in water or air, etc.). The Dortmund Databank (DDB) provide a wide range of database and software products related to fundamental properties of molecules and mixtures. With varying proprietary and free

[c4] ~~It isnt clear~~
[c5] *Text added.*
[c6] *Text added.*

educational access, their program package ARTIST was developed for the estimation of pure component properties from molecular structure. In the UK the National Chemical Database Service (CDS) provides free access to web-based services including ACD/Labs Inc Physchem and NMR predictions (http://cds.rsc.org/). Services specifically tailored to atmospheric studies include the E-AIM community model for calculating gas/solid/liquid partitioning (http://www.aim.env.uea.ac.uk/aim/aim.php) and the AIOMFAC portal for calculating activity coefficients in mixed inorganic/organic liquid systems (http://www.aiomfac.caltech.edu/).

In this paper we describe the development and application of a new web based facility, UManSysProp, to tackle such issues. Current facilities include: pure component vapour pressures, critical properties and sub-cooled densities of organic molecules; activity coefficient predictions for mixed inorganic-organic liquid systems; hygroscopic growth factors and CCN activation potential of mixed inorganic/organic aerosol particles with associated [c7c8]K(kappa)-Köhler values (Kreidenweis et al., 2005); absorptive partitioning calculations with/without a treatment of non-ideality. UManSysProp automatically extracts the relevant information for calculations. Built using open-source chemical informatics, described in section 2, the facilities are provided via a browser and device-friendly web-interface. In section 3, examples of each prediction are given along with reference to our existing publications that use these tools. Providing a wide range of comparisons between predictions and measurements of each property is outside of the scope of this paper given all of the potential subtleties associated with measurement data (e.g. Topping and McFiggans (2012)). Nonetheless, by providing a minimum set of examples for each case, the ability to perform such comparisons and act as the community's point of reference is demonstrated. Relevant inputs to replicate these examples are given in the text, with larger files to upload provided in the Appendix. If you want to access UMansSysProp without using a web-browser we also provide a programmer friendly JSON API that enables you to call our suite of tools from your own code. This is described in detail on our ReadTheDocs.org webpage (https://umansysprop.readthedocs.org/) with an example provided in the Appendix. We also provide the source code for all predictive techniques provided on the site, covered by the GNU GPL license to encourage development of a user community. We have released this via a Github repository https://github.com/loftytopping/UManSysProp_public.git, that has an associated DOI for the exact model version given in this paper as provided by the Zenodo service (DOI 10.5281/zenodo.45143).

2 Chemo-informatics base of UManSysProp

The discipline of chemo-informatics typically concerns the use of both software and computational hardware techniques applied to a range of problems in chemistry. The emergence of the open-source movement has lead to a wealth of chemo-informatics software made available, including OpenBa-

[c7] ~~K(kappa)~~
[c8] *Text added.*

bel, which acts as the molecular parsing software behind the service described here. OpenBabel (O'Boyle et al., 2008) is a cross-platform suite of tools. Features include the ability to interchange chemical file formats and sub-structure searching, the latter particularly relevant. For more information, the reader is referred to the OpenBabel wiki $(http://openbabel.org/)$.

OpenBabel comes with wrappers for numerous languages including Perl, Ruby, Java and Python. Here we use the Python extensions of OpenBabel, Pybel, with the Flask $(http://flask.pocoo.org/)$ python web-application framework to provide a user friendly and device compatible interface. Figure 1 displays a basic schematic of user interaction with the site to perform specific calculations for a compound represented as a SMILES string.

All calculations rely on a representation of individual compounds, be it inorganic ions or neutral organic molecules. Raw model or measurement molecular information needs to be converted into an appropriate format for use in property predictions. Common molecular file formats include Wiswesser Line Notation (WLN), ROSDAL and SYBL. In addition, IUPAC and NIST recently developed the IUPAC International Chemical Identifier (InChl). Another linear notation using short ASCII strings is the SMILES format (Simplified Molecular Input Line Entry System), a simplified chemical notation that allows a user to represent a two dimensional chemical structure in linear textual form. For example, the SMILES notation for carbon dioxide is $O=C=O$, whereas cyclohexane is represented as $C1CCCCC1$. UMansSysProp uses SMILES for several reasons. The notation is commonly employed in commercial and public software for prediction of chemical properties. It can be imported by most molecule editors for conversion into 2D / 3D models and has a wide base of software support and extensive theoretical backing $(www.daylight.com)$. Common database searches for organic molecules include NIST $(http://webbook.nist.gov/chemistry/)$ and the National Chemical Database service $(http://cds.rsc.org/)$. From these, the SMILES representation of individual molecules can be found. In table 1, SMILES for common inorganic ions are provided along with a selection of organic compounds used by O'meara et al. (2014) in their saturation vapour pressure review paper. The CAS registry number, a unique identifier assigned to every chemical substance described in the open scientific literature, is also given. SMILES and CAS numbers can often be used interchangeably for searching specific compounds on the internet.

2.1 Parsing

To use the SMILES format requires the ability to extract substructure information from each string that is meaningful to each property predictive technique. OpenBabel has the ability to filter and search molecular files using the SMARTS format (created by Daylight Chemical Information Systems, Inc alongside the SMILES format). One can understand the role of SMARTS in the following sequential bullet points:

- Estimation methods within UManSysProp are based on the group contribution method
- Groups must therefore be automatically and unambiguously inferred from the SMILES strings
- SMARTS strings are used within UManSysProp to identify all groups (or substructures) required to estimate all provided properties
- The nomenclature for SMARTS string is described in the daylight theory webpages with many examples ($www.daylight.com$)
- Caution was given to identify the appropriate SMARTS string matching the groups (descriptors) included In the various predictive techniques selected within UManSysProp

Regarding the last point, it is important to note that the SMARTS used are highly specific to the property estimation method. For example, the canonical and isomeric SMILES string for Succinic acid is C(CC(=O)O)C(=O)O. By visiting the Daylight Theory webpage, generic examples on a variety of SMARTS are given, copied into table 2.

Caution must be used however with such generic SMARTS, depending on the expected range of molecules to be passed by the parsing routine. For techniques used in UManSysProp, an extensive manual analysis of compounds used in the MCM (Jenkin et al., 2012), and a subset of GECKO mechanism (Aumont et al., 2005), were used to validate derived SMARTS libraries. Table 3 is replicated from the Supplementary material of Barley et al. (2011) to illustrate the careful design of SMARTS for the vapour pressure technique by Nannoolal et al. (2008), hereafter referred to as the 'Nannoolal' method.[c1] Nannoolal et al. (2008)) [c2][c3]It is easy enough to identify all primary alcohols (SMARTS a in Table 3) but the Nannoolal method requires primary alcohols to be split between NG_35 (on a carbon chain of 5 or more atoms) and NG_36 (primary alcohols on a C4 or smaller chain) although the exact criteria for this split is not clear in the literature. In our work (Barley et al., 2011) [c4]the allocation of primary alcohols is achieved using a set of five SMARTS. SMARTS b in Table 3 identifies whether the primary alcohol is on a carbon chain of 5 or more atoms. This chain has to be terminated by carbon atoms (which may bear functional groups that are not part of this count), but the intermediate atoms can be N or O as well as C. Hence (using SMILES notation) OCCCO and OCCCCO would both have two alcohol groups belonging to NG_36 while OCCCCC, OCCOCC and OCCN(C)CC would have primary alcohols belonging to NG_35. The other three SMARTS account for the possible branching of this heavy atom chain:- thus OCC(C)(C) and OCN(C)C would both be NG_36 alcohols while OCC(C)(C)C and OCN(C)CC would be NG_35 alcohols. Each predictive

[c1] ~~As noted in that paper, whilst it is 'easy' to identify all primary alcohols (SMARTS 'a' in the table), the Nannoolal method requires primary alcohols to be split between NG 35 (carbon chain of 5 or more atoms with nomenclature defined in~~

[c2] ~~and NG 36 (primary alcohols on a C4 or smaller chain) although the exact criteria for this split is not clear in the literature. For our applications, the allocation of primary alcohols is achieved using a set of five SMARTS.~~

[c3] *Text added.*

[c4] *Text added.*

technique then has an appropriate library of SMARTS (figure 1). What happens if a technique does not capture all features of a molecule that might be passed for parsing? For example, as noted by Barley et al. (2011), alcohol groups attached to a carbon-carbon double bond (vinyl alcohols) are not covered by the Nannoolal method. SMARTS 'D' in table 3 are used to identify vinyl alcohols which are then treated like secondary alcohols within the predictive technique. For the AIOMFAC activity coefficient model (Zuend et al., 2008), care has been taken with the use of specific CHn-OH interaction terms. In the literature there are multiple choices for parameters representing these groups. For AIOMFAC, the distinction of the terms presented by Marcolli and Peter (2005) are only made in the case of pure alcohols and polyols, whereas in other cases the specific CHn groups are dropped. In the case of pure alcohols/polyols the categorisation of groups is solved by assuming all alkyl CHn are in a hydrocarbon tail unless a) they bear an -OH group; b) they are a methyl group attached to a CHn bearing an -OH group; and c) they are in a ring, aromatic, C=C or C#C group (Dr Andreas Zuend, pers comms).

[c1] All of the above checks of specificity were carried out by hand for atmospheric chemical mechanisms. Whilst the current facilities check for under- or over-counting of atoms for any given set of functional groups, a future development would need an automatic method of checking specificity for compounds falling outside of this subset following the discussions presented by Ruggeri and Takahama (2015).

3 Calculations currently provided

The facilities provided on UManSysProp are split into pure component properties and predictions of bulk and single particle aerosol behaviour. Pure component properties are limited to 5000 compounds, predictions involving activity coefficients limited to 1000 compounds [c2] at any one time via the web portal (not through direct use of the source code). Limitations on the number for species are largely down to computational cost considerations for calculations involving activity coefficients when providing this through a web portal. Optimising these calculations using external computational accelerators including GPUs is the subject of ongoing work and will be reported in a future publication. These are listed below along with the associated options, as displayed on the homepage:

- Equilibrium absorptive partitioning (Pankow, 1994; Donahue et al., 2006) calculations as a function of relative humidity (RH) and temperature. These allow users to account for 2000 species with gas phase abundances, entered manually or via a file upload, including an inorganic core, an involatile inert core, and treatment of non-ideality if required. Options for

[c1] *Text added.*
[c2] *Text added.*

vapour pressure predictive techniques are also provided. Available techniques are provided in drop-down menus described shortly

- Activity coefficients in liquid mixtures as a function of temperature. Separated into mixed organic and mixed organic/inorganic, users can apply both the AIOMFAC (Zuend et al., 2008) and UNIFAC (Fredenslund et al., 1975) activity coefficient models.

- Hygroscopic growth factors as a function of RH and temperature. Separated into inorganic and mixed inorganic/organic systems, users have the option to manually enter or upload compound definitions, selecting variable techniques for calculating densities, vapour pressures and activity coefficients. As part of these simulations, [c1][c2]K(kappa)-Köhler values (Kreidenweis et al., 2005) are provided, including an estimate of the equilibrium vapour pressure of organic compounds above the solution following the co-condensation hypothesis of Topping et al. (2013).

- Critical properties of organic compounds. Used in multiple density predictive techniques (Barley et al., 2013), predictions of Critical Volume, Critical Temperature and Pressure are given (Nannoolal et al., 2007; Myrdal and Yalkowsky, 1997; Joback and Reid, 1987).

- Sub-cooled liquid density predictions of organic compounds as a function of temperature, again via manual entry or file upload (Girolami, 1994; Bas, 1915; Bruce E. Poling, 2001).

- Pure component vapour pressures of organic compounds as a function of temperature via manual entry or file upload (Nannoolal et al., 2008, 2004; Joback and Reid, 1987; Myrdal and Yalkowsky, 1997; Stein and Brown, 1994; Compernolle et al., 2011)

The provision of any given property predictive technique on the portal is dictated by it having been subject to the peer review process where possible. For the pure component properties, this has included a critical review of vapour pressure (O'meara et al., 2014) and density techniques (Barley et al., 2013). The activity coefficient methods AIOMFAC (Zuend et al., 2008) and UNIFAC (Fredenslund et al., 1975) are discussed extensively in the literature. The theory behind hygroscopic growth calculations and absorptive partitioning simulations are also extensively covered in various papers (e.g. McFiggans et al. (2010)), with appropriate references provided on the website.

3.1 User interface and file formats

The UManSysProp website first provides a portal where users can enter or upload a SMILES string and predict the property of interest. Examples of supplying SMILES strings via the input are given in section 3.2 and 3.3. Whilst users have the option to display output on a new webpage via HTML

[c1] ~~K(kappa)~~
[c2] *Text added.*

as the default option, the following download options are also available. For more information on their use, please refer to the references given in [c3] [c4]parentheses:

- HTML (view in web browser)
- Excel file
- Python pickle file ($https://docs.python.org/2/library/pickle.html$)
- XML file ($http://en.wikipedia.org/wiki/XML$)
- Zipped CSV file
- JSON file ($http://en.wikipedia.org/wiki/JSON$)

If you want to access UManSysProp without using a web-browser we also provide a programmer friendly JSON API that enables you to call our suite of tools from your own code. This is described in detail on our ReadTheDocs.org webpage (https://umansysprop.readthedocs.org/) with an example provided in the Appendix where we briefly discuss future expansions.We also provide the source code for all predictive techniques provided on the site, covered by the GNU GPL license to encourage development of a user community. We have released this via a Github repository as detailed in section 5.

3.2 Pure component properties

Figure 2 displays the range of pure component vapour pressure predictions for a random subset of 30 compounds derived from the MCM compound dataset studied by Barley et al. (2011) at 298.15K. To generate the data, after clicking on the link to 'Pure component vapour pressures of organic compounds', a text-file with SMILES string was uploaded using the 'upload' facility. The graph was created separately using the IgorPro package, the predictive techniques covering the combined vapour pressure and boiling point methods of Nannoolal et al. (2008) and Nannoolal et al. (2004) ($Vp(N)Tb(N)$), Nannoolal et al. (2008) and Joback and Reid (1987) ($Vp(N)Tb(JR)$) and (Myrdal and Yalkowsky, 1997) with Stein and Brown (1994) ($Vp(MY)Tb(SB)$)). The list of SMILES is provided in table A1 of the Appendix for replicating the results. Simply copy and paste the SMILES provided and save as a text file to upload. The figure highlights general features discussed in the recent review by Bilde et al. (2015) in which the use of the boiling point method by Joback and Reid (1987) leads to much lower values, the discrepancy between all methods increasing as the vapour pressures decrease.

Figure 3 displays a range of pure component density predictions, for the methods reviewed by Barley et al. (2013), for the same 30 MCM compound dataset at 298.15K. As with the vapour pressure

[c3] ~~parenthese~~

[c4] *Text added.*

predictions, after clicking on the link to [c1][c2]'Sub-cooled liquid density' link on the homepage, a text-file with SMILES string was uploaded using the 'upload' facility.

3.3 Bulk partitioning predictions and single particle hygroscopic growth factors

For predictions of absorptive partitioning, the molar based partitioning model described by Barley et al. (2009) is used (equation 1-3):

$$C_{OA} = \sum_i C_i \varepsilon_i \tag{1}$$

$$\varepsilon_i = \left(1 + \frac{C_i^*}{C_{OA}}\right)^{-1} \tag{2}$$

$$C_i = \frac{10^6 \gamma_i P_i^o}{RT} \tag{3}$$

where C_i is the total loading of component i ($\mu moles.m^3$), P_i^o is the saturation vapour pressure of component i (atm), R is the ideal gas constant (8.2057.10^{-5}m$^3 atm.mol^{-1}.K^{-1}$), T is the temperature (K), γ_i is the activity coefficient of component i in the liquid phase and C_i^* is the effective saturation concentration of component i ($\mu moles.m^3$). To the best of our knowledge, only Schell et al. (2001) refer to using Newtons method for solving the equilibrium concentration. For the case of ideal solution thermodynamics ($\gamma_i = 1$), the root of the partitioning equation 1 is similarly solved here using Newtons method. This is applicable to any number of components and typically this results in 6-10 iterations to arrive at a solution for the total molar concentrations of secondary organic material. When including non-ideality, an iterative method is used where the value of C_{OA} is nudged at each iteration using a weighted average of the previous value. As before, the final solution satisfies the constraint that chemical potentials are equal for each component. On UManSysProp it is possible to include an inorganic core by specifying concentrations of the ions. The user can assume solution ideality or non-ideality by selecting the appropriate selection from the drop-down menu. In all cases it is assumed that concentrations of the ions remain fixed and there is no loss of semi-volatile components such as nitric or hydrochloric acid. These will be added in a future release, along with an account for multiple liquid phase partitioning (see section 4). In addition, it is possible to specify the concentration of an unidentified water soluble or water insoluble compound with a specific molecular weight that is included in the partitioning calculations.

As an example, table 4 displays the predicted equilibrium SOA mass loadings using the 30 most abundant compounds within a scaled biogenic simulation described by Barley et al. (2011) using

[c1] ~~Sub-cooled liquid~~
[c2] *Text added.*

the MCM. For our simulation we kept the temperature at 298.1K, varying the relative humidity between 50 and 90 %. A 2 $\mu g.m^{-3}$ core, with variable inorganic composition defined in table 3, was used to demonstrate the effect of assuming solution non-ideality with full inorganic-organic interactions, using the AIOMFAC model (Zuend et al., 2008), or assuming ideality. For the vapour pressure predictions, the vapour pressure and boiling point techniques of Nannoolal et al. (2008) and Nannoolal et al. (2004) were used. To conduct the partitioning simulations, the input file consists of the SMILES string of each compound in the left hand column and total concentration, in $molecules/cc$, in the right hand column. For a given relative humidity (RH), the abundance of water vapour and saturation vapour pressures are calculated implicitly as described in Barley et al. (2011). The input file used in these simulations can be found in table A2 of the Appendix. To replicate the predicted non-ideal mass at 50%RH click on the equilibrium absorptive partitioning link. To add an $(NH_4)_2SO_4$ inorganic core, first enter the SMILES string for the ammonium ion [NH4+] in the text entry box for 'Inorganic ions' with a concentration of 0.0303 $\mu moles.m^{-3}$. Next, click on the 'Add' button to create another entry for the sulphate ion. Enter the SMILES string [O-]S(=O)(=O)[O-] in the text entry box with a concentration of 0.0151 $\mu moles.m^{-3}$, consistent with a concentration of 2 $\mu g.m^{-3}(NH_4)_2SO_4$ core. For the organic compound click on the 'upload file' option and select the text file created from information provided in the Appendix. In the options for 'Interaction model' select 'Assume non-ideal interactions using the AIOMFAC model, using the default Vapour pressure method options. Click on the 'Calculate' button to retrieve predictions of total mass loadings, concentration of each component in the condensed phase and its activity coefficient. Results in table 4 demonstrate the influence of assuming ideality, or not, on calculated mass loadings as a function of RH. Whilst all cases demonstrate an increase in mass at higher humidities (Topping and McFiggans, 2012), the composition of the core has a noticeable effect on the magnitude of 'salting in' relative to the inert non-ideal test case. Following Topping et al. (2013), the assumption of solution non-ideality acts to 'buffer' the increase in mass relative to the ideal test case. Note that each scenario will be sensitive to the range of functionalities in compounds of interest, the relative abundance of each condensate (Topping and McFiggans, 2012) and the volatility profile (Topping et al., 2013), the example here simply acting as an example of how to use the partitioning simulations in UManSysProp. Figure 4 displays the range of predicted activity coefficients for each organic compound at equilibrium as a function of RH and predicted saturation vapour pressure for the same scenario with an $NaCl$ core. We have plotted the range of activity coefficients as a function of predicted P_{sat} as an illustration that, for specific cases, there may be no general trend, despite attempts in the literature to generalise more complex mixtures (Donahue et al., 2011). In this case, at higher RH, the activity coefficients of each component increases, explaining the reduced predicted mass compared to the ideal case for this specific simulation.

Predictions of aerosol hygroscopicity have been covered extensively in the literature, ranging from

detailed explicit thermodynamic models (Topping et al., 2005) to empirically determined parameter representations of water uptake (Kreidenweis et al., 2005). Topping and McFiggans (2012) discussed the potential problems associated with co-condensation of organic semi-volatile compounds on retrieved hygroscopicity in instruments and potential effects on cloud microphysics (Topping et al., 2013). The true effect of semi-volatile partitioning can only be predicted using a dynamic framework that accounts for the amount of absorptive mass, size dependencies through the Kelvin effect and instrument configuration. Rather than provide full dynamic simulations, users are provided with a 'potential' for semi-volatile loss from the assumed fixed non-aqueous composition through provision of equilibrium vapour pressures above the solution. UManSysProp hygroscopicity calculations assume water is the only compound that can re-partition between the gas and condensed phase, providing growth factors, [c1][c2]K(kappa)-Köhler values, solute mass fraction and equilibrium vapour pressures above the solution. Figure 5 displays predicted growth factors and [c3][c4]K(kappa)-Köhler values, using the AIOMFAC activity coefficient model, for $(NH_4)_2SO_4$, and $NaCl$ at 3 dry diameters of 100, 50 and 20nm assuming a surface tension of 72.224 $mN.m^{-1}$. In each case the relative molar concentration of ions must be used to define the 'dry' composition. For example, for $(NH_4)_2SO_4$, the SMARTS [NH4+] and [O-]S(=O)(=O)[O-] with relative molar concentrations of 2.0 and 1.0 are used and simulations run across a range of relative humidities from 50 to 95 $\%$.

Solute mass fractions are often compared to measurements derived from an Electrodynamic Balance, or EDB. Figure 6 compares the predicted mass increase with the measured data presented by Choi and Chan (2002) for an equimolar $(NH_4)_2SO_4$-Glutaric acid solutions. For more complex systems, table 5 displays the variable growth factor, with and without solution non-ideality, between 50 to 90 $\%$RH, of a mixed aerosol comprised of $(NH_4)_2SO_4$ and the 90 organic compounds, assuming an equimolar mixture, presented by O'meara et al. (2014) for their vapour pressure predictive technique evaluation study. The inputs used for these simulations can be found in table A3 of the Appendix.

Predictions of CCN activation potential are also provided. In these calculations, the maximum point of the Kohler curve is calculated using the secant method since the Kohler curve function is continuous and has only one maximum when water is the only semi-volatile allowed to re-equilibrate. Table 6 displays predicted [c1][c2]K(kappa)-Köhler values derived from the predicted critical point for an equimolar Succinic acid - $(NH_4)_2SO_4$ aerosol, and the two separate components, setting the surface tension to 72 $mN.m^{-1}$ as a function of dry size. [c3][c4]K(kappa)-Köhler values assuming solution ideality are also given, the values constant as one would expect without accounting for the effect of molecular interactions. This simply demonstrates that by using the AIOMFAC

[c1] ~~K(kappa)~~
[c2] *Text added.*
[c3] ~~K(kappa)~~
[c4] *Text added.*
[c1] ~~K(kappa)~~
[c2] *Text added.*
[c3] ~~K(kappa)~~
[c4] *Text added.*

activity coefficient model, even at the point of activation there is a significant deviation from ideality, due to organic-inorganic interactions, in a system in which solutes are 'forced' to remain in the condensed phase. As stated in the introduction, it is not the purpose of this paper to provide a full sensitivity analysis of such effects but rather provide researchers with the facility to do similar in specific case studies. Following Topping and McFiggans (2012), we also provide predictions of the equilibrium vapour pressure of the organic solutes, when present, to assess the potential for evaporation or condensation, similar to the predictions of sub-saturated hygroscopicity.

4 Future work

Alongside relevant property predictive techniques, all current aerosol particle predictions are based on equilibrium thermodynamics with single particle or bulk representations. Whilst providing useful insights into the role of composition dependent processes, capturing the evolution of an aerosol population requires dynamic ensemble frameworks (Topping et al., 2013). To meet these demands, future capabilities will include gas-particle box-model frameworks with a range of complexity with regards to the number of compounds and processes included in calculations. Regarding the latter aspect, current work involves profiling the use of external computational accelerators for mitigating the cost of accounting for solution non-ideality in future variants of UManSysProp to increase the maximum number of compounds allowed in subsequent calculations. [c1]Where property measurements are available, these might prove more accurate than any given estimation technique. With this in mind, in addition to extending the range of predictions provided, UManSysProp will also be linked to a standardized database of property measurements.

[c2] In addition to providing the online portal for users who do not want to use source code, and the JSON API for linking with our web portal without using a web browser, we also provide the source code for all predictive techniques provided on the site, covered by the GNU GPL license to encourage development of a user community. We have released this via a Github repository

The authors would like to acknowledge NERC grants NE/H002588/1, NE/J009202/1 and NE/J02175X/1 for enabling Dr Mark Barley to perform SMARTS library constructions and property prediction comparisons. Dr Topping was similarly funded through the National Centre for Atmospheric Science (NCAS). The authors would like to thank Prof Andreas Zuend, of McGill University, for his discussions on automating functional group selections for the AIOMFAC method. [c3]The authors would also like to thank Prof Markus Petters, of North Carolina State University, for discussions on the benefit of open source.

[c1] *Text added.*
[c2] *Text added.*
[c3] *Text added.*

7 Appendix

7.1 Accessing predictions outside of a web-browser

Here we provide brief details on how to call UManSysProp from your own code, without the need for a web-browser. For full details, please refer to our documentation on our ReadTheDocs.org webpage (https://umansysprop.readthedocs.org/). It is recommended that you use the local client installation UManSysProp from within the IPython shell. This is simply because the API is designed with documentation built in which can be queried iverom within the environment, and this is considerably easier from within the IPython shell. The client component of UManSysProp can be installed on any machine with Python available, provided you have Python 2.7 of greater. This includes Microsoft Windows, Mac OS X and other operating systems. On Ubuntu, the waveform PPA can be used for simple installation:

```
$$ sudo add-apt-repository ppa:waveform/ppa
$$ sudo apt-get update
$$ sudo apt-get install python-umansysprop
```

On other platforms, the package can be installed from PyPI. Specify the client option to pull in all dependencies required by the client component:

```
$$ sudo pip install "umansysprop[client]"
```

The first step in using the UManSysProp system is creating a UManSysProp instance, as demonstrated in the Python code snippet given below. By default this requires the URL of the UManSysProp server. Currently this is http://umansysprop.seaes.manchester.ac.uk

```
>>>import umansysprop.client
>>>client = umansysprop.client.UManSysProp('http://umansysprop.seaes.manchester.ac.uk')
```

Once you have a client instance, you can query it to find out what methods are available from the web API. Within the IPython shell this can be done simply by entering client. and pressing the Tab key twice. Alternatively, the following one-liner in the regular Python shell can be used to query non-private methods:

```
>>>[m for m in dir(client) if not m.startswith('_')]
>>>['absorptive_partitioning', 'sub_cooled_density', 'test', 'vapour_pressure']
```

Once youe selected a method to call you can discover what parameters it takes and what it expects in those parameters by querying the method documentation. Within the IPython shell this can be viewed simply by appending [c1][c2]'?' to the method name. Alternatively, the help() function can be used in a regular Python shell:

[c1] ?

[c2] *Text added.*

```
>>>help(client.vapour_pressure)
Help on method vapour_pressure in module umansysprop.client:
client.vapour_pressure(self, compounds, temperatures, vp_method,
    bp_method)...
    Calculates vapour pressures for all specified *compounds*
    (given as a sequence of SMILES strings) at all given
    *temperatures* (a sequence of floating point values giving
     temperatures in degrees Kelvin). The
    *vp_method* parameter is one of the strings:

    * 'nannoolal'
    * 'myrdal_and_yalkowsky'
    * 'evaporation'
    ...
```

The documentation for each tool can viewed on the UManSysProp API documentation page. Calling any of the tools will (in the event of success, given a valid SMILES or temperature quantity, for example) return a Result instance. This is simply a list() which contains a sequence of Table instances. Each table has a name and this can be used to access the table in the owning Result list. For example:

```
>>>result = client.vapour_pressure(['CCCC', 'C(CC(=O)O)C(=O)O',
'C(=O)(C(=O)O)O'], [298.15, 299.15, 300.15, 310.15], 'nannoolal',
'nannoolal')
>>> result
[<Table name="pressures">]
>>> result.pressures
<Table name="pressures">
```

Table instances have a friendly string representation which can be used at the command line for quick evaluation of the contents:

```
>>> print(result.pressures)
        |              CCCC | C(CC(=O)O)C(=O)O | C(=O)(C(=O)O)O
--------+-------------------+------------------+---------------
298.15  | 0.220914923012    |   -6.33293991048 | -5.19636054531
299.15  | 0.235479319348    |   -6.28117761855 | -5.15170377256
300.15  | 0.249933657549    |   -6.22986499517 | -5.10742877511
310.15  | 0.388688301563    |   -5.74023509659 | -4.68464352888
```

7.2 Inputs for replicating results presented in the main paper

SMILES string
COO
CC1(C2CCC(=O)C1C2)C
C(C(=O)OON(=O)=O)C1C(C(C(=O)C)C1)(C)C
CC(=O)OO
CCO
C(=O)(OON(=O)=O)CO
C1(C(C(CC=O)C1)(C)C)C(=O)C
CC(=O)C=C
C(=O)O
C=C(C)C=C
CC(=O)CO
C12C(C(CC(C1(C)OO)O)C2)(C)C
O=C1C2CC(C(=O)C1)C2(C)C
CC(=O)O
OOC1(CCC2C(C1C2)(C)C)CO
CC1(C2CCC(CO)(C1C2)ON(=O)=O)C
OCC(=O)C(C)(C)O
CC(=O)C
C(OO)C1C(C(C(=O)C)C1)(C)C
CCCC
O=CCC(=O)OON(=O)=O
CC=O
OCC(C=C)(C)OO
CC(=O)OON(=O)=O
N(=O)(=O)OC
OCC(=O)O
C12C(C(CC=C1C)C2)(C)C
C(O)C1C(C(C(=O)C)(C1)OO)(C)C
C(O)C=O
C=O

Table A1: To replicate the predicted vapour pressure and density predictions covered in section 3.1, copy and paste these SMILES strings into a text file and follow the procedures outlined in the main body of text. Be sure to copy just the SMILES strings.

SMILES string	
CC(=O)C	3.94381E+11
C=O	1.69175E+11
CC(=O)CO	79083500000
CC(=O)OON(=O)=O	77758900000
N(OC)(=O)=O	465002000
CC(=O)O	43340464110
C(O)C=O	42707600000
COO	35736100000
CC1(C2CCC(=O)C1C2)C	30157700000
CC(=O)OO	27534600000
C(OON(=O)=O)(=O)CO	24247000000
CC(=O)C=C	22076400000
C=C(C)C=C	21243400000
O=C1C2CC(C(C1)=O)C2(C)C	20484500000
OOC1(CCC2C(C1C2)(C)C)CO	16366200000
CC1(C2CCC(CO)(C1C2)ON(=O)=O)C	13559600000
OCC(=O)C(C)(C)O	12492500000
CCCC	12475100000
CC=O	12241000000
OCC(C=C)(C)OO	11486200000
C\12$C(C(C=C1)C2)(C)C$	11410300000
C(O)C1C(C(C(C)=O)(C1)OO)(C)C	11241600000
OCC(=O)O	10444600000
C(OO)C1C(C(C(C)=O)C1)(C)C	10090900000
C(C(=O)OON(=O)=O)C1C(C(C(C)=O)C1)(C)C	9914850000
CCO	9863210000
C1(C(C(CC=O)C1)(C)C)C(C)=O	9654730000
C(=O)O	9612616618
C12C(C(CC(C1(C)OO)O)C2)(C)C	9447850000
O=CCC(=O)OON(=O)=O	9287460000
CC	9233180000
CO	9195410000
C(=O)C=O	8495350000
CCC	8376030000
C12C(C(CC(C1(C)O)=O)C2)(C)C	8079110000

N(OC1(C2C(C(CC1O)C2)(C)C)C)(=O)=O	7921030000
CC1(C2C(C1C(=O)CC2=O)=O)C	7834600000
C=C(C)C=O	7401890000
OCC(=O)OO	7293180000
C=C1CCC2C(C1C2)(C)C	7086740000
CC(=O)C=O	6778590000
C=CC(C)(C)O	6760990000
C(ON(=O)=O)C1C(C(C(C)=O)C1)(C)C	6141130000
C=CC(=O)CO	6069190000
CC=C	5991610000
OCC(OO)(C)C(=O)CO	5955210000
OC(C(O)(C)C)COO	5722410000
CC(C(CC(CO)=O)=O)=O	5708910000
C=C(C(=O)OON(=O)=O)C	5368430000
C(=O)C1C(C(C(C)=O)C1)(C)C	5305790000
O(O)C1C(C(C(C)=O)C1)(C)C	5041190000
OCC(=O)C(C)=C	4959340000
CC(=O)C(=O)CC(=O)OON(=O)=O	4389470000
CC	4266050000
CC(=O)CC	4186060000
C12C(C(CC(C1(C)O)ON(=O)=O)C2)(C)C	4173050000
CC1(C2CC(=O)C(C1C2)=O)C	4164890000
C(C(=O)O)C1C(C(C(C)=O)C1)(C)C	4046940000
CC(C)(C(=O)OON(=O)=O)O	4036600000
CC(C)C	4015740000
CCCCC	3826030000
O=C(C)CON(=O)=O	3736020000
OC=C(/C)=O	3691840000
C=C	3681720000
C12C(C(CC(C1(C)OO)ON(=O)=O)C2)(C)C	3630020000
C(=O)CC=O	3506730000
CC(C)(O)C=O	3384520000
OC(=C=O)	3319200000
C(C(C(CO)OO)=O)O	3247530000
CC1(C2CCC1C2)C	3165480000

Table A2: SMILES strings and molecular abundance used for the partitioning predictions presented in section 3.2. To replicate the results, copy and past both SMILES and abundance information into a text file and following the procedures outlined in the main body of text. Please ensure there is space between a given SMILES string and the abundance.

SMILES string
O=C(O)CCCCCCCCCCCC
CCCCCCCCCCCCCC(=O)O
O=C(O)CCCCCCCCCCCCCC
CCCCCCCCCCCCCCCC(=O)O
O=C(O)CCCCCCCCCCCCCCCC
O=C(O)CCCCCCCCCCCCCCCCC
O=C(O)CCCCCCCCCCCCCCCCCC
O=C(O)CCCCCCCCCCCCCCCCCCC
C(=O)(C(=O)O)O
C(C(=O)O)C(=O)O
O=C(O)C(C(=O)O)C
O=C(O)C(O)C(=O)O
C(CC(=O)O)C(=O)O
O=C(O)CC(C(=O)O)C
O=C(O)C(O)(C)CC(=O)O
O=C(O)CC(O)C(=O)O
O=C(O)C(O)C(O)C(=O)O
C(C(C(=O)O)N)C(=O)O
O=C(O)C(=O)CC(=O)O
C(CC(=O)O)CC(=O)O
O=C(O)CCC(C(=O)O)C
O=C(O)CC(C)CC(=O)O
C(C(=O)O)C(CC(=O)O)(C(=O)O)O
C(CC(=O)O)C(C(=O)O)N
O=C(O)C(=O)CCC(=O)O
O=C(O)CC(=O)CC(=O)O
C(CCC(=O)O)CC(=O)O
O=C(O)CCCCCC(=O)O
O=C(O)CCCCCCC(=O)O
c1ccc(c(c1)C(=O)O)C(=O)O

O=C(O)c1cccc(C(=O)O)c1

c1cc(ccc1C(=O)O)C(=O)O

O=C(O)C1(C(=O)O)CC1

O=C(O)C1(C(=O)O)CCC1

O=C(O)C1CCCC1C(=O)O

O=C(O)C1CC(C(=O)O)CCC1

O=C(O)CCCCCCCC(=O)O

O=C(O)CCCCCCCCCC(=O)O

O=C(O)CCCCCCCCCCC(=O)O

O=C(O)CC1CC(C(=O)C)C1(C)C

COc1ccc(cc1)C(O)=O

COc1cc(ccc1O)C(=O)O

O=C(O)c1cc(OC)c(O)c(OC)c1

N(=O)(=O)c1cc(O)c(O)cc1

C1C2C(C(C(C(O1)O2)O)O)O

O=C(O)c1ccccc1N(C)C

O=C(O)c1cc(N(C)C)ccc1

OCC(O)CCC

C(C(CO)O)O

C(CCO)CO

CNCCO

OC(C)CC(O)C

Cc1c(cccc1(N(=O)(=O)))N(=O)(=O)

C(CO)N

c1(N(=O)(=O))ccccc1N

Clc1c(cc(OC)c(O)c1OC)C=O

ClC(C(=O)O)C

c1ccc(c(c1)C(=O)O)O

COCCOCCOCCOc1ccccc1Br

O=C(O)CCc1ccccc1OC

COC1=C(C=C(C=C1)CCC(=O)O)OC

Clc1ccc(cc1Cl)N(=O)(=O)

Clc1cc(O)c(O)cc1

Oc1c(cc(cc1O)C(C)(C)C)C(C)(C)C

O=CCC(CCCC(O)(C)C)C

COc1ccc(c(c1O)OC)Cl

Nc1cc(Cl)ccc1

NCCCO
CCC(CC)(c1ccc(cc1)N(=O)(=O))N(=O)(=O)
O=C(O)c1cccc(N(=O)=O)c1
N(=O)(=O)c1cccc(O)c1
O=C(O)c1ccc(N)cc1
COc1ccc(C=O)cc1
O=C(OCc1ccccc1)c2ccccc2O
CCCCOC(=O)c1ccccc1C(=O)OCCCC
O=Cc1cc(OCC)c(O)cc1
Oc1ccc(cc1OC)CC=C
O=C1OCC(O1)CO
c1cc2c(cc1C=O)OCO2
O=C(OCCC(C)C)c1ccccc1O
O=C1C(CCCC1)C2(O)CCCCC2
O=C(OC)c1ccccc1N
N(=O)(=O)c1c(c(cc(c1OC)C(C)(C)C)(N(=O)=O))C
OCCN(C)CCO
COc1ccc(cc1)C(C)=O
c1c(cc(cc1O)O)O
O=CCC1CC(C(=O)C)C1(C)C
OCCOCCOCCOCCO
CC(=O)OC(COC(=O)C)COC(C)=O
N(=O)(=O)OCCOCCOCCO(N(=O)(=O))

Table A3: 90 compounds used in an equimolar mixture in section 3.2 for calculating mixed inorganic/organic growth factors. Once again, to replicate those results, copy and paste these SMILES, with equal molar concentrations, into a text file and follow the procedure covered in the main body of text.

References

B. Aumont, S. Szopa, and S. Madronich. Modelling the evolution of organic carbon during its gas-phase tropospheric oxidation: development of an explicit model based on a self generating approach. *Atmospheric Chemistry and Physics*, 5:2497–2517, 2005. 966XT Times Cited:80 Cited References Count:106.

M. Barley, D. O. Topping, M. E. Jenkin, and G. McFiggans. Sensitivities of the absorptive partitioning model of secondary organic aerosol formation to the inclusion of water. *Atmospheric Chemistry and Physics*, 9(9): 2919–2932, 2009. 447KJ Times Cited:22 Cited References Count:42.

M. H. Barley, D. Topping, D. Lowe, S. Utembe, and G. McFiggans. The sensitivity of secondary organic aerosol (soa) component partitioning to the predictions of component properties - part 3: Investigation of condensed compounds generated by a near-explicit model of voc oxidation. *Atmospheric Chemistry and Physics*, 11(24):13145–13159, 2011. 870KH Times Cited:6 Cited References Count:48.

M. H. Barley, D. O. Topping, and G. McFiggans. Critical assessment of liquid density estimation methods for multifunctional organic compounds and their use in atmospheric science. *Journal of Physical Chemistry A*, 117(16):3428–3441, 2013. 134LJ Times Cited:1 Cited References Count:67.

G Le Bas. *The Molecular Volume of Liquid Chemical Compounds*. Longmans, 1915.

M. Bilde, K. Barsanti, M. Booth, C. D. Cappa, N. M. Donahue, E. U. Emanuelsson, G. McFiggans, U. K. Krieger, C. Marcolli, D. Tropping, P. Ziemann, M. Barley, S. Clegg, B. Dennis-Smither, M. Hallquist, A. M. Hallquist, A. Khlystov, M. Kulmala, D. Mogensen, C. J. Percival, F. Pope, J. P. Reid, M. A. V. R. da Silva, T. Rosenoern, K. Salo, V. P. Soonsin, T. Yli-Juuti, N. L. Prisle, J. Pagels, J. Rarey, A. A. Zardini, and I. Riipinen. Saturation vapor pressures and transition enthalpies of low-volatility organic molecules of atmospheric relevance: From dicarboxylic acids to complex mixtures. *Chemical Reviews*, 115(10):4115–4156, 2015. Cj3kz Times Cited:1 Cited References Count:246.

John P. Oonnell Bruce E. Poling, John M. Prausnitz. *Properties of Gases and Liquids, Fifth Edition*. McGraw-Hill Education, New York, Chicago, San Francisco, Athens, London, Madrid, Mexico City, Milan, New Delhi, Singapore, Sydney, Toronto, 2001.

M. Y. Choi and C. K. Chan. The effects of organic species on the hygroscopic behaviors of inorganic aerosols. *Environmental Science Technology*, 36(11):2422–2428, 2002. 563UR Times Cited:172 Cited References Count:50.

S. Compernolle, K. Ceulemans, and J. F. Muller. Evaporation: a new vapour pressure estimation methodfor organic molecules including non-additivity and intramolecular interactions. *Atmospheric Chemistry and Physics*, 11(18):9431–9450, 2011. 826QL Times Cited:29 Cited References Count:75.

N. M. Donahue, A. L. Robinson, C. O. Stanier, and S. N. Pandis. Coupled partitioning, dilution, and chemical aging of semivolatile organics. *Environmental Science Technology*, 40(8):2635–2643, 2006. 035HX Times Cited:388 Cited References Count:49.

N. M. Donahue, S. A. Epstein, S. N. Pandis, and A. L. Robinson. A two-dimensional volatility basis set: 1. organic-aerosol mixing thermodynamics. *Atmospheric Chemistry and Physics*, 11(7):3303–3318, 2011. 750LN Times Cited:95 Cited References Count:83.

A. Fredenslund, R. L. Jones, and J. M. Prausnitz. Group-contribution estimation of activity-coefficients in non-ideal liquid-mixtures. *Aiche Journal*, 21(6):1086–1099, 1975. Ax703 Times Cited:1774 Cited References Count:32.

Gregory S. Girolami. A simple "back of the envelope" method for estimating the densities and molecular volume of liquids and volumes. *J. of Chemical Education*, 71(11):962–964, 1994.

M. Hallquist, J. C. Wenger, U. Baltensperger, Y. Rudich, D. Simpson, M. Claeys, J. Dommen, N. M. Donahue, C. George, A. H. Goldstein, J. F. Hamilton, H. Herrmann, T. Hoffmann, Y. Iinuma, M. Jang, M. E. Jenkin, J. L. Jimenez, A. Kiendler-Scharr, W. Maenhaut, G. McFiggans, T. F. Mentel, A. Monod, A. S. H. Prevot, J. H. Seinfeld, J. D. Surratt, R. Szmigielski, and J. Wildt. The formation, properties and impact of secondary organic aerosol: current and emerging issues. *Atmospheric Chemistry and Physics*, 9(14):5155–5236, 2009. 477RF Times Cited:904 Cited References Count:660.

M. E. Jenkin, K. P. Wyche, C. J. Evans, T. Carr, P. S. Monks, M. R. Alfarra, M. H. Barley, G. B. McFiggans, J. C. Young, and A. R. Rickard. Development and chamber evaluation of the mcm v3.2 degradation scheme for beta-caryophyllene. *Atmospheric Chemistry and Physics*, 12(11):5275–5308, 2012. 959AH Times Cited:13 Cited References Count:77.

K. G. Joback and R. C. Reid. Estimation of pure-component properties from group-contributions. *Chemical Engineering Communications*, 57(1-6):233–243, 1987. L4338 Times Cited:655 Cited References Count:21.

S. M. Kreidenweis, K. Koehler, P. J. DeMott, A. J. Prenni, C. Carrico, and B. Ervens. Water activity and activation diameters from hygroscopicity data - part i: Theory and application to inorganic salts. *Atmospheric Chemistry and Physics*, 5:1357–1370, 2005. 933FY Times Cited:91 Cited References Count:31.

C. Marcolli and T. Peter. Water activity in polyol/water systems: new unifac parameterization. *Atmospheric Chemistry and Physics*, 5:1545–1555, 2005. 936XT Times Cited:26 Cited References Count:38.

G. McFiggans, D. O. Topping, and M. H. Barley. The sensitivity of secondary organic aerosol component partitioning to the predictions of component properties - part 1: A systematic evaluation of some available estimation techniques. *Atmospheric Chemistry and Physics*, 10(21):10255–10272, 2010. 680CQ Times Cited:18 Cited References Count:62.

P. B. Myrdal and S. H. Yalkowsky. Estimating pure component vapor pressures of complex organic molecules. *Industrial Engineering Chemistry Research*, 36(6):2494–2499, 1997. Xc612 Times Cited:78 Cited References Count:15.

Y. Nannoolal, J. Rarey, D. Ramjugernath, and W. Cordes. Estimation of pure component properties part 1. estimation of the normal boiling point of non-electrolyte organic compounds via group contributions and group interactions. *Fluid Phase Equilibria*, 226:45–63, 2004. 886OC Times Cited:86 Cited References Count:11.

Y. Nannoolal, J. Rarey, and D. Ramjugernath. Estimation of pure component properties. part 2. estimation of critical property data by group contribution. *Fluid Phase Equilibria*, 252(1-2):1–27, 2007. 151WH Times Cited:48 Cited References Count:30.

Y. Nannoolal, J. Rarey, and D. Ramjugernath. 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(1-2):117–133, 2008. 331YC Times Cited:74 Cited References Count:16.

N. M. O'Boyle, C. Morley, and G. R. Hutchison. Pybel: a python wrapper for the openbabel cheminformatics toolkit. *Chemistry Central Journal*, 2, 2008. 346JV Times Cited:61 Cited References Count:13.

S. O'meara, A. M. Booth, M. H. Barley, D. Topping, and G. McFiggans. An assessment of vapour pressure estimation methods. *Physical Chemistry Chemical Physics*, 16(36):19453–19469, 2014. Ao4ia Times Cited:2 Cited References Count:63.

J. F. Pankow. An absorption-model of the gas aerosol partitioning involved in the formation of secondary organic aerosol. *Atmospheric Environment*, 28(2):189–193, 1994. Nr212 Times Cited:502 Cited References Count:10.

G. Ruggeri and S. Takahama. Technical note: Development of chemoinformatic tools to enumerate functional groups in molecules for organic aerosol characterization. *Atmospheric Chemistry and Physics Discussions*, 15(22):33631–33674, 2015. doi:10.5194/acpd-15-33631-2015. URL http://www.atmos-chem-phys-discuss.net/15/33631/2015/.

B. Schell, I J. Ackermann, H. Hass, F S. Binkowski, and A. Ebel. Modeling the formation of secondary organic aerosol within a comprehensive air quality model system. *Journal of Geophysical Research*, 106: 282758293, 2001.

S. E. Stein and R. L. Brown. Estimation of normal boiling points from group contributions. *Journal of Chemical Information and Computer Sciences*, 34(3):581–587, 1994. Nn745 Times Cited:137 Cited References Count:15.

D. Topping, P. Connolly, and G. McFiggans. Cloud droplet number enhanced by co-condensation of organic vapours. *Nature Geoscience*, 6(6):443–446, 2013. 154ET Times Cited:18 Cited References Count:30.

D. O. Topping and G. McFiggans. Tight coupling of particle size, number and composition in atmospheric cloud droplet activation. *Atmospheric Chemistry and Physics*, 12(7):3253–3260, 2012. 926BQ Times Cited:21 Cited References Count:42.

D. O. Topping, G. B. McFiggans, and H. Coe. A curved multi-component aerosol hygroscopicity model framework: Part 2 - including organic compounds. *Atmospheric Chemistry and Physics*, 5:1223–1242, 2005. 929YP Times Cited:94 Cited References Count:81.

A. Zuend, C. Marcolli, B. P. Luo, and T. Peter. A thermodynamic model of mixed organic-inorganic aerosols to predict activity coefficients. *Atmospheric Chemistry and Physics*, 8(16):4559–4593, 2008. 343OR Times Cited:55 Cited References Count:98.

SMILES strings		
Compound name	SMILES string	CAS number
Hydrogen ion	[H+]	12408-02-5
Ammonium ion	[NH4+]	14798-03-9
Sodium ion	[Na+]	7440-23-5
Calcium ion	[Ca+2]	7440-70-2
Sulphate ion	[O-]S(=O)(=O)[O-]	14808-79-8
Nitrate ion	[N+](=O)([O-])[O-]	14797-55-8
Chloride ion	[Cl-]	16887-00-6
Tridecanoic acid	O=C(O)CCCCCCCCCCCC	638-53-9
Tetradecanoic acid	CCCCCCCCCCCCCC(=O)O	544-63-8
Oxalic Acid	C(=O)(C(=O)O)O	144-62-7

Malonic acid	C(C(=O)O)C(=O)O	141-82-2
2-methyl malonic acid	O=C(O)C(C(=O)O)C	516-05-2
2-hydroxy malonic acid (tartonic)	O=C(O)C(O)C(=O)O	80-69-3
2-keto succinic acid	O=C(O)C(=O)CC(=O)O	328-42-7
Glutaric acid	C(CC(=O)O)CC(=O)O	110-94-1
Adipic acid	C(CCC(=O)O)CC(=O)O	124-04-9
1,1-cyclopropane dicarboxylic acid	O=C(O)C1(C(=O)O)CC1	598-10-7
1,1-cylcobutane dicarboxylic acid	O=C(O)C1(C(=O)O)CCC1	5445-51-2
nitrocatechol	N(=O)(=O)c1cc(O)c(O)cc1	3316-09-4
levoglucosan	C1C2C(C(C(C(O1)O2)O)O)O	498-07-7
1,2-Pentanediol	OCC(O)CCC	5345-92-0
3,5-di-tert-Butylcatechol	Oc1c(cc(cc1O)C(C)(C)C)C(C)(C)C	1020-31-1
Ethyl vanillin	O=Cc1cc(OCC)c(O)cc1	121-32-4
Eugenol	Oc1ccc(cc1OC)CC=C	97-53-0
Glycerine carbonate	O=C1OCC(O1)CO	931-40-8
Heliotropin	c1cc2c(cc1C=O)OCO2	120-57-0
Pinonaldehyde	O=CCC1CC(C(=O)C)C1(C)C	2704-78-1
Tetraethylene glycol	OCCOCCOCCOCCO	112-60-7
Triacetin	CC(=O)OC(COC(=O)C)COC(C)=O	102-76-1

Table 1: Example SMILES of common inorganic ions and organic compounds with associated CAS numbers. [c2]Please note, CAS numbers cannot be used directly in the prediction facility.

SMARTS	Description
[CH2]	aliphatic carbon with two hydrogens (methylene carbon)
[!C;R]	(NOT aliphatic carbon) AND in ring
[!C;!R0]	same as above ("!R0" means not in zero rings)
[c,n&H1]	any aromatic carbon OR H-pyrrole nitrogen

[35*]	any atom of mass 35

Table 2: Generic SMARTS strings taken from the Daylight information webpage

Functional group	Nannoolal group	SMARTS
A -COOH	NG 44	a:-[#6][CX3](=[OX1])[OX2;H1]
B -OOH	New group	a:-[#6;!$([CX3]=[OX1])][OX2][OX2;H1]
C -OH (primary)	NG 35 or NG 36	a:-[OX2;H1][CX4;H2,H3] b:-[OX2;H1;!$(O[#6][#6,#7,#8][#6,#7,#8][#6,#7,#8][#6])][CX4;H2;H3;!$(O[#6][#6,#7,#8][#6,#7,#8][#6,#7,#8][#6]) c:[OX2;H1;!$(O[#6][#6,#7,#8][#6,#7]([#6])[#6])][CX4;H2,H3;!$(O[#6][#6,#7,#8][#6,#7]([#6])[#6])] d:-[OX2;H1;!$(O[#6][#6,#7]([#6])[#6,#7,#8][#6])][CX4;H2,H3;!$(O[#6][#6,#7]([#6])[#6,#7,#8][#6])

D -OH (vinyl)	assigned to OH (sec) NG 34	e:-[OX2;H1;!$(O[#6][#6][([#6])([#6])([#6])][CX4;H2,H3;!$(O[#6][#6]([#6])([#6])[#6])][OX2;H1;$([oX2;H1][CX3]=[CX3])]

Table 3: SMARTS for Nanoolal groups, as copied from the supplementary material of Barley et al. (2011).

RH (%)	$(NH_4)_2SO_4$ [non-ideal]	$NaCl$ [non-ideal]	NH_4NO_3 [non-ideal]	Inert core [non-ideal]	Inert core [ideal]
50	0.1181	0.2371	0.2179	0.0045	0.1547
60	0.1164	0.2157	0.1771	0.0054	0.1965
70	0.1260	0.2354	0.1655	0.0070	0.2691
80	0.1584	0.3138	0.1867	0.0104	0.4261
90	0.2845	0.5807	0.3104	0.0206	1.0008

Table 4: Predicted total organic mass loadings ($\mu g.m^{-3}$) from the most abundant 30 compounds generated from a gas phase degradation mechanism including both ideal and non-ideal solution thermodynamics. The composition of the core, with an abundance of 2 $\mu g.m^{-3}$, along with the assumption of solution ideality/non-ideality, is given above each column. For the 'inert core' a molecular weight of 200 $g.mol^{-1}$ was used, with solution thermodynamics accounting for interactions only between water and organic condensates.

RH (%)	GF (non-ideal)	[c1][c2]K(kappa)-Köhler (non-ideal)	GF (Ideal)	[c3][c4]K(kappa)-Köhler
50	1.0249	0.0796	1.0379	0.1226
60	1.0323	0.0698	1.0554	0.1226
70	1.0414	0.0589	1.0829	0.1226

[c2] ~~K(kappa)~~
[c2] *Text added.*
[c4] ~~K(kappa)~~
[c4] *Text added.*

80	1.0534	0.0461	1.1322	0.1226
90	1.0715	0.0302	1.2496	0.1226

Table 5: Variable growth factor, with and without solution non-ideality, between 50 to 90 %RH, of a mixed aerosol comprised of $(NH_4)_2SO_4$ and the 90 organic compounds, assuming an equimolar mixture, presented by O'meara et al. (2014) for their vapour pressure predictive technique evaluation study

	[c1][c2]K(kappa)-Köhler			Critical saturation ratio (%)	
Dry [c3][c4]diameter (nm)	Succinic - $(NH_4)_2SO_4$	$(NH_4)_2SO_4$	Succinic	$(NH_4)_2SO_4$-72mn/m	$(NH_4)_2SO_4$ -50 mN/m
100	0.6585	0.6317	0.3443	0.1463	0.0835
200	0.6824	0.6646	0.3473	0.0505	0.0290
300	0.6920	0.6788	0.3481	0.0272	0.0156
400	0.6973	0.6869	0.3485	0.0176	0.0101
500	0.7007	0.6922	0.3486	0.0125	0.0072
600	0.7032	0.6961	0.3487	0.0095	0.0055
700	0.7050	0.6989	0.3488	0.0075	0.0043
800	0.7064	0.7012	0.3489	0.0062	0.0035
900	0.7075	0.7030	0.3489	0.0051	0.0030
1000	0.7085	0.7045	0.3489	0.0044	0.0025
	Ideal [c5][c6]K(kappa)-Köhler values				
	0.4331	0.7235	0.3490		

[c2] ~~K(kappa)~~
[c2] *Text added.*
[c4] ~~size~~
[c4] *Text added.*
[c6] ~~K(kappa)~~
[c6] *Text added.*

Table 6: Predicted [c11c12]K(kappa)-Köhler values, with and without accounting for solution non-ideality, derived from the predicted critical point for an equimolar Succinic acid - $(NH_4)_2SO_4$ aerosol, and the two separate components, setting the surface tension to 72 $mN.m^{-1}$ as a function of dry [c13c14]diameter.

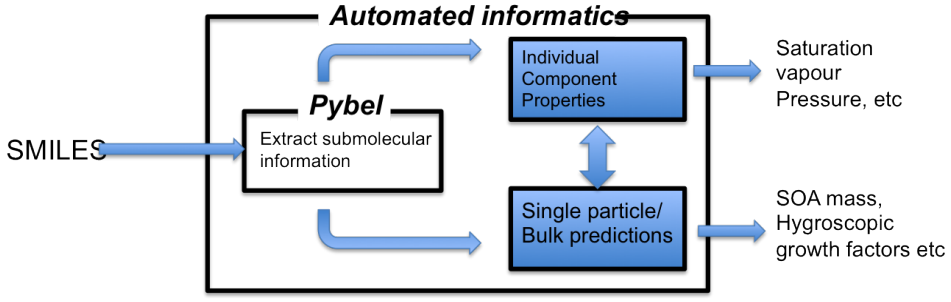

Figure 1. Workflow of calculations based on SMILES representation and the Pybel parsing module.

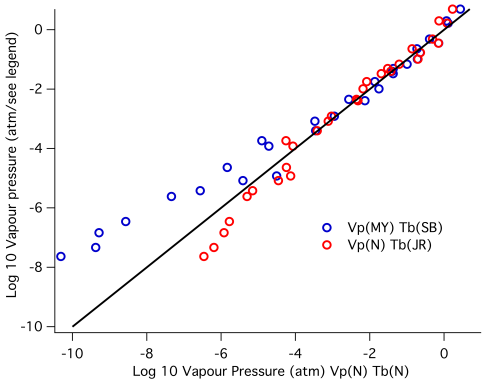

Figure 2. Predictions of pure component vapour pressures [c1](Log10 (atm)) at 298.15K using a subset of 30 compounds described by Barley et al. (2013), the list provided in the Appendix. The straight lines highlight the 1:1 relationship between predictions. As noted in the main body of text, the predictive techniques cover the combined vapour pressure and boiling point methods of Nannoolal et al. (2008) and Nannoolal et al. (2004) $(Vp(N)Tb(N))$, Nannoolal et al. (2008) and Joback and Reid (1987) $(Vp(N)Tb(JR))$ and (Myrdal and Yalkowsky, 1997) with Stein and Brown (1994) $(Vp(MY)Tb(SB))$.

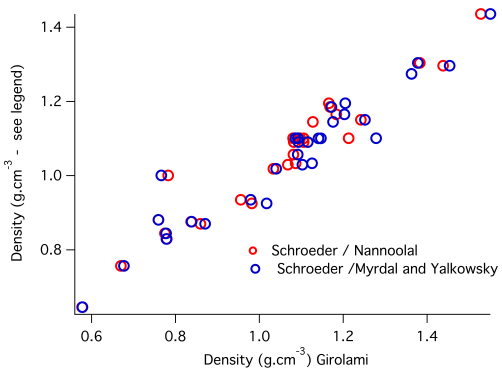

Figure 3. Predictions of pure component density [c2](g.cm-3) at 298.15K using the compounds described by Barley et al. (2011), the compound list provided in the Appendix. Methods used include those of Schroeder (Bruce E. Poling, 2001) combined with critical property estimation by both Joback and Reid (1987) and Nannoolal et al. (2008), compared to the method by Girolami (1994).

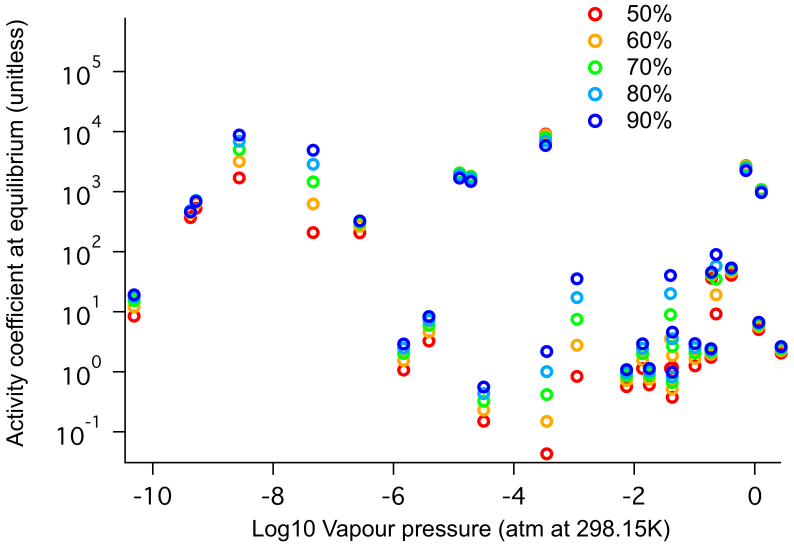

Figure 4. The range of predicted activity coefficients [c1](unitless on a mole fraction scale) for each organic compound as a function of saturation vapour pressure and RH.

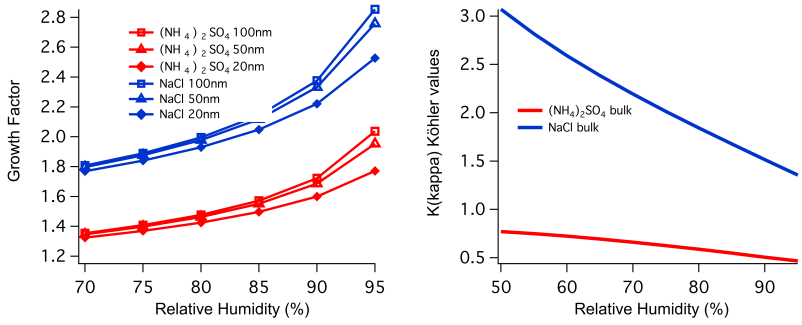

Figure 5. [left panel] Growth factor predictions for $(NH_4)_2SO_4$ and $NaCl$ particles of size100, 50 and 20nm diameter. [right panel] [c2][c3]K(kappa)-Köhler value predictions for $(NH_4)_2SO_4$ and $NaCl$ particles as a function of RH. All simulations use the AIOMFAC activity coefficient model.

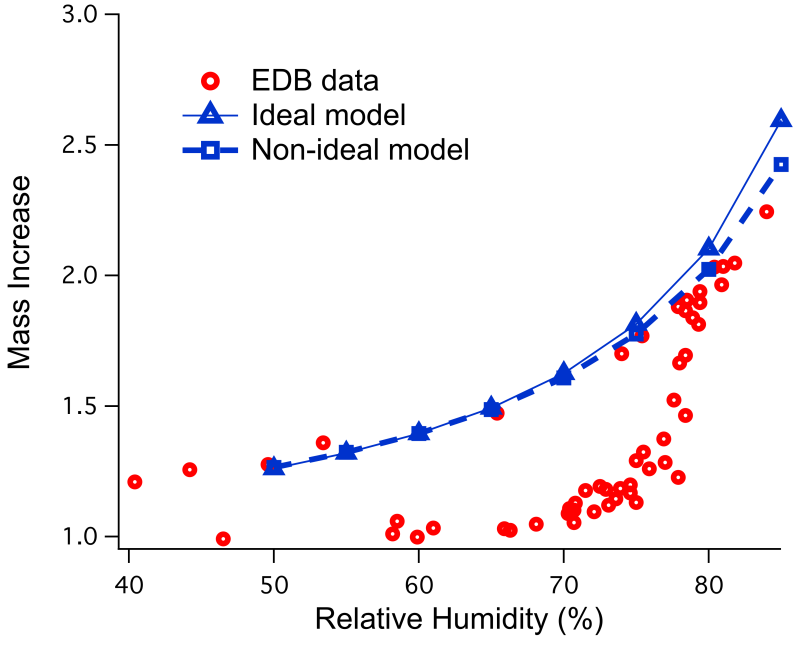

Figure 6. Comparisons of predicted water uptake, as particle mass increase (fraction), with EDB measurements on equimolar $(NH_4)_2SO_4$-Glutaric acid systems.