Atmos. Chem. Phys. Discuss., 15, 33631–33674, 2015
www.atmos-chem-phys-discuss.net/15/33631/2015/
doi:10.5194/acpd-15-33631-2015

Atmospheric Chemistry and Physics Discussions

Technical Note: Development of chemoinformatic tools to enumerate functional groups in molecules for organic aerosol characterization

G. Ruggeri and S. Takahama

ENAC/IIE Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland

Received: 1 October 2015 – Accepted: 26 October 2015 – Published: 27 November 2015

Correspondence to: S. Takahama (satoshi.takahama@epfl.ch)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

Functional groups (FGs) can be used as a reduced representation of organic aerosol composition in both ambient and environmental controlled chamber studies, as they retain a certain chemical specificity. Furthermore, FG composition has been informative for source apportionment, and various models based on a group contribution framework have been developed to calculate physicochemical properties of organic compounds. In this work, we provide a set of validated chemoinformatic patterns that correspond to: (1) groups incorporated in the SIMPOL.1 vapor pressure estimation model, (2) FGs that are measurable by Fourier transform infrared spectroscopy (FTIR), (3) a complete set of functional groups that can entirely describe the molecules comprised in the α-pinene and 1,3,5-trimethylbenzene MCMv3.2 oxidation schemes, and (4) bonds necessary for the calculation of carbon oxidation state. We also provide example applications for this set of patterns. We compare available aerosol composition reported by chemical speciation measurements and FTIR for different emission sources, and calculate the FG contribution to the O : C ratio of simulated gas phase composition generated from α-pinene photooxidation (using MCMv3.2 oxidation scheme).

1 Introduction

Atmospheric aerosols are complex mixtures of inorganic salts, mineral dust, sea salt, black carbon, metals, organic compounds, and water (Seinfeld and Pandis, 2006). Of these components, the organic fraction can comprise as much as 80 % of the aerosol mass (Lim and Turpin, 2002; Zhang et al., 2007), and yet eludes definitive characterization due to the number and diversity of molecule types. There have been many proposals for reducing representations in which a mixture of 10 000+ different types of molecules (Hamilton et al., 2004) are represented by some combination of their molecular size, carbon number, polarity, or elemental ratios (Pankow and Barsanti, 2009;

Kroll et al., 2011; Daumit et al., 2013; Donahue et al., 2012); many of which are associated with observable quantities (e.g., by aerosol mass spectrometry, AMS; Jayne et al., 2000, gas chromatography mass spectrometry, GC-MS and GCxGC-MS; Rogge et al., 1993; Hamilton et al., 2004). Molecular bonds or organic functional groups (FGs), which is the focus of this manuscript, can also be used to provide reduced representations for mixtures, and has been shown useful for organic mass (OM) quantification, source apportionment, and prediction of hygroscopicity and volatility (e.g., Russell, 2003; Donahue, 2011; Russell et al., 2011; Suda et al., 2014). Examples of property estimation methods include models for pure component vapor pressure (Pankow and Asher, 2008; Compernolle et al., 2011), UNIFAC and its variations for activity coefficients and viscosity (Ming and Russell, 2001; Griffin et al., 2002; Zuend et al., 2008, 2011). The FGs that can be detected or quantified by measurement vary widely by analytical technique, which include Fourier transform infrared spectroscopy (FTIR, Maria et al., 2002), Raman spectroscopy (Craig et al., 2015), nuclear magnetic resonance (NMR, Decesari et al., 2000; Cleveland et al., 2012), and gas chromatography with mass spectrometry and derivatization (Dron et al., 2010).

Projecting specific molecular information available through various forms of mass spectrometry (e.g., Williams et al., 2006; Kalberer et al., 2006; Laskin et al., 2012; Chan et al., 2013; Nguyen et al., 2013; Vogel et al., 2013; Yatavelli et al., 2014; Schilling Fahnestock et al., 2015; Chhabra et al., 2015) or model simulations employing explicit chemical mechanisms (e.g., Jenkin, 2004; Aumont et al., 2005; Herrmann et al., 2005) to a reduced dimensional space represented by some combination of FGs can be useful for measurement intercomparisons, or model-measurement comparisons. For this task, the aerosol community can benefit from developments in the chemoinformatics community. If the structure of a substance is described through its molecular (also referred to as chemical) graph (Balaban, 1985) – which is a set of atoms and their association through bonds – the abundance of arbitrary substructures (also called fragments) can be estimated through pattern matching algorithms called subgraph isomorphisms (Barnard, 1993; Ehrlich and Rarey, 2012; Kerber et al., 2014).

Structural information of molecules can be encoded in various representations, including a linear string of ASCII characters denoted as SMILES (Weininger, 1988). A corresponding set of fragments can be specified by SMARTS, which is a superset of the SMILES specification (DAYLIGHT Chemical Information Systems, Inc., 2015). There are many chemoinformatic packages that implement algorithms for pattern matching – for instance, OpenBabel (O'Boyle et al., 2011), Chemistry Development Kit (Steinbeck et al., 2003), OEChem (Openeye Scientific Software, Inc.), RDKit (Landrum, 2015), Indigo (GGA Software Services). The concept of using SMILES and SMARTS patterns have been reported for applications in the atmospheric chemistry community (Barley et al., 2011; COBRA, Fooshee et al., 2012). While some sets of SMARTS patterns for substructure matching can additionally be found in literature (Hann et al., 1999; Walters and Murcko, 2002; Olah et al., 2004; Enoch et al., 2008; Barley et al., 2011; Kenny et al., 2013) or on web databases – e.g., DAYLIGHT Chemical Information Systems, Inc. (DAYLIGHT Chemical Information Systems, Inc., 2015) – knowledge regarding the extent of specificity and validation of the defined patterns is not available.

In this work, we report specifications for four specific sets of substructures: (1) molecular fragments used by SIMPOL.1 for estimation of pure organic compound vapor pressures, (2) FGs contained in α-pinene and 1,3,5-trimethylbenzene photooxidation products defined in MCMv3.2 (Jenkin et al., 1997, 2003; Saunders et al., 2003; Bloss et al., 2005), obtained via http://mcm.leeds.ac.uk/MCM, (3) FGs that are measured or measurable (i.e., have absorption bands) for FTIR analysis (Pavia et al., 2008); and (4) bonds used for calculation of carbon oxidation state (Kroll et al., 2011). As there are several ways to define SMARTS patterns for substructure matching, we more generally prescribe a method for formulating patterns in such a way that permits a user to match and test not only the total number of FGs within a molecule, but to confirm that all atoms within molecule are classified uniquely into a set of FGs (except polyfunctional carbon, which can be associated with many FGs). We present a validation test for the groups defined, and show example applications for mapping molecules onto 2D-VBS space, inter-measurement comparison between OM composition reported by

GC-MS and FTIR for several source classes, and discuss implications for further applications. The patterns and software written for this manuscript are provided in a version controlled repository (Appendix C).

2 Methods

In this section, we present a series of patterns corresponding to substructures useful for vapor pressure estimation of FGs in molecules defined by measurements and chemical mechanisms (Sect. 2.1) as well as the methods and compound sets used for their validation (Sect. 2.2). We further describe the data set used for constructing a few example applications (Sect. 2.3).

2.1 Pattern specification for matching substructures

Four groups of patterns are defined: the first group (Table 1, substructures 1–33) corresponding to the complete set of FGs that can be found in the MCMv3.2 α-pinene and 1,3,5-trimethylbenzene oxidation scheme (Jenkin et al., 1997; Saunders et al., 2003), the second group corresponding to the FGs used to build the SIMPOL.1 model (Pankow and Asher, 2008) to predict pure components vapor pressures that are not present in the first set of patterns (Table 2), the third group used to study the FG abundance associated with FTIR measurements (FGs not specified before, containing Carbon, Oxygen and Nitrogen; Table 1, substructures 34–57) and the fourth group used to calculate the oxidation state of carbon atoms (Table 4). The OpenBabel toolkit (O'Boyle et al., 2011) is called through the pybel library (O'Boyle et al., 2008) in Python to search and enumerate abundances of fragments (most of which are specified by SMARTS) in each molecule (specified by SMILES). A few groups for which SMARTS patterns were difficult to obtain were calculated through algebraic relations specified through the string formatting syntax of the python programming language. In this syntax, values pre-computed through SMARTS matching are combined together to estimate properties for another group. While SMARTS can also describe ring definitions, ring perception is a difficult task partly due to the varying definitions of a ring, which must consider definition of aromaticity (tautomerism must also be considered) (Berger et al., 2004; May and Steinbeck, 2014). In this work, we use the smallest set of smallest rings (SSSR) (Downs et al., 1989) as defined by OpenBabel and many chemoinformatic software packages to enumerate the number of aromatic rings in this work. Ring enumeration is the only task specific to the software implementation, but otherwise the patterns specified can be ported to other software packages. The full implementation of patterns and scripts described in this manuscript are made available through an online repository (Appendix C).

We adapt chemoinformatic tools for use with SIMPOL.1 partly because the portable SMARTS pattern approach is more readily compatible with this model parameterization. We note that EVAPORATION vapor pressure model is fitted to more recent diacid measurements and includes positional information and non-linear interactions among FGs (Compernolle et al., 2011). Chemoinformatic tools can also be adapted for use with EVAPORATION by querying more specific structural information from the internal representations of molecular graphs defined by various software packages. With regards to the use of SIMPOL.1, vapor pressure predictions can also be improved by updating coefficients for the model with new estimates (Yeh and Ziemann, 2015).

SMARTS patterns for tallying the number of FGs can be formulated in many ways. We provide an illustration for the aldehyde FG group to illustrate the SMARTS development process used here. Propionaldehyde is used here as an example of an aldehyde. For this specific example, when querying for the aldehyde FG (substructure 9 in Table 1). The matched atoms will be 3, 4, 10 (as labeled in Fig. 1a), therefore the carbon atom to which the oxygen is attached, the oxygen and the hydrogen bonded to the sp^2 carbon. This SMARTS pattern was built in order to avoid matching molecules that are present in the α-pinene MCMv3.2 degradation mechanism such as the example in Fig. 1b, whose peroxide particular structure would be matched if the string `!$([O][O])` has not been explicitly specified. In addition, all atoms in the group are matched instead of just the identifying carbon or oxygen. The advantage of this strict protocol is that we can perform a validation to check that each atom is accounted for by at least one and only one group. From this validation, for a set of molecules we can confirm that we have defined a set of groups which accounts for all of the atoms in the system and that each atom (except polyfunctional carbon) is not claimed by multiple groups (Appendix A). This convention provides a means for apportioning FG contributions to atomic ratios (e.g., O : C, N : C) commonly used by the community. SMARTS patterns and coefficients associated with each bond type used to calculate the carbon oxidation state are reported in Table 4. Where the carbon oxidation state is calculated as the sum of the coefficients corresponding to the bonds to which the specific carbon is associated.

2.2 Data sets for validation

The first and the second groups of SMARTS patterns were validated against a set of 99 compounds (Table B1 in the Appendix) selected either from the compounds used in the development of the SIMPOL.1 method, either found to occur in atmospheric aerosol (Sect. 2.3) (Fraser et al., 2003, 1998; Grosjean et al., 1996), either selected from the ChemSpider database (Pence and Williams, 2010) to test for specific functionalities (eg. secondary amide), or selected from the MCMv3.2 α-pinene oxidation scheme. The patterns corresponding to the first group were further tested against the complete set of compounds present in the α-pinene and 1,3,5-trimethylbenzene MCMv3.2 oxidation schemes (408 compounds) in order to achieve a complete counting of all the atoms (carbon, oxygen, nitrogen and hydrogen atoms) and to avoid accounting heteroatoms to multiple FGs. The third group (Table 1, substructures 34–57) of SMARTS patterns was tested on a set of 26 compounds (Table B2) selected from the ChemSpider database and the fourth group (Table 4) was tested on a subset of 3 compounds extracted from the set of compounds used for the validation of the first group.

2.3 Data sets for example applications: molecules identified by GC-MS measurements and α-pinene and 1,3,5-TMB photooxidation products specified by the MCMv3.2 mechanism

A classic data set of organic compounds in primary organic aerosol (OA) from automobile exhaust (Rogge et al., 1993) and wood combustion (Rogge et al., 1998) quantified with GC-MS have been analyzed in order to retrieve the FG abundance of the mixture. Each compound, reported by common name in the literature, was converted to its corresponding SMILES string by querying the ChemSpider database with the Python ChemSpipy package (Swain, 2015), which wraps the ChemSpider application programming interface. FG composition, $\overline{\mathrm{OS}}_\mathrm{C}$ and pure component vapor pressure for each compound in the different reported mixture types was estimated using the substructure search algorithm described above. The algorithm previously described was applied to calculate the pure component vapor pressure for each compound i with the SIMPOL.1 model (Pankow and Asher, 2008). The total concentration in both gas phase and particle phase of the compounds reported by Rogge et al. (1993, 1998), and Hildemann et al. (1991) was used to estimate the OA concentration considering a seed concentration (C_OA) in the predilution channel of 10 mg m^{-3}, assuming fresh cooled emissions (Donahue et al., 2006). The total OA was then diluted of a factor of 1000 and the compounds partitioned between the two phases calculating their pure component saturation concentration (C_i^0) as calculated by Donahue et al. (2006), to derive the partitioning coefficient ξ_i.The results presented here refer to the calculated diluted aerosol concentration.

FG abundance of the set of compounds incorporated in the MCMv3.2 1,3,5-trimethylbenzene and α-pinene oxidation schemes was analyzed to demonstrate our validation scheme. Furthermore, the gas phase composition generated by α-pinene photoxidation in low NO$_x$ condition (α-pinene/NO$_x$ of 1.25), in the presence of propene as radical initiator was simulated using the Kinetic Pre-Processor (KPP, Damian et al., 2002; Sandu and Sander, 2006; Henderson, 2015) incorporating mechanistic information taken from MCMv3.2. Completeness and uniqueness requirements were tested and matched also for the α-pinene and propene MCMv3.2 degradation scheme. Initial concentrations of 240 ppb of α-pinene and 300 ppb of propene, a relative humidity of 61 % and a continuous irradiation were chosen as simulation conditions.

3 Results

3.1 Validation

Figure 2 shows that the enumerated FGs used by the SIMPOL.1 method (Table 2) are identical to the values enumerated manually. Matched FTIR FGs in Table 1 (substructures 34–57) are also identical to the true number of FGs in the set of compounds used for evaluation (Table B2), but is not shown as each group except alkane CH is matched at most once and a similar plot is uninformative. Figure 3 shows the completeness condition met, and Fig. 4 shows the specificity criterion fulfilled of the first set of chemoinformatic patterns (Table 1, substructures 1–33). The carbon atoms can be accounted by multiple FGs if polyfunctional: methylene and methyl groups are matched 2 and 3 times respectively by alkane CH group (substructure 1 in Table 1), while the carbon atoms in small molecules included in the test set have only 1 carbon atom that is matched 4 times (e.g. methanol, which has 3 alkane CH and 1 alcohol substructures).

3.2 Example applications

3.2.1 Mapping composition in 2-D volatility basis set space

The algorithm described has been used to project molecular composition to 2D-VBS space delineated by carbon oxidation and pure component saturation concentration (C^0). The properties of primary compounds measured in vehicle related OA by GC-MS are consistent with the hydrocarbon-like OA (HOA, Fig. 5) estimated from PMF analysis of AMS spectra, while the aerosol derived from wood combustion has char-

acteristics that overlap with HOA and biomass burning OA (BBOA) (also derived from PMF analysis of AMS spectra; Donahue et al., 2012) in $\overline{\mathrm{OS}}_\mathrm{C} - \log_{10}C^0$ space, and span a wide range of pure component saturation concentration. The carbon oxidation state and pure component saturation concentration of the compounds included in the α-pinene and 1,3,5-trimethylbenzene MCMv3.2 oxidation mechanisms are also reported in Fig. 5. It can be seen that the greatest abundance corresponds to intermediate volatility organic compounds (IVOC), and there is only a small overlap with the SVOC region. The lack of compounds in the LVOC region has to be taken into account when interpreting simulation results of MCMv3.2 with gas/particle partitioning; without inclusion of condensed-phase reactions.

Analyzing the oxidation state of the POA mixtures (Fig. 5) we can understand how the carbon oxidation state relates to the chemical characteristics of the mixture and in particular to the FG distribution. In the mixtures where more than 60 % of the carbon atoms are associated with alkane CH bonds, the highest number of carbons will have an oxidation state of −2, which corresponds to alkane CH associated with methylene groups (-CH_2-) (prevailing in the long alkane chains characteristic of HOA factor and vehicle exhaust primary OA). Where the carbon atoms are mostly associated with aromatic CH bonds a high fraction of the carbon atoms have oxidation state of −1 (carbon bonded to 1 hydrogen and 2 carbon atoms). Whereas when a non negligible fraction of the carbon in CH is associated with methyl groups (-CH_3), for example, in branched hydrocarbons; carbon atoms in −3 oxidation state reduce the mean oxidation state of the compound to which they are associated, and therefore the mean oxidation state of the mixture. A mean carbon oxidation state for a molecule greater than zero is due to a higher fraction of carbons attached to more electronegative atoms such as oxygen and nitrogen in this data set, and more generally includes sulfur and phosphorous atoms.

3.2.2 Source apportionment

In Fig. 6, the FG distributions of aerosol collected during biomass burning and vehicle emission studies (Rogge et al., 1998, 1993) have been compared to estimates from FTIR measurements of ambient samples separated by factor analytic decomposition (Positive Matrix Factorization or PMF; Paatero and Tapper, 1994) during September 2008 study period in California (Hawkins and Russell, 2010). The FTIR factor components from this study are consistent with similarly labeled factors from other field campaigns (Russell et al., 2011). The GC-MS reports approximately 20 % of the OA mass (Fine et al., 2002), while the FTIR quantifies around 90 % (Maria et al., 2003). For the study using FTIR, the biomass burning fraction was approximately 50 % of the total OA during intensive fire periods, and the fossil fuel combustion comprised 95 % of the overall OA during the campaign (Hawkins and Russell, 2010); these fractions form the bases for comparisons. The highest abundance of alkane-CH bonds in the compounds measured with GC-MS can be explained by the nature of the emissions and the preference of this analytical method to characterize the least oxidized fraction of the collected aerosol. Amine compounds and levoglucosan, a compound commonly associated with burning and decomposition of cellulose reported in modern measurements (Simoneit, 1999), are not reported for this study. High abundance of levoglucosan near particular fuel sources may be found in supermicron diameter particles (Radzi bin Abas et al., 2004), which may lead to higher concentrations of alcohol-COH. However, reported values by Hawkins and Russell (2010) are for submicron measurements, so low relative abundance of alcohol-COH is not surprising. Leithead et al. (2006) reported less than 2 % of mass contribution of levoglucosan to OA in urban and rural regions influenced by biomass burning. The high abundance of carbonyl-CO is associated with biomass burning aerosol and observed in FTIR measurements (Liu et al., 2009; Russell et al., 2009; Hawkins and Russell, 2010), but is not reported in this GC-MS analysis study. More recent methods including advanced derivatization (Dron et al., 2010) incorporate

analysis of compounds containing carbonyl-CO, and would be candidate studies for future comparisons.

FG distributions from vehicle sources are composed of greater than 90 % of alkane-CH by mass according to both estimation methods. While the fraction characterized by GC-MS and FTIR with PMF have associated uncertainties from derivatization and thermal separation in the chromatography column or in statistical separation, respectively, and lead to different fractions of mass reported. However, the approximate consistency in FG abundances estimated by the two methods, suggest that the fraction not analyzed by the GC-MS may not vary significantly from the measured fraction by FTIR in these aerosol types.

3.2.3 Oxygenated FG contribution to O : C ratio

Using the first set of SMARTS patterns we are able to match all the oxygen atoms, accounting them to specific FGs, in the α-pinene and 1,3,5-trimethylbenzene MCMv3.2 oxidation mechanisms. We can therefore calculate the contribution of each FG to the total O : C ratio of the gas phase mixture. In Fig. 7, contributions of FGs to the O : C ratio of the gas phase mixture generated by α-pinene photooxidation in low NO_x conditions (Sect. 2.3) is reported as a function of irradiation time. A singular peroxyacyl nitrate accounts for 26 % of the total gas phase mass (peroxyacetyl nitrate). It can be seen that the peroxyacyl nitrate functional group, although, accounts for the greatest fraction of the total O : C ratio after 20 h of simulation (53 % of the total O : C), as this FG contains five oxygen atoms per FG. A full analysis on oxidation products with gas/particle partitioning is discussed by Ruggeri et al. (2015). This type of analysis can provide intermediate information that is useful to suggest constraints on the form of oxygenation (and resulting change in organic mixture vapor pressure) assumed by simplified models such as the Statistical Oxidation Model (Cappa and Wilson, 2012).

4 Conclusions

In this study, we used chemoinformatic tools to create a validated set of patterns that allows us to perform substructure matching in molecules. With the chemoinformatic patterns developed we were able to calculate the pure component vapor pressure, the carbon oxidation state, the functional groups (FGs) associated with a set of compounds measured using GC-MS and to account for all the carbon, oxygen, nitrogen and hydrogen atoms in MCMv3.2 α-pinene and 1,3,5-trimethylbenzene oxidation mechanisms, without multiple matching.

With this method we were able to compare a chemical speciation technique (GC-MS) with FTIR spectroscopy and to project the different mixtures composition into the space of pure component saturation concentration and mean carbon oxidation state space. The FG distributions of vehicle exhaust and biomass burning emissions measured with GC-MS (Rogge et al., 1993, 1998) and statistically separated from ambient FTIR measurements (Hawkins and Russell, 2010) were compared. General agreement was observed for fossil fuel combustion associated with vehicle exhaust with $> 90\,\%$ of the mass attributed to alkane-CH, but for biomass burning the oxygenated fraction of OA reported by FTIR were 27 % higher. Comparison of this approach with more modern GC-MS measurements may be informative.

Pure component saturation concentration and carbon oxidation state were analyzed for two sets of compounds: the compounds measured using GC-MS (Rogge et al., 1993, 1998) in vehicle exhaust and biomass burning emission experiments and the compounds included in the MCMv3.2 oxidation schemes of α-pinene and 1,3,5-trimethylbenzene. From the analysis of this example data set, we relate FG distribution of the mixture to the carbon oxidation state. With the developed chemoinformatic patterns we were able to calculate the FG contribution to the overall O : C ratio of the gas phase generated by photooxidation of α-pinene in low-NO$_x$ conditions, simulated using the MCMv3.2 degradation mechanism (Jenkin et al., 1997; Saunders et al., 2003).

These applications can be further adapted for other methods developed to match substructures for other measurements, or enumerate groups used for the prediction of pure component vapor pressures (e.g., EVAPORATION, Compernolle et al., 2011) and activity coefficients (e.g., AIOMFAC Zuend et al., 2011). The proposed validation approach can also be followed to define FG patterns containing sulfur and halide bonds that absorb in the infrared region presently not included in this work.

Appendix A: Group validation

Let us consider a set of atoms A in molecule k and a set of FGs G. $\{a : \in A_k,\ a \in g\}$ denotes the set of atoms in molecule k which also is a member of group g, where $g \in G$. Completeness of G is defined by the condition that the combination of atoms matched by all groups in G comprises the full set of atoms A_k for every molecule:

$$\bigcup_{g \in G} \{a : a \in A_k,\ a \in g\} = A_k \quad \forall k$$

Specificity or minimal redundancy in G is defined by the condition that the intersection of atoms from all groups, excluding the set of polyfunctional carbon atoms $C_k^{\mathrm{p}} \subset A_k$, comprises the empty set:

$$\bigcap_{g \in G} \{a : a \in A_k,\ a \in g\} \backslash C_k^{\mathrm{p}} = \varnothing \quad \forall k$$

Appendix B: Compounds used for testing the chemoinformatic patterns

Tables B1 and B2 list compounds used for validation of patterns describing substructures 34–57 in Table 1 and all substructures in Table 2.

Appendix C: Software program

ASCII tables of the SMARTS patterns and the python program assembled for this work is released as Python program, APRL-SSP (APRL Substructure Search Program; https://github.com/stakahama/aprl-ssp), licensed under the GNU Public License version 3.0. In this program, series of scripts allow users to access the functionality of pybel and ChemSpiPy through input and output files defined as CSV-formatted tables.

Acknowledgements. The authors acknowledge funding from the Swiss National Science Foundation (200021_143298). The authors would like to thank S. Shipley for her initial contributions to the SMARTS pattern definitions, and to B. Henderson for his KPP code (http://github.com/barronh/kpp) and initial guidance.

References

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, doi:10.5194/acp-5-2497-2005, 2005. 33633

Balaban, A. T.: Applications of graph theory in chemistry, J. Chem. Inf. Comp. Sci., 25, 334–343, doi:10.1021/ci00047a033, 1985. 33633

Barley, M. H., Topping, D., Lowe, D., Utembe, S., and McFiggans, G.: 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, Atmos. Chem. Phys., 11, 13145–13159, doi:10.5194/acp-11-13145-2011, 2011. 33634

Barnard, J. M.: Substructure searching methods: old and new, J. Chem. Inf. Comp. Sci., 33, 532–538, doi:10.1021/ci00014a001, 1993. 33633

Berger, F., Flamm, C., Gleiss, P. M., Leydold, J., and Stadler, P. F.: Counterexamples in chemical ring perception, J. Chem. Inf. Comp. Sci., 44, 323–331, doi:10.1021/ci030405d, 2004. 33636

Bloss, C., Wagner, V., Jenkin, M. E., Volkamer, R., Bloss, W. J., Lee, J. D., Heard, D. E., Wirtz, K., Martin-Reviejo, M., Rea, G., Wenger, J. C., and Pilling, M. J.: Development of a

detailed chemical mechanism (MCMv3.1) for the atmospheric oxidation of aromatic hydrocarbons, Atmos. Chem. Phys., 5, 641–664, doi:10.5194/acp-5-641-2005, 2005. 33634

Brown, W. H., Foote, C. S., Iverson, B. L., and Anslyn, E. V.: Organic Chemistry, Books/Cole, Cengage learning, Belmont, USA, 2012. 33659

Cappa, C. D. and Wilson, K. R.: Multi-generation gas-phase oxidation, equilibrium partitioning, and the formation and evolution of secondary organic aerosol, Atmos. Chem. Phys., 12, 9505–9528, doi:10.5194/acp-12-9505-2012, 2012. 33642

Chan, M. N., Nah, T., and Wilson, K. R.: Real time in situ chemical characterization of submicron organic aerosols using Direct Analysis in Real Time mass spectrometry (DART-MS): the effect of aerosol size and volatility, Analyst, 138, 3749–3757, doi:10.1039/C3AN00168G, 2013. 33633

Chhabra, P. S., Lambe, A. T., Canagaratna, M. R., Stark, H., Jayne, J. T., Onasch, T. B., Davidovits, P., Kimmel, J. R., and Worsnop, D. R.: Application of high-resolution time-of-flight chemical ionization mass spectrometry measurements to estimate volatility distributions of α-pinene and naphthalene oxidation products, Atmos. Meas. Tech., 8, 1–18, doi:10.5194/amt-8-1-2015, 2015. 33633

Cleveland, M. J., Ziemba, L. D., Griffin, R. J., Dibb, J. E., Anderson, C. H., Lefer, B., and Rappengluck, B.: Characterization of urban aerosol using aerosol mass spectrometry and proton nuclear magnetic resonance spectroscopy, Atmos. Environ., 54, 511–518, doi:10.1016/j.atmosenv.2012.02.074, 2012. 33633

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, doi:10.5194/acp-11-9431-2011, 2011. 33633, 33636, 33644

Craig, R. L., Bondy, A. L., and Ault, A. P.: Surface enhanced Raman spectroscopy enables observations of previously undetectable secondary organic aerosol components at the individual particle level, Anal. Chem., 87, 7510–7514, doi:10.1021/acs.analchem.5b01507, 2015. 33633

Damian, V., Sandu, A., Damian, M., Potra, F., and Carmichael, G. R.: The kinetic preprocessor KPP-a software environment for solving chemical kinetics, Comput. Chem. Eng., 26, 1567–1579, doi:10.1016/S0098-1354(02)00128-X, 2002. 33638

Daumit, K. E., Kessler, S. H., and Kroll, J. H.: Average chemical properties and potential formation pathways of highly oxidized organic aerosol, Faraday Discuss., 165, 181–202, doi:10.1039/C3FD00045A, 2013. 33633

DAYLIGHT Chemical Information Systems, Inc.: available at: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, (last access: 30 September 2015), 2015. 33634

Decesari, S., Facchini, M. C., Fuzzi, S., and Tagliavini, E.: Characterization of water-soluble organic compounds in atmospheric aerosol: a new approach, J. Geophys. Res.-Atmos., 105, 1481–1489, doi:10.1029/1999JD900950, 2000. 33633

Donahue, N. M.: Atmospheric chemistry: the reaction that wouldn't quit, Nature Chemistry, 3, 98–99, doi:10.1038/nchem.941, 2011. 33633

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, 2635–2643, doi:10.1021/es052297c, 2006. 33638

Donahue, N. M., Henry, K. M., Mentel, T. F., Kiendler-Scharr, A., Spindler, C., Bohn, B., Brauers, T., Dorn, H. P., Fuchs, H., Tillmann, R., Wahner, A., Saathoff, H., Naumann, K.-H., Moehler, O., Leisner, T., Mueller, L., Reinnig, M.-C., Hoffmann, T., Salo, K., Hallquist, M., Frosch, M., Bilde, M., Tritscher, T., Barmet, P., Praplan, A. P., DeCarlo, P. F., Dommen, J., Prevot, A. S. H., and Baltensperger, U.: Aging of biogenic secondary organic aerosol via gas-phase OH radical reactions, P. Natl. Acad. Sci. USA, 109, 13503–13508, doi:10.1073/pnas.1115186109, 2012. 33633, 33640, 33672

Downs, G. M., Gillet, V. J., Holliday, J. D., and Lynch, M. F.: Review of ring perception algorithms for chemical graphs, J. Chem. Inf. Comp. Sci., 29, 172–187, doi:10.1021/ci00063a007, 1989. 33636

Dron, J., El Haddad, I., Temime-Roussel, B., Jaffrezo, J.-L., Wortham, H., and Marchand, N.: Functional group composition of ambient and source organic aerosols determined by tandem mass spectrometry, Atmos. Chem. Phys., 10, 7041–7055, doi:10.5194/acp-10-7041-2010, 2010. 33633, 33641

Ehrlich, H.-C. and Rarey, M.: Systematic benchmark of substructure search in molecular graphs – from Ullmann to VF2, Journal of Cheminformatics, 4, 13, doi:10.1186/1758-2946-4-13, 2012. 33633

Enoch, S. J., Madden, J. C., and Cronin, M. T. D.: Identification of mechanisms of toxic action for skin sensitisation using a SMARTS pattern based approach, SAR QSAR Environ. Res., 19, 555–578, doi:10.1080/10629360802348985, 2008. 33634

Fine, P. M., Cass, G. R., and Simoneit, B. R. T.: Chemical characterization of fine particle emissions from the fireplace combustion of woods grown in the southern United States, Environ. Sci. Technol., 36, 1442–1451, doi:10.1021/es0108988, 2002. 33641

Fooshee, D. R., Nguyen, T. B., Nizkorodov, S. A., Laskin, J., Laskin, A., and Badi, P.: COBRA: a computational brewing application for predicting the molecular composition of organic aerosols, Environ. Sci. Technol., 46, 6048–6055, doi:10.1021/es3003734, 2012. 33634

Fraser, M. P., Cass, G. R., Simoneit, B. R. T., and Rasmussen, R. A.: Air quality model evaluation data for organics. 5. C_6–C_{22} nonpolar and semipolar aromatic compounds, Environ. Sci. Technol., 32, 1760–1770, doi:10.1021/es970349v, 1998. 33637

Fraser, M. P., Cass, G. R., and Simoneit, B. R. T.: Air quality model evaluation data for organics. 6. C_3–C_{24} organic acids, Environ. Sci. Technol., 37, 446–453, doi:10.1021/es0209262, 2003. 33637

Griffin, R. J., Dabdub, D., Kleeman, M. J., Fraser, M. P., Cass, G. R., and Seinfeld, J. H.: Secondary organic aerosol – 3. Urban/regional scale model of size- and composition-resolved aerosols, J. Geophys. Res.-Atmos., 107, 4334, doi:10.1029/2001JD000544, 2002. 33633

Grosjean, E., Grosjean, D., Fraser, M. P., and Cass, G. R.: Air quality model evaluation data for organics. 3. Peroxyacetyl nitrate and peroxypropionyl nitrate in Los Angeles air, Environ. Sci. Technol., 30, 2704–2714, doi:10.1021/es9508535, 1996. 33637

Hamilton, J. F., Webb, P. J., Lewis, A. C., Hopkins, J. R., Smith, S., and Davy, P.: Partially oxidised organic components in urban aerosol using GCXGC-TOF/MS, Atmos. Chem. Phys., 4, 1279–1290, doi:10.5194/acp-4-1279-2004, 2004. 33632, 33633

Hann, M., Hudson, B., Lewell, X., Lifely, R., Miller, L., and Ramsden, N.: Strategic pooling of compounds for high-throughput screening, J. Chem. Inf. Comp. Sci., 39, 897–902, doi:10.1021/ci990423o, 1999. 33634

Hawkins, L. N. and Russell, L. M.: Oxidation of ketone groups in transported biomass burning aerosol from the 2008 Northern California Lightning Series fires, Atmos. Environ., 44, 4142–4154, doi:10.1016/j.atmosenv.2010.07.036, 2010. 33641, 33643, 33673

Henderson, B. H.: Kinetic Pre-Processor with updates to allow working with MCM, available at: http://github.com/barronh/kpp (last access: 30 September 2015), 2015. 33638

Herrmann, H., Tilgner, A., Barzaghi, P., Majdik, Z., Gligorovski, S., Poulain, L., and Monod, A.: Towards a more detailed description of tropospheric aqueous phase organic chemistry: CAPRAM 3.0, Atmos. Environ., 39, 4351–4363, doi:10.1016/j.atmosenv.2005.02.016, 2005. 33633

Hildemann, L. M., Markowski, G. R., and Cass, G. R.: Chemical-composition of emissions from urban sources of fine organic aerosol, Environ. Sci. Technol., 25, 744–759, doi:10.1021/es00016a021, 1991. 33638

Jayne, J. T., Leard, D. C., Zhang, X. F., Davidovits, P., Smith, K. A., Kolb, C. E., and Worsnop, D. R.: Development of an aerosol mass spectrometer for size and composition analysis of submicron particles, Aerosol Sci. Tech., 33, 49–70, doi:10.1080/027868200410840, 2000. 33633

Jenkin, M. E.: Modelling the formation and composition of secondary organic aerosol from α- and β-pinene ozonolysis using MCM v3, Atmos. Chem. Phys., 4, 1741–1757, doi:10.5194/acp-4-1741-2004, 2004. 33633

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, doi:10.1016/S1352-2310(96)00105-7, 1997. 33634, 33635, 33643

Jenkin, M. E., Saunders, S. M., Wagner, V., and Pilling, M. J.: Protocol for the development of the Master Chemical Mechanism, MCM v3 (Part B): tropospheric degradation of aromatic volatile organic compounds, Atmos. Chem. Phys., 3, 181–193, doi:10.5194/acp-3-181-2003, 2003. 33634

Kalberer, M., Sax, M., and Samburova, V.: Molecular size evolution of oligomers in organic aerosols collected in urban atmospheres and generated in a smog chamber, Environ. Sci. Technol., 40, 5917–5922, doi:10.1021/es0525760, 2006. 33633

Kenny, P. W., Montanari, C. A., and Prokopczyk, I. M.: ClogPalk: a method for predicting alkane/water partition coefficient, J. Comput. Aid. Mol. Des., 27, 389–402, doi:10.1007/s10822-013-9655-5, 2013. 33634

Kerber, A., Laue, R., Meringer, M., Raocker, C., and Schymanski, E.: Mathematical Chemistry and Chemoinformatics: Structure Generation, Elucidation and Quantitative Structure-Property Relationships, Walter de Gruyter, 2014. 33633

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, doi:10.1038/nchem.948, 2011. 33633, 33634

Landrum, G.: RDKit: Open-source cheminformatics, available at: http://www.rdkit.org (last access: 30 September 2015), 2015. 33634

Laskin, J., Eckert, P. A., Roach, P. J., Heath, B. S., Nizkorodov, S. A., and Laskin, A.: Chemical analysis of complex organic mixtures using reactive nanospray desorption electrospray ionization mass spectrometry, Anal. Chem., 84, 7179–7187, doi:10.1021/ac301533z, 2012. 33633

Leithead, A., Li, S.-M., Hoff, R., Cheng, Y., and Brook, J.: Levoglucosan and dehydroabietic acid: Evidence of biomass burning impact on aerosols in the Lower Fraser Valley, Atmos. Environ., 40, 2721–2734, doi:10.1016/j.atmosenv.2005.09.084, 2006. 33641

Lim, H. J. and Turpin, B. J.: Origins of primary and secondary organic aerosol in Atlanta: results' of time-resolved measurements during the Atlanta supersite experiment, Environ. Sci. Technol., 36, 4489–4496, doi:10.1021/es0206487, 2002. 33632

Liu, S., Takahama, S., Russell, L. M., Gilardoni, S., and Baumgardner, D.: Oxygenated organic functional groups and their sources in single and submicron organic particles in MILAGRO 2006 campaign, Atmos. Chem. Phys., 9, 6849–6863, doi:10.5194/acp-9-6849-2009, 2009. 33641

Maria, S. F., Russell, L. M., Turpin, B. J., and Porcja, R. J.: FTIR measurements of functional groups and organic mass in aerosol samples over the Caribbean, Atmos. Environ., 36, 5185–5196, doi:10.1016/S1352-2310(02)00654-4, 2002. 33633

Maria, S. F., Russell, L. M., Turpin, B. J., Porcja, R. J., Campos, T. L., Weber, R. J., and Huebert, B. J.: Source signatures of carbon monoxide and organic functional groups in Asian Pacific Regional Aerosol Characterization Experiment (ACE-Asia) submicron aerosol types, J. Geophys. Res.-Atmos., 108, 8637, doi:10.1029/2003JD003703, 2003. 33641

May, J. W. and Steinbeck, C.: Efficient ring perception for the Chemistry Development Kit, Journal of Cheminformatics, 6, 3, doi:10.1186/1758-2946-6-3, 2014. 33636

Ming, Y. and Russell, L. M.: Predicted hygroscopic growth of sea salt aerosol, J. Geophys. Res.-Atmos., 106, 28259–28274, doi:10.1029/2001JD000454, 2001. 33633

Nguyen, T. B., Nizkorodov, S. A., Laskin, A., and Laskin, J.: An approach toward quantification of organic compounds in complex environmental samples using high-resolution electrospray ionization mass spectrometry, Analytical Methods, 5, 72–80, doi:10.1039/c2ay25682g, 2013. 33633

Nic, M., Jirat, J., and Kosata, B.: IUPAC Compendium of Chemical Terminology – the Gold Book, available at: http://goldbook.iupac.org (last access: 30 September 2015), 2014. 33659

O'Boyle, N. M., Morley, C., and Hutchison, G. R.: Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit, Chem. Cent. J., 2, 5, doi:10.1186/1752-153X-2-5, 2008. 33635

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R.: Open Babel: an open chemical toolbox, Journal of Cheminformatics, 3, 33, doi:10.1186/1758-2946-3-33, 2011. 33634, 33635

Olah, M., Bologa, C., and Oprea, T.: An automated PLS search for biologically relevant QSAR descriptors, J. Comput. Aid. Mol. Des., 18, 437–449, doi:10.1007/s10822-004-4060-8, 2004. 33634

Paatero, P. and Tapper, U.: Positive matrix factorization – a nonnegative factor model with optimal utilization of error-estimates of data values, Environmetrics, 5, 111–126, doi:10.1002/env.3170050203, 1994. 33641

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, 2773–2796, doi:10.5194/acp-8-2773-2008, 2008. 33633, 33635, 33638, 33660, 33665, 33669

Pankow, J. F. and Barsanti, K. C.: The carbon number-polarity grid: A means to manage the complexity of the mix of organic compounds when modeling atmospheric organic particulate matter, Atmos. Environ., 43, 2829–2835, doi:10.1016/j.atmosenv.2008.12.050, 2009. 33632

Pavia, D., Lampman, G., and Kriz, G.: Introduction to Spectroscopy, Brooks/Cole Pub Co., Belmont, CA (USA), 2008. 33634

Pence, H. E. and Williams, A.: ChemSpider: an online chemical information resource, J. Chem. Educ., 87, 1123–1124, doi:10.1021/ed100697w, 2010. 33637

Radzi bin Abas, M., Oros, D. R., and Simoneit, B. R. T.: Biomass burning as the main source of organic aerosol particulate matter in Malaysia during haze episodes, Chemosphere, 55, 1089–1095, doi:10.1016/j.chemosphere.2004.02.002, 2004. 33641

Rogge, W. F., Hildemann, L. M., Mazurek, M. A., Cass, G. R., and Simoneit, B. R. T.: Sources of fine organic aerosol. 2. Noncatalyst and catalyst-equipped automobiles and heavy-duty diesel trucks, Environ. Sci. Technol., 27, 636–651, doi:10.1021/es00041a007, 1993. 33633, 33638, 33641, 33643, 33672, 33673

Rogge, W. F., Hildemann, L. M., Mazurek, M. A., Cass, G. R., and Simoneit, B. R. T.: Sources of fine organic aerosol. 9. Pine, oak and synthetic log combustion in residential fireplaces, Environ. Sci. Technol., 32, 13–22, doi:10.1021/es960930b, 1998. 33638, 33641, 33643, 33672, 33673

Ruggeri, G., Alexander, F. B., Takahama, S., and Henderson, B. H.: Comparison of simulation and measurements of α-pinene and 1,3,5-trimethylbenzene degradation using functional group abundance as a metric, in preparation, 2015. 33642

Russell, L. M.: Aerosol organic-mass-to-organic-carbon ratio measurements, Environ. Sci. Technol., 37, 2982–2987, doi:10.1021/es026123w, 2003. 33633

Russell, L. M., Bahadur, R., Hawkins, L. N., Allan, J., Baumgardner, D., Quinn, P. K., and Bates, T. S.: Organic aerosol characterization by complementary measurements of chemical bonds and molecular fragments, Atmos. Environ., 43, 6100–6105, doi:10.1016/j.atmosenv.2009.09.036, 2009. 33641

Russell, L. M., Bahadur, R., and Ziemann, P. J.: Identifying organic aerosol sources by comparing functional group composition in chamber and atmospheric particles, P. Natl. Acad. Sci. USA, 108, 3516–3521, doi:10.1073/pnas.1006461108, 2011. 33633, 33641

Sandu, A. and Sander, R.: Technical note: Simulating chemical systems in Fortran90 and Matlab with the Kinetic PreProcessor KPP-2.1, Atmos. Chem. Phys., 6, 187–195, doi:10.5194/acp-6-187-2006, 2006. 33638

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, doi:10.5194/acp-3-161-2003, 2003. 33634, 33635, 33643

Schilling Fahnestock, K. A., Yee, L. D., Loza, C. L., Coggon, M. M., Schwantes, R., Zhang, X., Dalleska, N. F., and Seinfeld, J. H.: Secondary organic aerosol composition from C12 alkanes, J. Phys. Chem. A, 119, 4281–4297, doi:10.1021/jp501779w, 2015. 33633

Seinfeld, J. H. and Pandis, S. N.: Atmospheric Chemistry and Physics: from Air Pollution to Climate Change, 2nd edn., John Wiley & Sons, New York, 2006. 33632

Simoneit, B. R. T.: A review of biomarker compounds as source indicators and tracers for air pollution, Environ. Sci. Pollut. R., 6, 159–169, doi:10.1007/BF02987621, 1999. 33641

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, J. Chem. Inf. Comp. Sci., 43, 493–500, doi:10.1021/ci025584y, 2003. 33634

Suda, S. R., Petters, M. D., Yeh, G. K., Strollo, C., Matsunaga, A., Faulhaber, A., Ziemann, P. J., Prenni, A. J., Carrico, C. M., Sullivan, R. C., and Kreidenweis, S. M.: Influence of functional groups on organic aerosol cloud condensation nucleus activity, Environ. Sci. Technol., 48, 10182–10190, doi:10.1021/es502147y, 2014. 33633

Swain, M.: ChemSpiPy, available at: http://chemspipy.readthedocs.org (last access: 30 September 2015), 2015. 33638

Vogel, A. L., Äijälä, M., Corrigan, A. L., Junninen, H., Ehn, M., Petäjä, T., Worsnop, D. R., Kulmala, M., Russell, L. M., Williams, J., and Hoffmann, T.: In situ submicron organic aerosol characterization at a boreal forest research station during HUMPPA-COPEC 2010 using soft and hard ionization mass spectrometry, Atmos. Chem. Phys., 13, 10933–10950, doi:10.5194/acp-13-10933-2013, 2013. 33633

Walters, W. and Murcko, M. A.: Prediction of "drug-likeness", Adv. Drug Deliver. Rev., 54, 255–271, doi:10.1016/S0169-409X(02)00003-0, 2002. 33634

Weininger, D.: Smiles, a chemical language and information-system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comp. Sci., 28, 31–36, doi:10.1021/ci00057a005, 1988. 33634

Williams, B. J., Goldstein, A. H., Kreisberg, N. M., and Hering, S. V.: An in-situ instrument for speciated organic composition of atmospheric aerosols: thermal desorption aerosol GC/MS-FID (TAG), Aerosol Sci. Tech., 40, 627–638, doi:10.1080/02786820600754631, 2006. 33633

Yatavelli, R. L. N., Stark, H., Thompson, S. L., Kimmel, J. R., Cubison, M. J., Day, D. A., Campuzano-Jost, P., Palm, B. B., Hodzic, A., Thornton, J. A., Jayne, J. T., Worsnop, D. R., and Jimenez, J. L.: Semicontinuous measurements of gas–particle partitioning of organic acids in a ponderosa pine forest using a MOVI-HRToF-CIMS, Atmos. Chem. Phys., 14, 1527–1546, doi:10.5194/acp-14-1527-2014, 2014. 33633

Yeh, G. K. and Ziemann, P. J.: Gas-wall partitioning of oxygenated organic compounds: measurements, structure–activity relationships, and correlation with gas chromatographic retention factor, Aerosol Sci. Tech., 49, 727–738, doi:10.1080/02786826.2015.1068427, 2015. 33636

Zhang, Q., Jimenez, J. L., Canagaratna, M. R., Allan, J. D., Coe, H., Ulbrich, I., Alfarra, M. R., Takami, A., Middlebrook, A. M., Sun, Y. L., Dzepina, K., Dunlea, E., Docherty, K., DeCarlo, P. F., Salcedo, D., Onasch, T., Jayne, J. T., Miyoshi, T., Shimono, A., Hatakeyama, S., Takegawa, N., Kondo, Y., Schneider, J., Drewnick, F., Borrmann, S., Weimer, S., Demerjian, K., Williams, P., Bower, K., Bahreini, R., Cottrell, L., Griffin, R. J., Rautiainen, J., Sun, J. Y., Zhang, Y. M., and Worsnop, D. R.: Ubiquity and dominance of oxygenated species in organic aerosols in anthropogenically-influenced Northern Hemisphere midlatitudes, Geophys. Res. Lett., 34, L13801, doi:10.1029/2007GL029979, 2007. 33632

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, doi:10.5194/acp-8-4559-2008, 2008. 33633

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, doi:10.5194/acp-11-9155-2011, 2011. 33633, 33644

Table 1. Substructures matched in order to account for the complete set of carbons and oxygen atoms in the set of compounds constituting the *α*-pinene and 1,3,5-trimethylbenzene degradation scheme in MCM v3.2 (substructures 1–33) and extra molecular substructures measurable with FTIR (substructures 34–57). The SMARTS pattern corresponding to the FG aldehyde has been specified in order to explicitly match the hydrogen attached to the sp^2 carbon. In order to not double count the aldehydic carbonyl in case of formaldehyde, a different SMARTS pattern, specific for formaldehyde, has been formulated (substructure 15). In the case of the SMARTS pattern specified for counting aldehyde FGs in the SIMPOL.1 method, the hydrogen is not explicitly matched, and for this reason this problem is avoided. A SMARTS pattern specific for formic acid has been specified in order to explicitly match the hydrogen atom attached to the carbon of the carboxylic FG. The formulation of this extra SMARTS pattern is necessary to avoid explicitly counting an eventual carbon attached to the carboxylic carbon. For space constraints the SMARTS patterns have been reported on multiple lines, even if the SMARTS notation requires unique lines.

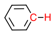
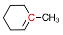
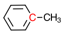
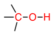
No.	Substructure	Definition	Chemoinformatic definition	Matched pattern
1	Quaternary carbon	A carbon atom bonded to four carbon atoms.[1]	`[$([C]([#6])([#6])([#6])[#6])]`	
2	Alkane CH	Hydrogen atom attached to a sp3 carbon atom.	`[CX4][H]`	
3	Alkene CH	Hydrogen atom attached to a non aromatic sp^2 carbon atom.	`[CX3;$(C=C)][H]`	
4	Aromatic CH	Hydrogen atom attached to an aromatic sp^2 carbon atom.	`[c][H]`	
5	C sp^2 non quaternary	A non aromatic sp^2 carbon atom bonded to three carbons.	`[CX3;$([C]([#6])(=[#6])[C])]`	
6	C sp^2 aromatic non quaternary	An aromatic sp^2 carbon atom bonded to three carbon atoms.	`[c;$([c](c)(c)[C])]`	
7	Alcohol OH	A compound containing an -OH (hydroxyl) group bonded to a tetrahedralcarbon atom.[1]	`[C;!$(C=O)][OX2H][H]`	

Table 1. Continued.

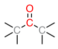
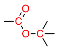
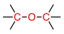
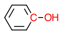
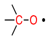
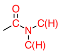
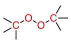
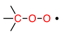
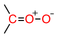
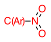
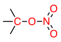
No.	Substructure	Definition	Chemoinformatic definition	Matched pattern
8	Ketone	A compound containing a carbonyl group bonded to two carbon atoms.[1]	`[CX3;$(C([#6])(=[O])[#6])](=[O;!$([O][O])]))`	
9	Aldehyde	A compound containing a -CHO group.[1] (excludes formaldehyde)	`[CX3;$(C([#1])(=[O])[#6])](=[O;!$([O][O])])[H]`	
10	Carboxylic acid	A compound containing a carboxyl, -COOH, group.[1] (excludes formic acid)	`[CX3;!$([CX3][H])](=O)[OX2H][H]`	
11	Formic acid	Formic acid compound.	`[CX3](=O)([H])[OX2H][H]`	
12	Acyloxy radical	Oxygen-centered radicals consisting of an acyl radical bonded to an oxygen atom.[2]	`[C;$(C=O)](=O)[OX2;!$([OX2][H]);!$([OX2][O]);!$([OX2][N]);!$([OX2]([#6])[#6])]`	
13	Ester	A derivative of a carboxylic acid in which H of the carboxyl group is replaced by a carbon.[1]	`[CX3H1,CX3](=O)[OX2H0][#6;!$([C]=[O])]`	
14	Ether	An -OR group, where R is an alkyl group.[1]	`[OD2]([#6;!$(C=O)])[#6;!$(C=O)]`	
15	Formaldehyde	Formaldehyde compound.	`[CX3;$(C(=[O])([#1])[#1])](=[O;!$([O][O])])([H])[H]`	
16	Phenol OH	Compounds having one or more hydroxy groups attached to a benzene or other arene ring.[2]	`[c;!$(C=O)][OX2H][H]`	
17	Oxy radical (alkoxy)	Oxygen centered radical consisting of an oxygen bonded to an alkyl.	`[#6;!$(C=O)][OX2;!$([OX2][H]);!$([OX2][O]);!$([OX2][N]);!$([OX2]([#6])[#6]);!$([OX2][S])]`	
18	Carboxylic amide (primary, secondary and tertiary)	A derivative of a carboxylic acid in which the -OH is replaced by an amine.[1]	`[CX3](=O)[NX3;!$(N=O)]([#6,#1])[#6,#1]`	
19	Peroxide	Compounds of structure ROOR in which R may be any organyl group.[2]	`[#6][OD2][OD2,OD1][#6]`	
20	Peroxy radical	Oxygen centered radical derived from an hydroperoxyde.	`[O;!$([O][#6]);!$([O][H]);!$([OX2][N]);!$(O=C)][O][#6;!$([C](=O)~OO)]`	

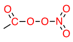
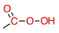
Table 1. Continued.

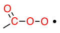
No.	Substructure	Definition	Chemoinformatic definition	Matched pattern

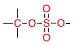
21	C=O$^+$-O$^-$ group	Group of the type C=O$^+$-O$^-$	`[O;!$([O][#6]);!$([O][H]);`
			`!$([OX2][N]);!$(O=C)]`
			`[O]=[#6;!$([C](=O)~OO)]`
			`([#6,#1])[#6,#1]`
				

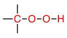
22	C-nitro	Compounds having the nitrogroup, -NO$_2$ (free valence on nitrogen), which is attached to a carbon.[2]	`[#6][$([NX3](=O)=O),`
			`$([NX3+](=O)[O-])](~[O])`
			`(~[O])`
				

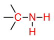
23	Organonitrate	Compounds having the nitrogroup, -NO$_2$ (free valence on nitrogen), which is attached to an oxygen.[2]	`[#6][O][$([NX3](=[OX1])`
			`(=[OX1])O),$([NX3+]([OX1-])`
			`(=[OX1])O)](~[O])(~[O])`
				

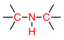
24	Peroxyacyl nitrate	Functional group containing a -COOONO$_2$.	`[C](=O)OO[N](~O)~[O]`	
25	Peroxy acid	Acids in which an acidic -OH group has been replaced by an -OOH group.[2]	`C(=O)O[O][H]`	

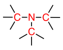
26	Acylperoxy radical	Oxygen centered radical derived from a peroxy acid.	`C(=O)O[O;!$([O][H]);`
			`!$([OX2][N])]`
				

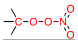
27	Organosulfate	Esters compounds derived from alcohol and sulfuric acids functional groups.	`[#6][O][SX4;`
			`$([SX4](=O)(=O)(O)O),`
			`$([SX4+2]([O-])([O-])(O)O)]`
			`(~[O])(~[O])(~[O])`
				

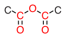
28	Hydroperoxide	A compound containing an -OOH group.[1]	`[#6;!$(C=O)][OD2]`
			`[OX2H,OD1][#1]`
				

29	Primary amine	An amine in which nitrogen is bonded to one carbon and two hydrogens.[1]	`[#6][NX3;H2;!$(NC=O)]`
			`([H])[H]`
				

30	Secondary amine	An amine in which nitrogen is bonded to two carbons and one hydrogen.[1]	`[#6][NX3;H;!$(NC=O)]`
			`([#6])[H]`
				

31	Tertiary amine	An amine in which nitrogen is bonded to three carbons.[1]	`[#6][NX3;H0;!$(NC=O);`
			`!$(N=O)]([#6])[#6]`
				

32	Peroxy nitrate	Functional group containing a COONO$_2$.	`[#6][O;!$(OOC(=O))]`
			`[O;!$(OOC(=O))][N](~O)~[O]`
				

33	Anhydride	Two acyl groups bonded to an oxygen atom.[1]	`[CX3](=O)[O][CX3](=O)`	

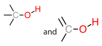
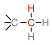
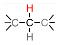
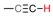
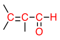
34	Alcohol O-H and Phenol O-H	Alcohol and phenol O-H.	`[OX2H;$([O]([#6])[H]);`
			`!$([O](C=O)[H])][H]`
				

Table 1. Continued.

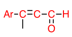
No.	Substructure	Definition	Chemoinformatic definition	Matched pattern
35	Alkane C-H in -CH$_3$	C-H bonds in CH$_3$ group.	`[CX4;$(C([H])([H])[H])][H]`	

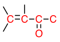
36	Alkane C-H in -CH$_2$	C-H bonds in CH$_2$ group.	`[CX4;$(C([H])([H])`
			`([!#1])[!#1])][H]`
				

37	Alkynes C-H	Hydrogen bonded to a sp carbon in an alkyne group.	`[C;$(C#C)][H]`	
38	Alkynes C≡C	Two carbons that are triple bonded.	`[C]#[C]`	
39	Aromatic C=C	Two aromatic carbons bonded with an aromatic bond.	`c:c`	

40	Conjugated aldehyde C=O and α,β C=C	An aldehyde C=O conjugated with an alkene C=C in α and β positions.	`[CX3;$(C(=[O])([#1])[C]=[C])]`
			`([C]=[C;!$(Cc)])`
			`(=[O;!$([O][O])])[H]`
				

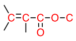
41	Conjugated aldehyde C=O and phenyl	An aldehyde C=O conjugated with a phenyl group.	`[CX3;$(C(=[O])([#1])`
			`[c;$(c1cc[c]cc1)])]([#6,#1])`
			`(=[O;!$([O][O])])[H]`	

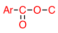
42	Conjugated aldehyde C=O and α,β C=C and phenyl	An aldehyde C=O conjugated with alkene C=C in α and β positions and a phenyl group.	`[CX3;$(C(=[O])([#1])[C]=[C]`
			`[c;$(c1cc[c]cc1)])]`
			`([C])(=[O;!$([O][O])])[H]`
				

43	Conjugated ketone C=O and α,β C=C	A ketone C=O conjugated with an alkene C=C in α and β positions.	`[CX3;$(C([#6])(=[O])`
			`[C]=[C])]([C])`
			`(=[O;!$([O][O])])[C]`
				

44	Conjugated ketone C=O and phenyl	A ketone C=O conjugated with a phenyl group.	`[CX3;$(C([C])(=[O])`
			`[c;$(c1cc[c]cc1)])]([C])`
			`(=[O;!$([O][O])])[c]`	

45	Conjugated ketone C=O and two phenyl	A ketone C=O conjugated with two phenyl groups.	`[CX3;$(C([c,$(c1cc[c]cc1)])`
			`(=[O])[c;$(c1cc[c]cc1)])]`
			`([c])(=[O;!$([O][O])])[c]`	

46	Conjugated ester C=O and α,β C=C	An ester C=O conjugated with alkene C=C in α and β positions .	`[C;!$(Cc)]=[C]`
			`[CX3;$([C]([O][C])`
			`(=[O])[C]=[C])]([O][C])`
			`(=[O;!$([O][O])])`
				

47	Conjugated ester C=O and phenyl	A ester C=O conjugated with a phenyl group.	`[CX3;$([C]([O][C])(=[O])`
			`[c,$(c1cc[c]cc1)])]([O][C])`
			`(=[O;!$([O][O])])`
				

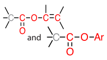
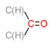
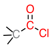
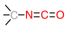
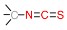
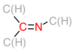
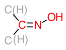
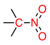
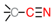
48	Conjugated ester C-O with C=C or phenyl	An ester C=O conjugated with alkene C=C in α and β positions and a phenyl group.	`[CX3;$([C]([#6])(=[O])[O]`
			`[C]=[C]),$([C]([#6])(=[O])`
			`[O][c;$(c1cc[c]cc1)])]`
			`(=[O;!$([O][O])])[O]`
			`[#6;$(C=C),$(c1cc[c]cc1)]`
				

Table 1. Continued.

No.	Substructure	Definition	Chemoinformatic definition	Matched pattern
49	Nonacid carbonyl	Carbonyl group in ketones and aldehydes.	[CX3;$(C([#6,#1])(=[O]) [#6,#1])](=[O;!$([O][O])])	
50	Acyl Chloride	An acyl group bonded to a chloride atom.	[C,$([C]([#6])(=[O]))] (=O)[Cl]	
51	Isocyanate	An -N=C=O group.	[N;$([N]([#6])=[C]=[O])] =[C]=[O]	
52	Isothiocyanate	An -N=C=S group.	[N;$([N]([#6])=[C]=[S])] =[C]=[S]	
53	Imine	A carbon- nitrogen double bond, R_2C=NR.	[C;$(C([#6,#1])([#6,#1])=[N])] =[N][#1,#6]	
54	Oxime	A carbon-nitrogen double bond, R_2C=NOH.	[C;$(C([#6,#1])([#6,#1]) =[N][O][H])]=[N][O][H]	
55	Aliphatic nitro	Compounds having the nitro group, -NO_2 (free valence on nitrogen), which is attached to an alifatic carbon.	[C][$([NX3](=O)=O), $([NX3+](=O)[O-])] (~[O])(~[O])	
56	Aromatic nitro	Compounds having the nitro group, -NO_2 (free valence on nitrogen), which is attached to an aromatic carbon.	[c][$([NX3](=O)=O), $([NX3+](=O)[O-])] (~[O])(~[O])	
57	Nitrile	A carbon atom bonded to a nitrogen atom with a triple bond.	[C;$([C]#[N])]#[N]	

[1] Brown et al. (2012).
[2] IUPAC Compendium of Chemical Terminology – the Gold Book (Nic et al., 2014).

Table 2. Chemical substructures required by SIMPOL.1 model (Pankow and Asher, 2008). *K* in the table corresponds to *k* in Pankow and Asher (2008), Table 5. For the calculation of the ester (SIMPOL.1), the generic ester specified in Table 1 (substructure 13) is specified. The group named "Carbon number on the OH side of an amide" is used in the calculation of the "carbon number on the acid side of an amide" but is not present in the SIMPOL.1 groups indicated by Pankow and Asher (2008).

Groups	Chemoinformatic definition or reference to Table 1	*k*
Carbon number	[#6]	1
Carbon number on the acid side of an amide[a,b]	Carbon number- Carbon number on the OH side of an amide-1 if (Amide, primary+Amide, secondary +Amide, tertiary> 0) else 0	2
Aromatic ring[c]	count_aromatic_rings(molecule)	3
Non aromatic ring[c]	count_nonaromatic_rings(molecule)	4
C=C (non aromatic)	C=C	5
C=C-C=O in non-aromatic ring	[$(C=CC=O);A;R]	6
Hydroxyl (alkyl)	Table 1, number 7	7
Aldehyde	[CX3;$(C([#1])(=[O])[#6,#1])](=[O;!$([O][O])])	8
Ketone	Table 1, number 8	9
Carboxylic acid	[CX3](=O)[OX2H][H]	10
Ester (SIMPOL.1)[b]	Ester-Nitroester	11
Ether (SIMPOL.1)	[OD2]([C;!R;!$(C=O)])[C;!R;!$(C=O)]	12
Ether, alicyclic	[OD2;R]([C;!$(C=O);R])[C;!$(C=O);R]	13
Ether, aromatic	c~[O,o]~[c,C&!$(C=O)]	14
Nitrate	Table 1, number 23	15
Nitro	Table 1, number 22	16
Aromatic hydroxyl (e.g. phenol)	Table 1, number 16	17
Amine, primary	[C][NX3;H2;!$(NC=O)]([H])[H]	18
Amine, secondary	[C][NX3;H;!$(NC=O)]([C])[H]	19

Table 2. Continued.

Groups	Chemoinformatic definition or reference to Table 1	*k*
Amine, tertiary	`[C][NX3;H0;!$(NC=O);!$(N=O)]([C])[C]`	20
Amine, aromatic	`[N;!$(NC=O);!$(N=O);$(Na)]`	21
Amide, primary	`[CX3;$(C(=[O])[NX3;!$(N=O)])](=[O])[N]([#1])[#1]`	22
Amide, secondary	`[CX3;$(C(=[O])[NX3;!$(N=O)]([#6])[#1])](=[O])[N][#1]`	23
Amide, tertiary	`[CX3;$(C(=[O])[NX3;!$(N=O)]([#6])[#6])](=[O])[N]`	24
Carbonylperoxynitrate	Table 1, number 24	25
Peroxide	Table 1, number 19	26
Hydroperoxide	Table 1, number 28	27
Carbonylperoxyacid	Table 1, number 25	28
Nitrophenol[c]	`count_nitrophenols(molecule,’Phenol,’Nitro)`	29
Nitroester[a]	`[#6][OX2H0][CX3,CX3H1](=O)[C;$(C[N]([O])~[O]),$(CC[N]([O])~[O]),$(CCC[N]([O])~[O]),$(CCCC[N]([O])~[O]),$(CCCCC[N]([O])~[O])]`	30
Carbon number on the OH side of an amide	`[C;$(C[NX3][CH,CC](=O)),$(CC[NX3][CH,CC](=O)),$(CCC[NX3][CH,CC](=O)),$(CCCC[NX3][CH,CC](=O)),$(CCCCC[NX3[CH,CC](=O))]`	

[a] In the case of the calculations of the number of carbons on the acid side of an amide and for nitroester is this table, these patterns provide correct counting for compounds with a maximum of 5 carbon atoms on the acid side of an amide or in between the ester and the nitro group respectively. To match cases with higher number of carbon atoms, it is necessary to repeat the specified pattern with an augmented number of carbons specified in the code.

[b] Quantities are calculated from other groups; the code shown is executable string formatting syntax of the Python programming language. Entries in braces `{}` are replaced by the number of matched groups designated by name.

[c] User-defined functions which access additional molecular structure information for ring structures. `molecule` is a reserved name indicating an object of the Molecule class defined by the pybel library for our implementation, and entries in quoted braces `'{}` passed as arguments correspond to the matched substructure prior to enumeration. These functions are provided as part of the companion program (Appendix C). This functional interface abstracts the calculation such that the patterns above can be used with any chemoinformatic software package provided that the implementation of ring enumeration functions are changed accordingly.

Table 3. Absorption bands in the infrared region of different FGs and the correspondence in Table 1.

No.	Functional group and functional groups pattern	Wavenumber (cm^{-1})
2, 35, 36	Alkane C-H	2900 (C-H stretch), 1450 and 1375 (bend in CH$_3$), 1465 (bend in CH$_2$)
3	Alkene C-H	3100 (C-H stretch), 720 (Bend, rocking), 100–650 (Out of plane bend)
37	Alkyne C-H	3300 (Stretch)
4	Aromatic C-H	3000 (C-H stretch), 900–690 (Out of plane bend)
38	Alkyne C≡C	2150 (CC stretch)
39	Aromatic C=C	1600 and 1475 (Stretch)
7, 16, 34	Alcohol and phenol	3400 (O-H stretch), 1440–1220 (C-O-H bend), 1260–1000 (C-O stretch)
10, 11	Carboxylic acid COOH	3400–2400 (O-H stretch), 1730–1700 (C=O stretch), 1320–1210 (stretch)
8, 9, 15, 49	Aldehyde and ketone	1740 (aldehyde C=O stretch), 1720–1708 (ketone C=O stretch), 1300–1100 (ketone C(C=O)C bend), 2860–2800 and 2760–1200 (aldehyde C-H stretch)
29, 30, 31	Amines	1640–1560 (N-H bend, in primary amines), 3500–3300 (secondary and primary amines N-H stretch), 1500 (secondary amines N-H bend), 800 (secondary and primary amines N-H out of plane bend), 1350–1000 (C-N stretch)
14	Ether	1300–1000 (C-O stretch)
13	Ester	1750–1735 (C=O stretch), 1300–1000 (C-O stretch)

Table 3. Continued.

No.	Functional group and functional groups pattern	Wavenumber (cm^{-1})
18, (SIMPOL.1 groups)	Amide	1680–1630 (C=O stretch), 3350 and 3180 (primary amide N-H stretch), 3300 (secondary amide N-H stretch), 1640–1550 (primary and secondary amide N-H bend)
27	Organosulfate	876 (C-O-S stretch)
23	Organonitrate	1280 (symmetric NO_2 stretch)
50	Acid Chloride	1850–1775 (C=O stretch), 730–550 (C-Cl stretch)
22, 55, 56	Nitro	1600–1640 (aliphatic nitro -NO_2 asymmetric stretch), 1390–1315 (aliphatic nitro -NO_2 symmetric stretch), 1550–1490 (aromatic nitro -NO_2 asymmetric stretch), 1355–1315 (aromatic nitro -NO_2 symmetric stretch)
57	Nitrile	2250 (stretch, if conjugated 1780–1760)
51	Isocyanate	2270 (stretch)
52	Isothiocyanate	2125 (stretch)
53	Imine	1690–1640 (stretch)
33	Anhydride	1830–1800 (C=O stretch), 1775–1740 (C-O stretch)
40, 41, 42	Conjugated aldehyde	1700–1680 and 1640 (conjugated aldehyde C=O with C=C in α and β), 1700–1660 and 1600–1450 (conjugated aldehyde C=O with phenyl), 1680 (conjugated aldehyde C=O with C=C and phenyl)
43, 44, 45	Conjugated ketone	1700–1675 and 1644–1617 (conjugated ketone C=O and α,β C=C), 1700–1680 and 1600–1450 (conjugated ketone C=O with phenyl), 1670–1600 (conjugated ketone and two phenyl)
46, 47, 48	Conjugated ester	1740–1715 and 1640–1625 (conjugated ester C=O and α, β C=C), 1740–1715 and 1600–1450 (conjugated ester C=O and phenyl), 1765–1762 (conjugated ester C-O with C=C or phenyl)

Table 4. List of SMARTS patterns and coefficients associated with each bond type, used to calculate the carbon oxidation state as described in the Sect. 2.

Bond	SMARTS pattern	Coefficient
C-H	`[#6][H]`	−1
C-C	`[#6]-[#6]`	0
C=C	`[#6]=[#6]`	0
C≡C	`[#6]#[#6]`	0
C-O	`[#6]-[#8]`	1
C=O	`[#6]=[#8]`	2
C-N	`[#6]-[#7]`	1
C=N	`[#6]=[#7]`	2
C≡N	`[#6]#[#7]`	2
C-S	`[#6]-[#16]`	1
C=S	`[#6]=[#16]`	2
C≡S	`[#6]#[#16]`	3

Table B1. List of the compounds used to test the chemoinformatic patterns used in the SIMPOL.1 (Pankow and Asher, 2008) group contribution method to calculate pure component vapor pressure (Table 2).

Compound or MCMv3.2 internal name	Smiles
2,2-dimethyl pentane	CCCC(C) (C)C
1,1-dimethyl cyclohexane	CC1(CCCCC1)C
cyclobutanol	C1CC(C1)O
1,2-pentanediol	CCCC(CO)O
butanal	CCCC=O
2-octanone	CCCCCCC(=O)C
heptanal	CCCCCCC=O
ethanoic acid	CC(=O)O
butanoic acid	CCCC(=O)O
4-oxo-pentanoic acid	CC(=O)CCC(=O)O
2,4-hexadienal	C/C=C/C=C/C=O
3-butenoic-acid	C=CCC(=O)O
2-phenyl-propane	CC(C)C1=CC=CC=C1
2-phenyl-ethanol	C1=CC=C(C=C1)CCO
2-hydroxy-1-methyl-benzene	CC1=CC=CC=C1O
3-methyl-benzoic acid	CC1=CC(=CC=C1)C(=O)O
formamide	C(=O)N
dimethyl-acetamide	CC(C)C(=O)N
N,N-Dimethylacetamide	CC(=O)N(C)C
2-propylamine	CC(C)N
2-butylamine	CCC(C)N
4-amino-3-methylbenzoic acid	CC1=C(C=CC(=C1)C(=O)O)N
1-butoxy-2-ethoxyethane	O(CCCC)CCOCC
cis-2,4-dimethyl-1,3-dioxane	C[C@H]1OCC[C@@H](C)O1
3-methylbutyl nitrate	CC(C)CCO[N+](=O)[O-]
2-methyl-propyl ethanoate	CC(C)COC(=O)C
1-methyl-propyl butanoate	O=C(OC(CC)C)CCC
2-nitro-1-propanol	CC(CO)[N+](=O)[O-]
ethyl nitroacetate	CCOC(=O)C[N+](=O)[O-]
di-n-butyl peroxide	CC(C) (C)OOC(C) (C)C
peroxyacetylnitrate	CC(=O)OO[N+](=O)[O-]
ethyl-hydroperoxide	CCOO
butyl-hydroperoxide	CCCCOO
butanedioic acid	C(CC(=O)O)C(=O)O
methylbutanedioic acid	CC(CC(=O)O)C(=O)O
benzoic acid	C1=CC=C(C=C1)C(=O)O
1,3,5-benzenetricarboxylic acid	C1=C(C=C(C=C1C(=O)O)C(=O)O)C(=O)O
1,2,4,5-benzenetetracarboxylic acid	C1=C(C(=CC(=C1C(=O)O)C(=O)O)C(=O)O)C(=O)O
2,6-naphthalenedicarboxylic acid	C1=CC2=C(C=CC(=C2)C(=O)O)C=C1C(=O)O
dehydroabietic acid	CC(C)C1=CC2=C(C=C1)[C@]3(CCC[C@@]([C@@H]3CC2) (C)C(=O)O)C
dinitrophenol	C1=CC(=C(C(=C1)O)[N+](=O)[O-])[N+](=O)[O-]
perylene	C1=CC2=C3C(=C1)C4=CC=CC5=C4C(=CC=C5)C3=CC=C2
benzo[ghi]perylene	C1=CC2=C3C(=C1)C4=CC=CC5=C4C6=C(C=C5)C=CC(=C36)C=C2
benzo[ghi]fluoranthene	C1=CC2=C3C(=C1)C4=CC=CC5=C4C3=C(C=C2)C=C5
anthracene-9,10-dione	C1=CC=C2C(=C1)C(=O)C3=CC=CC=C3C2=O
n-pentacontane	C(CCCCCCCCCCCCCCCCCCCCCC)CCCCCCCCCCCCCCCCCCCCCCCCCCC
trans-2-butene	C/C=C/C
peroxyacetyl nitrate	CC(=O)OO[N+](=O)[O-]

Table B1. Continued.

Compound or MCMv3.2 internal name	Smiles
acetone	CC(=O)C
glyoxal	C(=O)C=O
crotonaldehyde	C/C=C/C=O
cyclohexanone	C1CCC(=O)CC1
cyclohex-2-eneone	C1CC=CC(=O)C1
1-(4-methyl-phenyl)-ethanone	Cc1ccc(cc1)C(=O)C
1-phenyl-1-butanone	CCCC(=O)c1ccccc1
cyclohexane	C1CCCCC1
1,1-dimethyl cyclopentane	CC1(CCCC1)C
3-ethyl-phenol	CCc1cccc(c1)O
p-hydroxybiphenyl	C1=CC=C(C=C1)C2=CC=C(C=C2)O
cis-2-butene-1,4-diol	C(/C=C/CO)O
oct-2-en-4-ol	OC(/C=C/C)CCCC
1,7-heptanediol	C(CCCO)CCCO
pinic acid	CC1(C(CC1C(=O)O)CC(=O)O)C
norpinic acid	CC1(C(CC1C(=O)O)C(=O)O)C
octadeca-9-enoic acid	CCCCCCCC/C=C/CCCCCCCC(=O)O
pentamethyl benzoic acid	Cc1c(c(c(c(c1C)C)C(=O)O)C)C
heptanamide	CCCCCCC(=O)N
diethyl-butanamide	CCC(CC) (CC)C(=O)N
n-ethyl-n-phenylamine	CCNc1ccccc1
triethanolamine	C(CO)N(CCO)CCO
methyl dimethoxyethanoate	COC(C(=O)OC)OC
methyl benzoate	COC(=O)c1ccccc1
2-methyl-propyl benzoate	CC(C)COC(=O)c1ccccc1
1,3-dioxolan	C1COCO1
2-phenyl-1,3-dioxolane	c1ccc(cc1)C2OCCO2
2,4-dimethoxybenzoic acid	COc1ccc(c(c1)OC)C(=O)O
phenylmethyl nitrate	C1=CC=C(C=C1)CO[N+](=O)[O-]
2,4-dinitrophenol	c1cc(c(cc1[N+](=O)[O-])[N+](=O)[O-])O
4-nitrophenol	c1cc(ccc1[N+](=O)[O-])O
2-methyl-6-nitrobenzoic acid	Cc1cccc(c1C(=O)O)[N+](=O)[O-]
di-(1-methyl-propyl) peroxide	CCC(C)OOC(C)CC
ethylbutanamide	CCCC(=O)NCC
C811CO3	[O]OC(=O)CC1CC(C(=O)O)C1(C)C
APINBOO	[O-][O+]=CCC1CC(C(=O)C)C1(C)C
C106O2	O=CCC(=O)CC(C(=O)C)C(C) (C)O[O]
C721O	OC(=O)C1CC([O])C1(C)C
2,2-Dimethylpropaneperoxoic acid	OOC(=O)C(C) (C)C
APINCO	CC1=CCC(CC1O)C(C) (C)[O]
C89CO2	O=CCC1CC(C(=O)[O])C1(C)C
C10PAN2	O=N(=O)OOC(=O)CC1CC(C(=O)C)C1(C)C
Pinanol	O=N(=O)OC1(C)C(O)CC2CC1C2(C)C
C811CO3H	OOC(=O)CC1CC(C(=O)O)C1(C)C
C106OOH	O=CCC(=O)CC(C(=O)C)C(C) (C)OO
Ethyl sulfate	CCOS(=O) (=O)O
Toluene	Cc1ccccc1
Nitroperoxymethane	COON(=O)=O
Diethylamine	CCNCC
Dimethylamine	CNC

Table B2. List of compounds used to test the substructures 34–57 in Table 1.

Compound name	Smiles
propane	CCC
pentyne	CCCC#C
benzene	c1ccccc1
pentenal	CC/C=C/C=O
benzaldehyde	c1ccc(cc1)C=O
cinnamaldehyde	c1ccc(cc1)C=CC=O
mesityloxide	CC(=CC(=O)C)C
acetophenone	CC(=O)c1ccccc1
benzophenone	c1ccc(cc1)C(=O)c2ccccc2
cyclopentanone	C1CCC(=O)C1
biacetyl	CC(=O)C(=O)C
pentadione	CC(=O)CC(=O)C
methylmethacrylate	CC(=C)C(=O)OC
methylbenzoate	COC(=O)c1ccccc1
vinylacetate	CC(=O)OC=C
butyrolactone	C1CC(=O)OC1
ethanoic anhydride	CC(=O)OC(=O)C
acetyl chloride	CC(=O)Cl
propionitrile	CCC#N
methyl isocyanate	CN=C=O
methyl isothiocyanate	CN=C=S
ethanimine	CC=N
acetone oxime	CC(=NO)C
nitrobenzene	c1ccc(cc1)[N+](=O)[O-]
nitropropane	CCC[N+](=O)[O-]

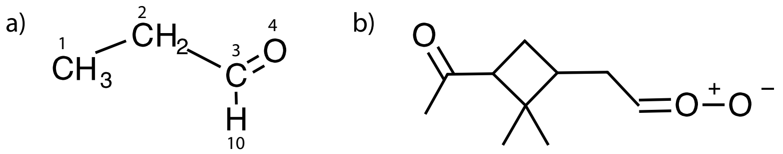


Figure 1. Propionaldehyde (**a**, SMILES code CCC=O) and compound named APINOOB in MCMv3.2 scheme (**b**, SMILES code [O-][O+]=CCC1CC(C(=O)C)C1(C)C). The carbon and oxygen atoms are enumerated, together with the hydrogen of the aldehyde group in compound **(a)**.

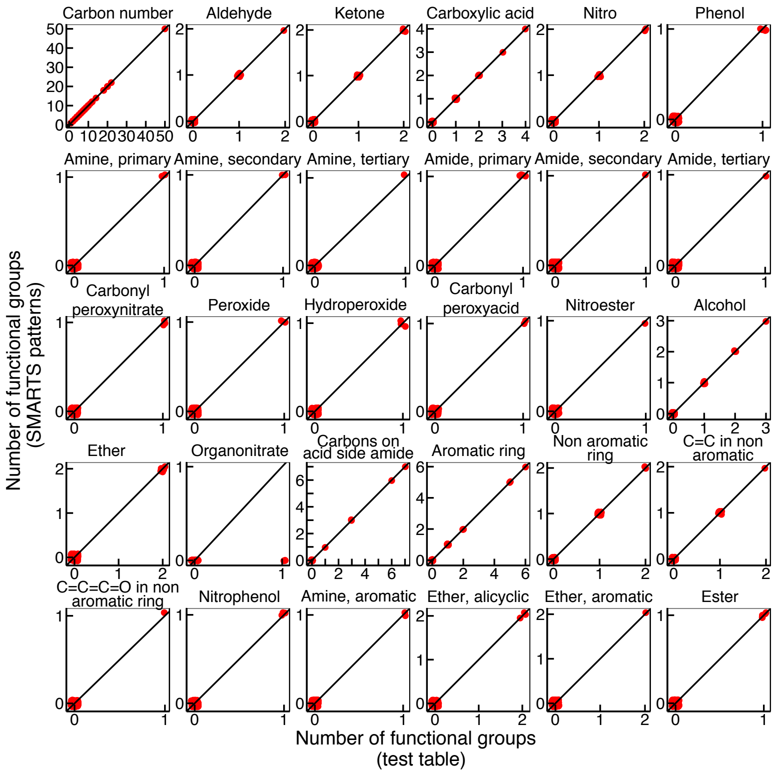


Figure 2. Validation of the developed chemoinformatic patterns for the chemical substructures required in the SIMPOL.1 model (Pankow and Asher, 2008). This validation set includes 99 compounds as described in Sect. 2.

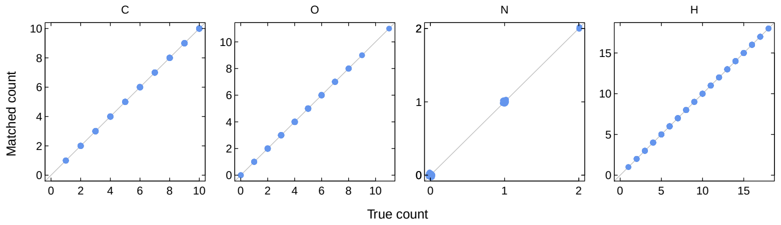


Figure 3. Test of the completeness of matching of all the atoms in the α-pinene and 1,3,5-trimethylbenzene degradation scheme in MCMv3.2 by the SMARTS patterns in Table 1, substructures 1–33.

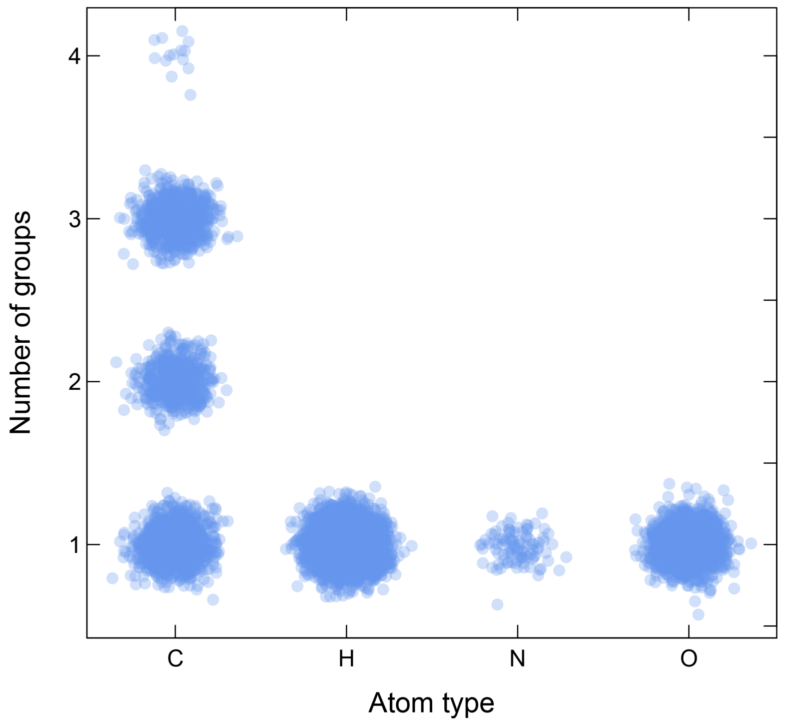


Figure 4. Test for the uniqueness of matching for each atom. Number of times a specific atom has been matched, in the *α*-pinene and 1,3,5-trimethylbenzene degradation scheme in MCMv3.2 by the SMARTS patterns in Table 1, substructures 1–33. Oxygen, nitrogen and hydrogen atoms are matched only once. The carbon atoms are matched multiple times when multifunctional.

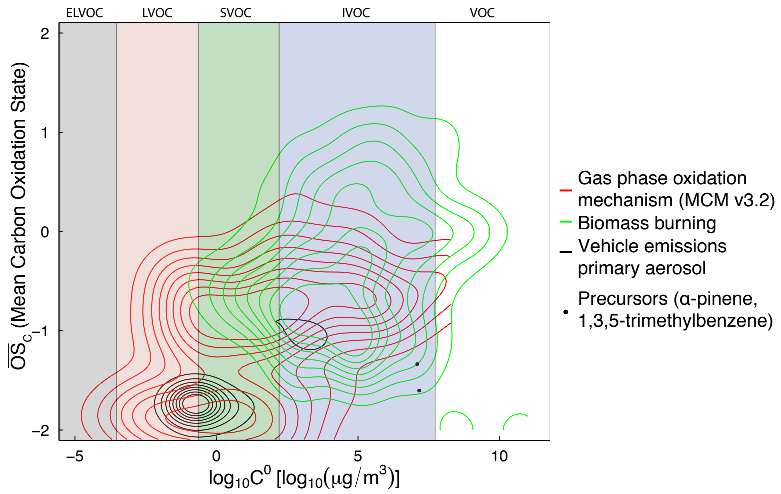


Figure 5. Logarithm of the pure component saturation concentration ($\log_{10}C^0$) and mean carbon oxidation state of each compound ($\overline{\mathrm{OS}}_\mathrm{C}$) measured by Rogge et al. (1993, 1998) for biomass burning and vehicle emissions sources (green and blue lines), and of each molecule constituting the MCMv3.2 gas phase oxidation mechanism of *α*-pinene and 1,3,5-trimethylbenzene. The lines in the plot denote isolines ($0, 0.1, \ldots, 0.9$) of the maximum density estimate for the different compound sets. The black dots indicate the position of *α*-pinene and 1,3,5-trimethylbenzene. The area of the plot is divided in volatility regions according to the classification of Donahue et al. (2012).

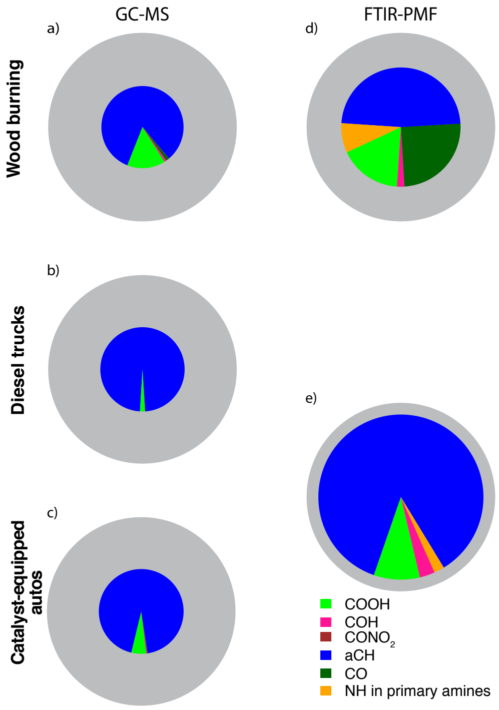


Figure 6. Comparison of the FG distribution of the quantified fraction measured by GC-MS (**a–c**; Rogge et al., 1998, 1993) and FTIR-PMF (**d, e**; Hawkins and Russell, 2010) in aerosol emitted by wood burning **(a, d)** and vehicle emission **(b, c, e)** sources. The grey area is the non-measured OA fraction by the two different analytical techniques used (around 80 % for GC-MS and around 55 and 20 % for FTIR-PMF).

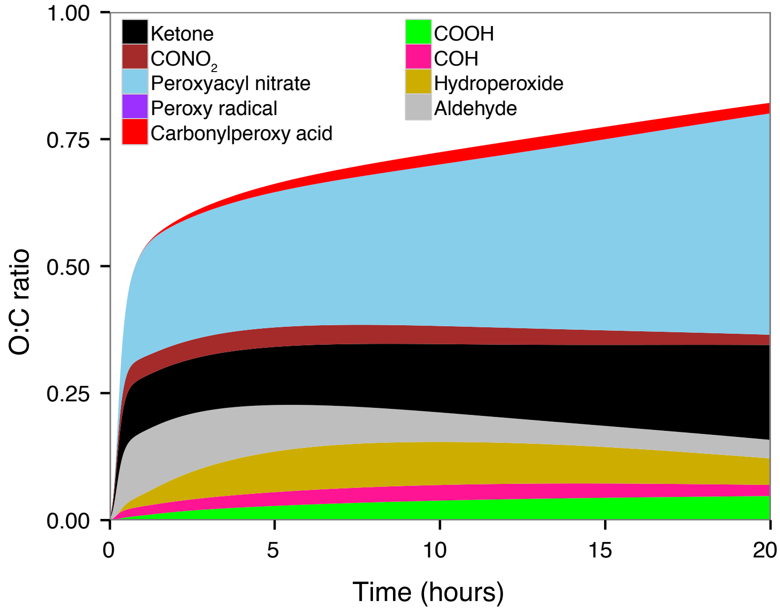


Figure 7. Time series of FG contributions to the total O : C of the gas phase generated by photooxidation of α-pinene in low-NO_x regime, simulated using the MCMv3.2 degradation scheme.