Comments to referees

Giulia Ruggeri

January 2016

We thank both reviewers for their extremely useful comments. Based on a general consensus that the focus of the Technical Note should be on the novelty of the tool itself and less on the example applications, we have made a few general changes:

- Methods (Section 2.1) describing the SMARTS pattern development is expounded

- Example Applications (Section 3.2) is made much more concise

- Conclusion (Section 4) is modified to summarize the chemoinformatic tools, more than interpretation from the specific applications presented

We include detailed responses to reviewer comments as a supplement.

1 Referee #1

1. **Comment** P. 33642: Usually one uses the term low-NOx reaction to mean that organic peroxy radicals have reacted with HO2 or other RO2 radicals, or perhaps by autooxidation. If an organic nitrate contributes 26% of the gas-phase mass of products then this does not seem to qualify. The authors should make clear what they mean by low-NOx

   Response We have omitted the definition of low NO_x in the manuscript and just specified that the α-pinene simulation was carried in the presence of NO_x (initial ratio α-pinene/NO_x is 1.25). In line 25, page 33638 we modified the sentence as follows:

   "Furthermore, the gas phase composition generated by α-pinene photooxidation in the presence of NO_x (α-pinene/NO_x ratio of 1.25), with propene as a radical initiator, was simulated using the Kinetic Pre-Processor (KPP, Damian et al., 2002; Sandu and Sander, 2006; Henderson, 2016) incorporating mechanistic information taken from MCMv3.2."

2. **Comment** My understanding is that the present classification method does not separate positional isomers. Is this correct?

   Response Yes this is the case for most of the chemoinformatic patterns described in this manuscript. Only in the case of conjugated aldehyde ketone and ester with an alkene or a benzene ring the relative position of these two functional groups is specifically described by the SMARTS pattern (Substructures 40-48). We have therefore added a statement in section 2.1:

   "Positional arguments can be included by querying specific structural information from the internal representations of molecular graphs according to implementations in various software packages, or formulating SMARTS patterns which require specificity in the arrangement of neighboring atoms (Barley et al., 2011; Topping et al., 2016). In this work, positional information of FGs are used only for conjugated aldehyde, ketone, and ester with an alkene or benzene ring (Table 1, substructures 40-48)."

3. **Comment** P. 33643: In the Conclusions it would be useful to summarize the value of this approach and clarify its advantages over getting this information other ways.

Response We have restructured the conclusion in order to give more importance to the chemoinformatic patterns developed than to the example application. In this way we hope to better convey the importance of the validation of the chemoinformatic patterns described in the manuscript. The Conclusion (Section 4) has been modified as follows:

"We introduced the application of chemoinformatic tools that allow us to perform substructure matching in molecules to enumerate FGs present in compounds relevant for organic aerosol chemistry. We developed 50+ substructure patterns and validated them over a list of 125 compounds that were selected in order to account for all the functional groups (FGs) represented. We demonstrate how these tools can facilitate intercomparisons between GC-MS and FTIR measurements, and mapping of compounds onto the VBS space described by pure component vapor pressure and oxidation state.

"We further introduce a novel approach for defining a set of patterns which accounts for each atom in a chemical system once and only once (except for polyfunctional carbon atoms associated with multiple FGs); this condition is confirmed by an atomic-level validation scheme applied to chemically explicit α-pinene and 1,3,5-TMB degradation mechanisms. This validation scheme provides an intermediate resolution between molecular speciation and atomic composition, and permits apportionment of conventionally aggregated quantities such as O:C, H:C, and N:C to contributions from individual FGs. We illustrate its application to the photochemical degradation of α-pinene from speciated simulations using MCMv3.2."

2 Referee #2

1. **Comment**: I am unsure whether Atmospheric Chemistry & Physics is the most suitable journal for this manuscript. Further, in its current form, the manuscript needs to be simplified to be an effective Technical Note. I found the diversity of datasets and example applications distracting. For example, in Fig. 6, the authors are comparing functional group distributions derived from GC /MS measurements made during the late 1990's with functional group distributions derived from FTIR measurements in 2010; also included are estimates of measured/unmeasured fractions for each instrument based on additional (and separate) publications. The authors spend one paragraph describing each of the measurements, the dominant functional groups, and the reasons for the discrepancies between the wood burning samples. In the conclusions section, the authors return to this discrepancy between the GC/MS and FTIR measurements for wood smoke, which has little to do with the overall contribution of the Ruggeri and Takahama work. It would be sufficient to demonstrate that functional group distributions can be obtained using their approach. Further comments are provided below on the inclusion of likely extraneous information that diminishes the potential contribution of this work. It is recommended that the authors consider simplifying the applications presented, and focus more discussion on validation.

Response: We propose this manuscript to Atmospheric Chemistry & Physics since its readers are the target audience for the new tools we introduce; as demonstrated in our example applications the tools can be used to analyze both measurements and model simulations. We recently also submitted a companion paper to Atmospheric Chemistry & Physics to further demonstrate its application in the analysis of model simulations of secondary organic aerosol formation, which is a prime topic for this journal. Furthermore we have now restructured the Conclusions and the Section 3.2 in order to give more importance to the chemoinformatic patterns developed and validated in this work.

Section 3.2.2: "From this comparison, we find that the oxidized fraction is much higher in the biomass burning aerosol composition estimated by FTIR. The high abundance of alkane CH bonds in the compounds reported by GC-MS can be explained by the preference of this analytical method to characterize the least oxidized fraction of the collected aerosol. While high abundance of carbonyl groups are reported in FTIR measurements of biomass burning aerosol (Liu et al., 2009; Russell et al., 2009; Hawkins and Russell, 2010), more recent methods including advanced derivatization (Dron et al., 2010) are necessary for quantification of carbonyl containing compounds by GC-MS. In addition, neither amine compounds nor levogucosan were reported in this GC-MS study. Levoglucosan is a polysaccharide

compound often used as a tracer for burning and decomposition of cellulose reported in modern GC-MS measurements (Simoneit, 1999). However, FTIR does not report high fraction of alcohol COH as levoglucosan near particular fuel sources may be found mostly in supermicron diameter particles (Radzi bin Abas et al., 2004) (submicron OA was analyzed by Hawkins and Russell, 2010), its degradation in the atmosphere is rapid (Hennigan et al., 2010; Cubison et al., 2011; Lai et al., 2014), and the overall mass contribution to biomass burning OA is small (less than 2% by mass, Leithead et al., 2006).

"Both estimation methods agree that more than 90% of OM mass is composed of alkane-CH for vehicle sources 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."

Section 4: "We introduced the application of chemoinformatic tools that allow us to perform substructure matching in molecules to enumerate FGs present in compounds relevant for organic aerosol chemistry. We developed 50+ substructure patterns and validated them over a list of 125 compounds that were selected in order to account for all the functional groups (FGs) represented. We demonstrate how these tools can facilitate intercomparisons between GC-MS and FTIR measurements, and mapping of compounds onto the VBS space described by pure component vapor pressure and oxidation state.

"We further introduce a novel approach for defining a set of patterns which accounts for each atom in a chemical system once and only once (except for polyfunctional carbon atoms associated with multiple FGs); this condition is confirmed by an atomic-level validation scheme applied to chemically explicit α-pinene and 1,3,5-TMB degradation mechanisms. This validation scheme permits apportionment of quantities such as O:C, H:C, and N:C to contributions from individual FGs, which are conventionally analyzed as aggregate measures. We illustrate its application to the photochemical degradation of α-pinene from speciated simulations using MCMv3.2."

2. **Comment**: p 33638, line 20-23: The description of the partitioning calculations is confusing. It is not clear what parameters were of interest, or how they were determined. The calculated pure compound vapor pressures could be used to calculate partitioning coefficients, which then could be used (with total OA) to calculate the fraction of each compound in the gas vs. particle phase based on Pankow 1994. Alternatively, calculated particle phase fractions (with total OA) could be used to calculate pure component saturation concentrations based on Donahue et al. 2006. It is suggested that the Greek symbol zeta not be used to define the partitioning coefficient, as that symbol traditionally has been used to define mole-fraction scale activity coefficients in partitioning calculations.

   Response The phrase has been restructured as suggested. The Greek symbol used here is actually xi and not zeta, but as it is difficult to discern it in the font used by the journal we specified that it is a xi after the symbol. The phrase has been restructured as: "After diluting the total OA of a factor of 1000 the compounds were partitioned between the two phases based on the partitioning coefficient ξ_i (xi) calculated from the pure component saturation concentration (C_i^0) as described by Donahue et al. (2006)."

3. **Comment** Page 33639/33640, discussion of Fig. 5: It appears the authors are mapping compounds measured by Rogge et al. in vehicle OA and woodsmoke OA, as well as compounds predicted by MCM. If this is correct, the detailed discussion of PMF/AMS HOA and BBOA in the text does not add anything, and only confuses what is presented in Fig. 5. Additionally, the legend should make it clearer that compounds from woodsmoke POA (and not biomass burning) and vehicle POA are plotted. It does not make sense to me that the compounds in woodsmoke POA span a higher volatility range than the compounds predicted by MCM. Particularly given that MCM is typically unable to predict measured OA loadings without invoking particle phase or aqueous phase chemistry. The authors do

3

reference the abundance of MCM compounds in the IVOC region and absence in the LVOC, as well as particle-phase chemistry. . .maybe the traces are mislabeled and the green trace is MCM?

Response We apologize for the mislabeling of the colors in the legend, which have now been corrected. In replotting, we also discovered an omission of four large, multifunctional compounds with log C^0 of -4 in the set of wood-burning compounds, which have also now been included. We have substantially shortened the discussion of the mapping in Section 3.2 to highlight the main points to be communicated by Figure 5:

"The algorithm described has been used to project molecular composition of GC-MS and MCM compounds to 2D-VBS space delineated by carbon oxidation and pure component saturation concentration (C^0) (Figure 5). The properties of vehicle-related primary OA and wood combustion compounds measured by GC-MS are generally consistent with those reported for hydrocarbon-like OA and biomass burning OA, respectively, derived from PMF analysis of AMS spectra (Donahue et al., 2012). The low oxidation state is observed on account of more than 60% of carbon atoms being associated with methylene groups (-CH$_2$-, oxidation state of -2) in long-chain hydrocarbon compounds, and an association to lesser degree with CH groups in aromatic rings (oxidation state of -1) and methyl groups (-CH$_3$, oxidation state of -3).

"Most compounds in the MCMv3.2 system correspond to intermediate volatility organic compounds (IVOC), with only a small fraction with the semivolatile organic compound (SVOC) regime. When using of MCMv3.2 for simulation of secondary OA formation, additional mechanisms (e.g., in the condensed phase) are necessary to introduce low volatility organic compounds (LVOC) as observed in atmospheric and environmental controlled chamber observations (Ehn et al., 2014; Shiraiwa et al., 2014). Higher oxidation states than for compounds in the GC-MS set are observed on account of the larger number of functional groups containing electronegative atoms (oxygen and nitrogen) bonded to carbon."

4. **Comment** The caption for table 1 is long and contains important information that is not thoroughly discussed in the text. It is recommended that the discussion in the text be expanded to cover the inclusion of special patterns for formaldehyde and formic acid.

Response We added this part in Section 2.1 and rephrased the text as follows:

"Therefore, we provide an example for the aldehyde FG group to illustrate the development process, with particular attention paid to the description of atoms returned in the matched set and how their bonding environments are defined. We first describe a formulation specific for fulfilling the atom-level validation which requires two patterns to account for all aldehyde groups in the system, and an alternate formulation for only enumerating FGs that requires only a single pattern.

"When applied to propionaldehyde, the set of atoms returned by matching the pattern for substructure 9 in Table 1 will be 3, 4, 10 (as labeled in Figure 1a). The first bracket `[CX3;$(C([#1])(=[O])[#6])]` describes the carbon atom to be matched and returned. `CX3` describes a carbon with 3 bonds (effectively sp^2); `$(C([#1])(=[O])[#6])` qualifies that it is bonded to hydrogen, oxygen, and another carbon. The expression `(=[O;!$([O][O])])` describes the double-bonded oxygen to this carbon atom; `!$([O][O])` excludes preventing matching of C=O$^+$-O$^-$ (defined as a separate group, substructure 21 in Table 1) that are present in other molecules (an example is provided in Figure 1b). The last bracket `[H]` is included to explicitly include the hydrogen atom in the returned set. While the sp^3 carbon attached to the sp^2 is not returned in the set of matched atoms, this additional specificity is necessary to prevent double counting of the same aldehydic group in the formaldehyde molecule, which contains two hydrogen atoms bonded to sp^2 carbon. A separate SMARTS pattern is defined for formaldehyde (Table 1 substructure 15). (For similar reasons, a SMARTS pattern specific for formic acid has been specified alongside the carboxylic FG.)

"In this approach, all atoms in the aldehyde group are matched instead of just the identifying carbon, oxygen, or hydrogen. The advantage of this strict protocol is that we can devise a validation such that each atom in a molecule or chemical system is accounted for by one and only one group — except

for polyfunctional carbon — for any proposed set of FGs (Appendix A). Fulfillment of this validation criterion provides a means for interpreting atomic ratios commonly used by the community (e.g., O:C, H:C, and N:C) through contributions of distinctly defined FGs.

"Revisiting the aldehyde FG example, an alternative pattern specified only for the purposes of counting FGs for use in SIMPOL.1 is shown in Table 2. We only describe the bonding environment of the sp^2 carbon and count the number of its occurrence, so a single pattern can be used for both formaldehyde and other aldehyde compounds."

5. **Comment** Table 3 is not referenced in the text. Is it necessary?

   Response We thank the reviewer for pointing out this omission. This table is meant to offer a connection between the FTIR measurements and the functional groups we can harvest from GC-MS and MCM. We have added the reference to it in Section 2.1 as follows: "The regions of absorption in the IR spectrum associated with FGs patterns are reported in Table 3 as an additional reference"

6. **Comment** It is recommended that the authors list the groups of patterns in the same order throughout the manuscript. For example, in the abstract and the introduction, SIMPOL.1 appears as the first group, but in section 2.1 it is the second group.

   Response We have changed the text in order to keep the consistency as suggested.

7. **Comment** p. 33637, lines 15-19: Awkward. It is recommended that the authors reword this description of the test compounds. If I understand, it seems sufficient to say that compounds were selected either from and then give the list (without repeating either or selected).

   Response The phrase has been reworded as suggested and now they keep the same consistency as in the abstract, that has been rephrased as:

   "1) 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, 2) FGs that are measurable by Fourier transform infrared spectroscopy (FTIR), 3) groups incorporated in the SIMPOL.1 vapor pressure estimation model, and 4) bonds necessary for the calculation of carbon oxidation state."

8. **Comment** p. 33639, line 10: Replace *is* with *are* (. . .but are shown. . .).

   Response The text has been modified as suggested: "Matched FTIR FGs in Table 1 (substructures 33-57) are also identical to the true number of FGs in the set of compounds used for evaluation (Table B2), but are not shown as each group except alkane CH is matched at most once and a similar plot is uninformative"

9. **Comment** p. 33642, lines 17-19: The discussion of the contribution of peroxacyl nitrate mass vs. O:C is confusing as written. It is suggested that the authors reword these sentences to more clearly articulate that while the total mass contribution is only 26%, the total O:C contribution is 53% due to the high O:C ratio of these functional groups (or the dominant compound).

   Response The phrase has been rewritten as suggested: "A singular peroxyacyl nitrate compound (peroxyacetyl nitrate) accounts for 26% of the total gas phase mass. The peroxyacyl nitrate functional group furthermore accounts for the greatest fraction of the total O:C ratio after 20 hours of simulation (53% of the total O:C), as it contains five oxygen atoms per FG."

10. **Comment** Fig. 6 caption: Was the OA fraction in gray truly unmeasured? Or was it unresolvable using the technique specified?

    Response The distinction between "measurable" and "resolvable" in this context is not entirely clear, but the gray fraction is the estimated fraction of OA that was not quantified by both techniques. It is possible that with additional derivatization (Dron et al., 2010), for instance, that the GC-MS can resolve or measure a higher fraction of mass. We have changed our terminology to refer to this as the "unresolved" fraction. We have also changed Fig. 6 in order to let the gray area represent the

unresolved fraction by both FTIR and GC-MS, so now the gray area in the second column corresponds to 10% (mass fraction not resolved by FTIR). We have specified the type of biomass burning in the figure and changed the caption:

"Comparison of the FG distribution of the quantified fraction measured by GC-MS (a,b and c; Rogge et al., 1998; Rogge et al., 1993) and FTIR-PMF (d and e; Hawkins and Russell, 2010) in aerosol emitted by biomass burning (a and d) and vehicle emission (b,c and e) sources. The gray area is the unresolved OA fraction by the two different analytical techniques used (around 80% for GC-MS and around 10% for FTIR). The type of biomass burning is specified in the pie charts a and d."

References

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, Atmospheric Chemistry and Physics, 11, 13 145–13 159, doi:10.5194/acp-11-13145-2011, 2011.

Cubison, M. J., Ortega, A. M., Hayes, P. L., Farmer, D. K., Day, D., Lechner, M. J., Brune, W. H., Apel, E., Diskin, G. S., Fisher, J. A., Fuelberg, H. E., Hecobian, A., Knapp, D. J., Mikoviny, T., Riemer, D., Sachse, G. W., Sessions, W., Weber, R. J., Weinheimer, A. J., Wisthaler, A., and Jimenez, J. L.: Effects of aging on organic aerosol from open biomass burning smoke in aircraft and laboratory studies, Atmospheric Chemistry and Physics, 11, 12 049–12 064, doi:10.5194/acp-11-12049-2011, URL http://www.atmos-chem-phys.net/11/12049/2011/, 2011.

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

Donahue, N. M., Robinson, A. L., Stanier, C. O., and Pandis, S. N.: Coupled partitioning, dilution, and chemical aging of semivolatile organics, Environmental Science & Technology, 40, 2635–2643, doi:10.1021/es052297c, 2006.

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, Proceedings of the National Academy of Sciences of the United States of America, 109, 13 503–13 508, doi:10.1073/pnas.1115186109, 2012.

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, Atmospheric Chemistry and Physics, 10, 7041–7055, doi:10.5194/acp-10-7041-2010, 2010.

Ehn, M., Thornton, J. A., Kleist, E., Sipilä, M., Junninen, H., Pullinen, I., Springer, M., Rubach, F., Tillmann, R., Lee, B., Lopez-Hilfiker, F., Andres, S., Acir, I.-H., Rissanen, M., Jokinen, T., Schobesberger, S., Kangasluoma, J., Kontkanen, J., Nieminen, T., Kurtén, T., Nielsen, L. B., Jø gensen, S., Kjaergaard, H. G., Canagaratna, M., Maso, M. D., Berndt, T., Petäjä, T., Wahner, A., Kerminen, V.-M., Kulmala, M., Worsnop, D. R., Wildt, J., and Mentel, T. F.: A large source of low-volatility secondary organic aerosol, Nature, 506, 476–479, doi:10.1038/nature13032, URL http://www.nature.com/nature/journal/v506/n7489/full/nature13032.html, 2014.

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

Henderson, B. H.: Kinetic Pre-Processor with updates to allow working with MCM, doi: 10.5281/zenodo.44682, URL `http://github.com/barronh/kpp`, 2016.

Hennigan, C. J., Sullivan, A. P., Collett, Jeffrey L., J., and Robinson, A. L.: Levoglucosan stability in biomass burning particles exposed to hydroxyl radicals, Geophysical Research Letters, 37, L09 806, doi: 10.1029/2010GL043088, 2010.

Lai, C., Liu, Y., Ma, J., Ma, Q., and He, H.: Degradation kinetics of levoglucosan initiated by hydroxyl radical under different environmental conditions, Atmospheric Environment, 91, 32–39, URL `http://www.sciencedirect.com/science/article/pii/S1352231014002398`, 2014.

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, Atmospheric Environment, 40, 2721–2734, doi:10.1016/j.atmosenv.2005.09.084, 2006.

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, Atmospheric Chemistry and Physics, 9, 6849–6863, doi:10.5194/acp-9-6849-2009, 2009.

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, 1089–95, doi: 10.1016/j.chemosphere.2004.02.002, 2004.

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, Environmental Science & Technology, 27, 636–651, doi:10.1021/es00041a007, 1993.

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, Environmental Science & Technology, 32, 13–22, doi:10.1021/es960930b, 1998.

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, Atmospheric Environment, 43, 6100–6105, doi:10.1016/j.atmosenv.2009.09.036, 2009.

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

Shiraiwa, M., Berkemeier, T., Schilling-Fahnestock, K. A., Seinfeld, J. H., and Pöschl, U.: Molecular corridors and kinetic regimes in the multiphase chemical evolution of secondary organic aerosol, Atmospheric Chemistry and Physics, 14, 8323–8341, doi:10.5194/acp-14-8323-2014, URL `http://www.atmos-chem-phys.net/14/8323/2014/`, 2014.

Simoneit, B. R. T.: A review of biomarker compounds as source indicators and tracers for air pollution, Environmental Science and Pollution Research, 6, 159–169, doi:10.1007/BF02987621, 1999.

Topping, D., Barley, M., Bane, M. K., Higham, N., Aumont, B., Dingle, N., and McFiggans, G.: UManSysProp v1.0: an online and open-source facility for molecular property prediction and atmospheric aerosol calculations, Geoscientific Model Development, 9, 899–914, doi:10.5194/gmd-9-899-2016, URL `http://www.geosci-model-dev.net/9/899/2016/`, 2016.

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

G. Ruggeri[1] and S. Takahama[1]

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

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

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) 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, 2) FGs that are measurable by Fourier transform infrared spectroscopy (FTIR), 3) groups incorporated in the SIMPOL.1 vapor pressure estimation model, 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

1

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

30   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), spectrophotometry (Aimanant and Ziemann, 2013), nuclear magnetic resonance (NMR, Decesari et al., 2000; Cleve-

35   land 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.,

40   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

45   (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

50   be specified by SMARTS, which is a superset of the SMILES specification (DAYLIGHT Chemical Information Systems, Inc.). 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), Indigo (GGA Software Services). The concept of using SMILES and SMARTS patterns have been reported

55   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; Kenny2013) or on web databases — e.g., DAYLIGHT Chemical Information Systems, Inc. (DAYLIGHT Chemical Information Systems, Inc.) — knowledge regarding the

60   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) FGs contained in α-pinene and 1,3,5-trimethylbenzene photooxidation products defined in MCMv3.2 (Jenkin

2

et al., 1997; Saunders et al., 2003; Jenkin et al., 2003; Bloss et al., 2005), obtained via http://mcm.leeds.ac.uk/MCM; 2) FGs that are measured or measurable (i.e., have absorption bands) for
65   FTIR analysis (Pavia et al., 2008); 3) molecular fragments used by SIMPOL.1 for estimation of pure organic compound vapor pressures; and 4) bonds used for calculation of carbon oxidation state ($\overline{OS}_C$) (Kroll et al., 2011, 2015). As there are several ways to define SMARTS patterns for substructure matching, we prescribe a general 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
70   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
75   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 (Section 2.1) as well as the methods and compound sets used for their validation (Section 2.2). We further
80   describe the data set used for constructing a few example applications (Section 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 used to study the
85   FG abundance associated with FTIR measurements (FGs not specified before, containing carbon, oxygen and nitrogen atoms; Table 1, substructures 33-57), the third 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) and the fourth group used to calculate the oxidation state of carbon atoms (Table 3). The regions of absorption in the IR spectrum
90   associated with FGs patterns are reported in Table 4 as an additional reference. 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
95   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

3

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

100   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 (Section C).

105   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 EVAP-ORATION vapor pressure model is fitted to more recent diacid measurements and includes positional information and non-linear interactions among FGs (Compernolle et al., 2011). Positional arguments can be included by querying specific structural information from the internal representa-

110   tions of molecular graphs according to implementations in various software packages, or formulating SMARTS patterns which require specificity in the arrangement of neighboring atoms (Barley et al., 2011; Topping et al., 2016). In this work, positional information of FGs are used only for conjugated aldehyde, ketone, and ester with an alkene or benzene ring (Table 1, substructures 40-48). With regards to the use of SIMPOL.1, vapor pressure predictions can also be improved by updating

115   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. Therefore, we provide an example for the aldehyde FG group to illustrate the development process, with particular attention paid to the description of atoms returned in the matched set and how their bonding environments are defined. We first describe a formulation specific for fulfilling the atom-level vali-

120   dation which requires two patterns to account for all aldehyde groups in the system, and an alternate formulation for only enumerating FGs that requires only a single pattern.

When applied to propionaldehyde, the set of atoms returned by matching the pattern for substructure 9 in Table 1 will be 3, 4, 10 (as labeled in Figure 1a). The first bracket `[CX3;$(C([#1])(=[O])[#6])]` describes the carbon atom to be matched and returned. `CX3`

125   describes a carbon with 3 bonds (effectively sp^2); `$(C([#1])(=[O])[#6])` qualifies that it is bonded to hydrogen, oxygen, and another carbon. The expression `(=[O;!$([O][O])])` describes the double-bonded oxygen to this carbon atom; `!$([O][O])` excludes preventing matching of $C=O^+-O^-$ (defined as a separate group, substructure 21 in Table 1) that are present in other molecules (an example is provided in Figure 1b). The last bracket `[H]` is included to explicitly in-

130   clude the hydrogen atom in the returned set. While the sp^3 carbon attached to the sp^2 is not returned in the set of matched atoms, this additional specificity is necessary to prevent double counting of the same aldehydic group in the formaldehyde molecule, which contains two hydrogen atoms bonded to sp^2 carbon. A separate SMARTS pattern is defined for formaldehyde (Table 1 substructure 15).

4

(For similar reasons, a SMARTS pattern specific for formic acid has been specified alongside the
135   carboxylic FG.)

In this approach, all atoms in the aldehyde group are ~~therefore~~ matched instead of just the identifying carbon, oxygen, or hydrogen. The advantage of this strict protocol is that we can devise a validation such that each atom in a molecule or chemical system is accounted for by one and only one group — except for polyfunctional carbon — for any proposed set of FGs (Appendix A). Ful-
140   fillment of this validation criterion provides a means for interpreting atomic ratios commonly used by the community (e.g., O:C, H:C, and N:C) through contributions of distinctly defined FGs.

Revisiting the aldehyde FG example, an alternative pattern specified only for the purposes of counting FGs for use in SIMPOL.1 is shown in Table 2. We only describe the bonding environment of the sp^2 carbon and count the number of its occurrence, so a single pattern can be used for both
145   formaldehyde and other aldehyde compounds.

A separate set of SMARTS patterns are defined for estimation of $\overline{OS}_C$. Instead of FGs, these patterns enumerate the type of bond and atom attached to a carbon atom, and its oxidation state is calculated as the sum of the coefficients corresponding to its bonds.

2.2 Data sets for validation

150   The first and the third groups of SMARTS patterns were validated against a set of 99 compounds (Table B1, Appendix B) selected from those used in the development of the SIMPOL.1 method, or occurring in atmospheric aerosol (Section 2.3) (Fraser et al., 2003; Grosjean et al., 1996; Fraser et al., 1998), or from the ChemSpider database (Pence and Williams, 2010) (to test for specific functionalities, eg. secondary amide) or from the MCMv3.2 α-pinene oxidation scheme. The patterns
155   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 second group (Table 1, substructures 33-57) of SMARTS patterns was tested on a set of 26 compounds (Table B2, Appendix B) selected from
160   the ChemSpider database and the fourth group (Table 3) 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

165   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), which wraps the ChemSpider
application programming interface. FG composition, $\overline{OS}_C$ 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), Rogge et al. (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, as-
suming fresh cooled emissions (Donahue et al., 2006). After diluting the total OA of a factor of 1000
the compounds were partitioned between the two phases based on the partitioning coefficient ξ_i (xi)
calculated from the pure component saturation concentration (C_i^0) as described by Donahue et al.
(2006).

FG abundance of the set of compounds incorporated in the MCMv3.2 α-pinene and 1,3,5-
trimethylbenzene oxidation schemes was analyzed to demonstrate our validation scheme. Further-
more, the gas phase composition generated by α-pinene photooxidation in the presence of NO$_x$
(α-pinene/NO$_x$ ratio of 1.25), with propene as a radical initiator, was simulated using the Kinetic
Pre-Processor (KPP, Damian et al., 2002; Sandu and Sander, 2006; Henderson, 2016) 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 concen-
trations 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

Figures 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 33-57) are also identical
to the true number of FGs in the set of compounds used for evaluation (Table B2), but are 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 Figure 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

Mapping composition in 2-D volatility basis set space. The algorithm described has been used to project molecular composition of GC-MS and MCM compounds to 2D-VBS space delineated by carbon oxidation and pure component saturation concentration (C^0) (Figure 5). The properties of vehicle-related primary OA and wood combustion compounds measured by GC-MS are generally consistent with those reported for hydrocarbon-like OA and biomass burning OA, respectively, derived from PMF analysis of AMS spectra (Donahue et al., 2012). The low oxidation state is observed on account of more than 60% of carbon atoms being associated with methylene groups (-CH_2-, oxidation state of -2) in long-chain hydrocarbon compounds, and an association to lesser degree with CH groups in aromatic rings (oxidation state of -1) and methyl groups (-CH_3, oxidation state of -3).

Most compounds in the MCMv3.2 system correspond to intermediate volatility organic compounds (IVOC), with only a small fraction with the semivolatile organic compound (SVOC) regime. When using of MCMv3.2 for simulation of secondary OA formation, additional mechanisms (e.g., in the condensed phase) are necessary to introduce low volatility organic compounds (LVOC) as observed in atmospheric and environmental controlled chamber observations (Ehn et al., 2014; Shiraiwa et al., 2014). Higher oxidation states than for compounds in the GC-MS set are observed on account of the larger number of functional groups containing electronegative atoms (oxygen and nitrogen) bonded to carbon.

Source apportionment In Figure 6, the FG distributions of aerosol collected during wood-burning and vehicle emission studies (Rogge et al., 1993; Rogge et al., 1998) 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 studies by Rogge et al. (1993, 1998) have been chosen as they have been used as a reference in the study of composition of organic aerosol from combustion sources (Heringa et al., 2012). 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); these fractions form the bases for comparisons. 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).

From this comparison, we find that the oxidized fraction is much higher in the biomass burning aerosol composition estimated by FTIR. The high abundance of alkane CH bonds in the compounds reported by GC-MS can be explained by the preference of this analytical method to characterize the least oxidized fraction of the collected aerosol. While high abundance of carbonyl groups are reported in FTIR measurements of biomass burning aerosol (Liu et al., 2009; Russell et al., 2009; Hawkins and Russell, 2010), more recent methods including advanced derivatization (Dron et al.,

2010) are necessary for quantification of carbonyl containing compounds by GC-MS. In addition, neither amine compounds nor levogucosan were reported in this GC-MS study. Levoglucosan is a polysaccharide compound often used as a tracer for burning and decomposition of cellulose reported in modern GC-MS measurements (Simoneit, 1999). However, FTIR does not report high fraction of alcohol COH as levoglucosan near particular fuel sources may be found mostly in supermicron diameter particles (Radzi bin Abas et al., 2004) (submicron OA was analyzed by Hawkins and Russell, 2010), its degradation in the atmosphere is rapid (Hennigan et al., 2010; Cubison et al., 2011; Lai et al., 2014), and the overall mass contribution to biomass burning OA is small (less than 2% by mass, Leithead et al., 2006).

Both estimation methods agree that more than 90% of OM mass is composed of alkane-CH for vehicle sources 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.

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 Figure 7, contributions of FGs to the O:C ratio of the gas phase mixture generated by α-pinene photooxidation in low NO_x conditions (Section 2.3) is reported as a function of irradiation time. A singular peroxyacyl nitrate compound (peroxyacetyl nitrate) accounts for 26% of the total gas phase mass. The peroxyacyl nitrate functional group furthermore accounts for the greatest fraction of the total O:C ratio after 20 hours of simulation (53% of the total O:C), as it contains five oxygen atoms per FG. A full analysis on oxidation products with gas/particle partitioning is discussed by Ruggeri et al. (2016). 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

We introduced the application of chemoinformatic tools that allow us to perform substructure matching in molecules to enumerate FGs present in compounds relevant for organic aerosol chemistry. We developed 50+ substructure patterns and validated them over a list of 125 compounds that were selected in order to account for all the functional groups (FGs) represented. We demonstrate how

these tools can facilitate intercomparisons between GC-MS and FTIR measurements, and mapping of compounds onto the VBS space described by pure component vapor pressure and oxidation state.

275   We further introduce a novel approach for defining a set of patterns which accounts for each atom in a chemical system once and only once (except for polyfunctional carbon atoms associated with multiple FGs). This condition is confirmed by an atomic-level validation scheme applied to chemically explicit α-pinene and 1,3,5-TMB degradation mechanisms. This validation scheme provides an intermediate resolution between molecular speciation and atomic composition, and permits

280 apportionment of conventionally aggregated quantities such as O:C, H:C, and N:C to contributions from individual FGs. We illustrate its application to the photochemical degradation of α-pinene from speciated simulations using MCMv3.2.

    These applications can be further adapted for other methods developed to match substructures for other measurements, or enumerate groups used in group contribution methods for estimation

285 of vapor pressures, activity coefficients, and Henry's law constants (Raventos-Duran et al., 2010; Compernolle et al., 2011; 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

290 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:

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

295 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:

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

9

Appendix B: Compounds used for testing the chemoinformatic patterns

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-]
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
2,4-dimethyl-benzaldehyde	CC1=CC(=C(C=C1)C=O)C
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 33-57 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-]

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 (<u>APRL</u> <u>S</u>ubstructure <u>S</u>earch <u>P</u>rogram; Takahama, 2015), 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.

References

310   Aimanant, S. and Ziemann, P. J.: Development of Spectrophotometric Methods for the Analysis of Functional Groups in Oxidized Organic Aerosol, Aerosol Science and Technology, 47, 581–591, doi:10.1080/02786826.2013.773579, http://dx.doi.org/10.1080/02786826.2013.773579, 2013.

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, Atmospheric
315   Chemistry and Physics, 5, 2497–2517, doi:10.5194/acp-5-2497-2005, 2005.

Balaban, A. T.: Applications of graph theory in chemistry, Journal of Chemical Information and Computer Sciences, 25, 334–343, doi:10.1021/ci00047a033, 1985.

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
320   condensed compounds generated by a near-explicit model of VOC oxidation, Atmospheric Chemistry and Physics, 11, 13 145–13 159, doi:10.5194/acp-11-13145-2011, 2011.

Barnard, J. M.: Substructure searching methods: Old and new, Journal of Chemical Information and Computer Sciences, 33, 532–538, doi:10.1021/ci00014a001, 1993.

Berger, F., Flamm, C., Gleiss, P. M., Leydold, J., and Stadler, P. F.: Counterexamples in Chemical Ring Percep-
325   tion, Journal of Chemical Information and Computer Sciences, 44, 323–331, doi:10.1021/ci030405d, 2004.

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, Atmospheric Chemistry and Physics, 5, 641–664, doi:10.5194/acp-5-641-2005, 2005.

330   Brown, W. H., Foote, C. S., Iverson, B. L., and Anslyn, E. V.: Organic Chemistry, Books/Cole, Cengage learning, 20 Davis Drive, Belmont, CA 94002-3098, USA, 2012.

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

335   Chan, M. N., Nah, T., and Wilson, K. R.: Real time in situ chemical characterization of sub-micron 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.

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 spectrome-
340   try 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.

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, Atmospheric Environment, 54, 511–518, doi:10.1016/j.atmosenv.2012.02.074, 2012.

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

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, Analytical Chemistry, 87, 7510–7514, doi:10.1021/acs.analchem.5b01507, 2015.

Cubison, M. J., Ortega, A. M., Hayes, P. L., Farmer, D. K., Day, D., Lechner, M. J., Brune, W. H., Apel, E., Diskin, G. S., Fisher, J. A., Fuelberg, H. E., Hecobian, A., Knapp, D. J., Mikoviny, T., Riemer, D., Sachse, G. W., Sessions, W., Weber, R. J., Weinheimer, A. J., Wisthaler, A., and Jimenez, J. L.: Effects of aging on organic aerosol from open biomass burning smoke in aircraft and laboratory studies, Atmospheric Chemistry and Physics, 11, 12 049–12 064, doi:10.5194/acp-11-12049-2011, http://www.atmos-chem-phys. net/11/12049/2011/, 2011.

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

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.

DAYLIGHT Chemical Information Systems, Inc.: http://www.daylight.com/dayhtml/doc/theory/theory.smarts. html, accessed 30 September 2015.

Decesari, S., Facchini, M. C., Fuzzi, S., and Tagliavini, E.: Characterization of water-soluble organic compounds in atmospheric aerosol: A new approach, Journal of Geophysical Research: Atmospheres, 105, 1481–1489, doi:10.1029/1999JD900950, 2000.

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

Donahue, N. M., Robinson, A. L., Stanier, C. O., and Pandis, S. N.: Coupled partitioning, dilution, and chemical aging of semivolatile organics, Environmental Science & Technology, 40, 2635–2643, doi:10.1021/es052297c, 2006.

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, Proceedings of the National Academy of Sciences of the United States of America, 109, 13 503–13 508, doi:10.1073/pnas.1115186109, 2012.

Downs, G. M., Gillet, V. J., Holliday, J. D., and Lynch, M. F.: Review of ring perception algorithms for chemical graphs, Journal of Chemical Information and Computer Sciences, 29, 172–187, doi:10.1021/ci00063a007, 1989.

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, Atmospheric Chemistry and Physics, 10, 7041–7055, doi:10.5194/acp-10-7041-2010, 2010.

Ehn, M., Thornton, J. A., Kleist, E., Sipilä, M., Junninen, H., Pullinen, I., Springer, M., Rubach, F., Tillmann, R., Lee, B., Lopez-Hilfiker, F., Andres, S., Acir, I.-H., Rissanen, M., Jokinen, T., Schobesberger, S., Kangasluoma, J., Kontkanen, J., Nieminen, T., Kurtén, T., Nielsen, L. B., Jø gensen, S., Kjaergaard, H. G., Canagaratna, M., Maso, M. D., Berndt, T., Petäjä, T., Wahner, A., Kerminen, V.-M., Kulmala, M., Worsnop, D. R.,

Wildt, J., and Mentel, T. F.: A large source of low-volatility secondary organic aerosol, Nature, 506, 476–479, doi:10.1038/nature13032, http://www.nature.com/nature/journal/v506/n7489/full/nature13032.html, 2014.

390   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.

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 and QSAR in Environmental Research, 19, 555–578, doi:10.1080/10629360802348985, 2008.

395   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, Environmental Science & Technology, 36, 1442–1451, doi:10.1021/es0108988, 2002.

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, Environmental
400   Science & Technology, 46, 6048–6055, doi:10.1021/es3003734, 2012.

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, Environmental Science & Technology, 32, 1760–1770, doi:10.1021/es970349v, 1998.

Fraser, M. P., Cass, G. R., and Simoneit, B. R. T.: Air quality model evaluation data for organics. 6. C-3-C-24
405   organic acids, Environmental Science & Technology, 37, 446–453, doi:10.1021/es0209262, 2003.

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, Journal of Geophysical Research-atmospheres, 107, 4334, doi:10.1029/2001JD000544, 2002.

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

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, Atmospheric Chemistry and Physics, 4, 1279–1290, doi:10.5194/acp-4-1279-2004, 2004.

415   Hann, M., Hudson, B., Lewell, X., Lifely, R., Miller, L., and Ramsden, N.: Strategic Pooling of Compounds for High-Throughput Screening, Journal of Chemical Information and Computer Sciences, 39, 897–902, doi:10.1021/ci990423o, 1999.

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

Henderson, B. H.: Kinetic Pre-Processor with updates to allow working with MCM, doi:10.5281/zenodo.44682, http://github.com/barronh/kpp, 2016.

Hennigan, C. J., Sullivan, A. P., Collett, Jeffrey L., J., and Robinson, A. L.: Levoglucosan stability in biomass burning particles exposed to hydroxyl radicals, Geophysical Research Letters, 37, L09 806,
425   doi:10.1029/2010GL043088, 2010.

Heringa, M. F., DeCarlo, P. F., Chirico, R., Lauber, A., Doberer, A., Good, J., Nussbaumer, T., Keller, A., Burtscher, H., Richard, A., Miljevic, B., Prevot, A. S. H., and Baltensperger, U.: Time-Resolved Charac-

terization of Primary Emissions from Residential Wood Combustion Appliances, Environmental Science & Technology, 46, 11 418–11 425, doi:10.1021/es301654w, http://dx.doi.org/10.1021/es301654w, pMID: 22970884, 2012.

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, Atmospheric Environment, 39, 4351–4363, doi:10.1016/j.atmosenv.2005.02.016, 2005.

Hildemann, L. M., Markowski, G. R., and Cass, G. R.: Chemical-composition of Emissions From Urban Sources of Fine Organic Aerosol, Environmental Science & Technology, 25, 744–759, doi:10.1021/es00016a021, 1991.

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 Science and Technology, 33, 49–70, doi:10.1080/027868200410840, 2000.

Jenkin, M. E.: Modelling the formation and composition of secondary organic aerosol from alpha- and beta-pinene ozonolysis using MCM v3, Atmospheric Chemistry and Physics, 4, 1741–1757, doi:10.5194/acp-4-1741-2004, 2004.

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

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, Atmospheric Chemistry and Physics, 3, 181–193, doi:10.5194/acp-3-181-2003, 2003.

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, Environmental Science & Technology, 40, 5917–5922, doi:10.1021/es0525760, 2006.

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.

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.

Kroll, J. H., Lim, C. Y., Kessler, S. H., and Wilson, K. R.: Heterogeneous Oxidation of Atmospheric Organic Aerosol: Kinetics of Changes to the Amount and Oxidation State of Particle-Phase Organic Carbon, J. Phys. Chem. A, 119, 10 767–10 783, doi:10.1021/acs.jpca.5b06946, http://dx.doi.org/10.1021/acs.jpca.5b06946, 2015.

Lai, C., Liu, Y., Ma, J., Ma, Q., and He, H.: Degradation kinetics of levoglucosan initiated by hydroxyl radical under different environmental conditions, Atmospheric Environment, 91, 32–39, http://www.sciencedirect.com/science/article/pii/S1352231014002398, 2014.

Landrum, G.: RDKit: Open-source cheminformatics, http://www.rdkit.org, accessed 30 September 2015.

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, Analytical Chemistry, 84, 7179–7187, doi:10.1021/ac301533z, 2012.

470 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, Atmospheric Environment, 40, 2721–2734, doi:10.1016/j.atmosenv.2005.09.084, 2006.

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, Environmental Science & Technology, 36, 4489–4496, doi:10.1021/es0206487, 2002.

475 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, Atmospheric Chemistry and Physics, 9, 6849–6863, doi:10.5194/acp-9-6849-2009, 2009.

Maria, S. F., Russell, L. M., Turpin, B. J., and Porcja, R. J.: FTIR measurements of functional groups and organic

480 mass in aerosol samples over the Caribbean, Atmospheric Environment, 36, 5185–5196, doi:10.1016/S1352-2310(02)00654-4, 2002.

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, Journal of Geophysical Research-atmospheres,

485 108, doi:10.1029/2003JD003703, 2003.

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.

Miloslav, N., Jiri, J., and Bedrich, K.: IUPAC Compendium of Chemical Terminology- the Gold Book, http://goldbook.iupac.org, accessed 30 September 2015.

490 Ming, Y. and Russell, L. M.: Predicted hygroscopic growth of sea salt aerosol, Journal of Geophysical Research-atmospheres, 106, 28 259–28 274, doi:10.1029/2001JD000454, 2001.

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.

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

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.

Olah, M., Bologa, C., and Oprea, T.: An automated PLS search for biologically relevant QSAR descriptors,

500 Journal of Computer-Aided Molecular Design, 18, 437–449, doi:10.1007/s10822-004-4060-8, 2004.

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.

Pankow, J. F. and Asher, W. E.: SIMPOL.1: a simple group contribution method for predicting vapor pressures and enthalpies of vaporization of multifunctional organic compounds, Atmospheric Chemistry and Physics,
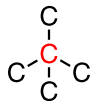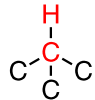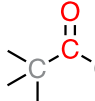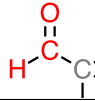
505 8, 2773–2796, doi:10.5194/acp-8-2773-2008, 2008.

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, Atmospheric Environment, 43, 2829–2835, doi:10.1016/j.atmosenv.2008.12.050, 2009.

Pavia, D., Lampman, G., and Kriz, G.: Introduction to Spectroscopy, Brooks/Cole Pub Co., 2008.

510   Pence, H. E. and Williams, A.: ChemSpider: An Online Chemical Information Resource, Journal of Chemical Education, 87, 1123–1124, doi:10.1021/ed100697w, 2010.

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, 1089–95, doi:10.1016/j.chemosphere.2004.02.002, 2004.

515   Raventos-Duran, T., Camredon, M., Valorso, R., Mouchel-Vallon, C., and Aumont, B.: Structure-activity relationships to estimate the effective Henry's law constants of organics of atmospheric interest, Atmospheric Chemistry and Physics, 10, 7643–7654, doi:10.5194/acp-10-7643-2010, 2010.

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, Environmental

520   Science & Technology, 27, 636–651, doi:10.1021/es00041a007, 1993.

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, Environmental Science & Technology, 32, 13–22, doi:10.1021/es960930b, 1998.

Ruggeri, G., Alexander, F. B., Takahama, S., and Henderson, B. H.: Model-measurement comparison of func-

525   tional group abundance in α-pinene and 1,3,5-trimethylbenzene secondary organic aerosol formation, In preparation, 2016.

Russell, L. M.: Aerosol organic-mass-to-organic-carbon ratio measurements, Environmental Science & Technology, 37, 2982–2987, doi:10.1021/es026123w, 2003.

Russell, L. M., Bahadur, R., Hawkins, L. N., Allan, J., Baumgardner, D., Quinn, P. K., and Bates, T. S.: Or-

530   ganic aerosol characterization by complementary measurements of chemical bonds and molecular fragments, Atmospheric Environment, 43, 6100–6105, doi:10.1016/j.atmosenv.2009.09.036, 2009.

Russell, L. M., Bahadur, R., and Ziemann, P. J.: Identifying organic aerosol sources by comparing functional group composition in chamber and atmospheric particles, Proceedings of the National Academy of Sciences of the United States of America, 108, 3516–3521, doi:10.1073/pnas.1006461108, 2011.

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

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 com-

540   pounds, Atmospheric Chemistry and Physics, 3, 161–180, doi:10.5194/acp-3-161-2003, 2003.

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, The Journal of Physical Chemistry A, 119, 4281–4297, doi:10.1021/jp501779w, 2015.

Seinfeld, J. H. and Pandis, S. N.: Atmospheric Chemistry and Physics: From Air Pollution to Climate Change,

545   John Wiley & Sons, New York, 2nd edition edn., 2006.

Shiraiwa, M., Berkemeier, T., Schilling-Fahnestock, K. A., Seinfeld, J. H., and Pöschl, U.: Molecular corridors and kinetic regimes in the multiphase chemical evolution of secondary organic aerosol, Atmospheric Chemistry and Physics, 14, 8323–8341, doi:10.5194/acp-14-8323-2014, http://www.atmos-chem-phys.net/14/8323/2014/, 2014.

550   Simoneit, B. R. T.: A review of biomarker compounds as source indicators and tracers for air pollution, Environmental Science and Pollution Research, 6, 159–169, doi:10.1007/BF02987621, 1999.

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., and Willighagen, E.: The Chemistry Development Kit (CDK):‚Äâ An Open-Source Java Library for Chemo- and Bioinformatics, Journal of Chemical Information and Computer Sciences, 43, 493–500, doi:10.1021/ci025584y, 2003.

555   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, Environmental Science & Technology, doi:10.1021/es502147y, 2014.

Swain, M.: ChemSpiPy, http://chemspipy.readthedocs.org, last accessed 30 September 2015.

Takahama, S.: APRL Substructure Search Program, doi:10.5281/zenodo.34975, https://github.com/stakahama/
560   aprl-ssp, 2015.

Topping, D., Barley, M., Bane, M. K., Higham, N., Aumont, B., Dingle, N., and McFiggans, G.: UManSysProp v1.0: an online and open-source facility for molecular property prediction and atmospheric aerosol calculations, Geoscientific Model Development, 9, 899–914, doi:10.5194/gmd-9-899-2016, http://www.geosci-model-dev.net/9/899/2016/, 2016.

565   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, 10 933–10 950, doi:10.5194/acp-13-10933-2013, 2013.

Walters, W. and Murcko, M. A.: Prediction of 'drug-likeness', Advanced Drug Delivery Reviews, 54, 255 –
570   271, doi:10.1016/S0169-409X(02)00003-0, 2002.

Weininger, D.: Smiles, A Chemical Language and Information-system .1. Introduction To Methodology and Encoding Rules, Journal of Chemical Information and Computer Sciences, 28, 31–36, doi:10.1021/ci00057a005, 1988.

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

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‚Äìparticle partitioning of organic acids in a ponderosa pine forest using a MOVI-
580   HRToF-CIMS, Atmos. Chem. Phys., 14, 1527–1546, doi:10.5194/acp-14-1527-2014, 2014.

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 Science and Technology, 49, 727–738, doi:10.1080/02786826.2015.1068427, 2015.

Zhang, Q., Jimenez, J. L., Canagaratna, M. R., Allan, J. D., Coe, H., Ulbrich, I., Alfarra, M. R., Takami,
585   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, Geophysical Research Letters, 34, L13 801, doi:10.1029/2007GL029979, 2007.

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

Zuend, A., Marcolli, C., Booth, A. M., Lienhard, D. M., Soonsin, V., Krieger, U. K., Topping, D. O., Mc-Figgans, 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, Atmospheric Chemistry and Physics, 11, 9155–9206, doi:10.5194/acp-11-9155-2011, 2011.

590

595

Tables

Table 1a. 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 33-57). [removed text from here and added in the Method section]For space constraints the SMARTS patterns have been reported on multiple lines, even if the SMARTS notation requires unique lines.

N°	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;!$(C=O)][OX2H][H]`	
8	Ketone	A compound containing a carbonyl group bonded to two carbon atoms.[1]	`[CX3;$(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]`	

[1] Brown et al. (2012)

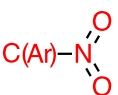600   [2] Miloslav et al.

23

Table 1b. Continued
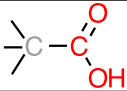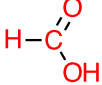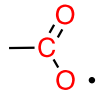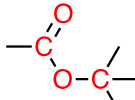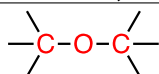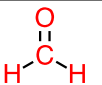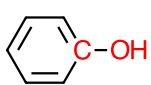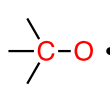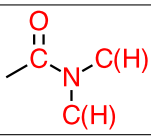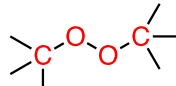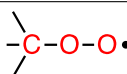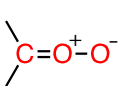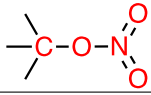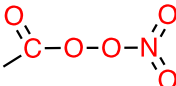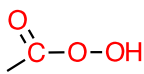
N°	Substructure	Definition	Chemoinformatic definition	Matched pattern
10	Carboxylic acid	A compound containing a carboxyl, -COOH, group.[1] (excludes formic acid)	`[CX3;!$([CX3][H])](=O)` `[OX2H][H]`	
11	Formic acid	Formic acid compound.	`[CX3](=O)([H])[OX2H][H]`	
12	Acyloxy radical	Oxygen-centered radicals consisting of an acyl radical bonded to an oxygen atom.[2]	`[C;$(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;!$(N=O)]` `([#6,#1])[#6,#1]`	
19	Peroxide	Compounds of structure ROOR in which R may be any organyl group.[2]	`[#6][OD2][OD2,OD1][#6]`	
20	Peroxy radical	Oxygen centered radical derived from an hydroperoxide.	`[O;!$([O][#6]);!$([O][H]);` `!$([OX2][N]);!$(O=C)][O]` `[#6;!$([C](=O)~OO)]`	
21	C=O+-O− group	Group of the type C=O+-O−	`[O;!$([O][#6]);!$([O][H]);` `!$([OX2][N]);!$(O=C)]` `[O]=[#6;!$([C](=O)~OO)]` `([#6,#1])[#6,#1]`	
22	C-nitro	Compounds having the nitrogroup, -NO2 (free valence on nitrogen), which is attached to a carbon.[2]	`[#6][$([NX3](=O)=O),` `$([NX3+](=O)[O-])](~[O])` `(~[O])`	
23	Organonitrate	Compounds having the nitrogroup, -NO2 (free valence on nitrogen), which is attached to an oxygen.[2]	`[#6][O][$([NX3](=[OX1])` `(=[OX1])O),$([NX3+]([OX1-])` `(=[OX1])O)](~[O])(~[O])`	
24	Peroxyacyl nitrate	Functional group containing a -COOONO2.	`[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]`	

24

Table 1c. Continued

N°	Substructure	Definition	Chemoinformatic definition	Matched pattern
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]`	
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,β positions and a phenyl group.	`[CX3;$(C(=[O])([#1])[C]=[C] [c;$(c1cc[c]cc1)])] ([C])(=[O;!$([O][O])])[H]`	

Table 1d. Continued

N°	Substructure	Definition	Chemoinformatic definition	Matched pattern
43	Conjugated ketone C=O and α,α 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]`	C=C-C-C / O
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;!$([O][O])])[c]`	Ar-C-C / O
45	Conjugated ketone C=O and two phenyl	A ketone C=O conjugated with two phenyl groups.	`[CX3;$(C([c,$(c1cc[c]cc1)]) (=[O])[c;$(c1cc[c]cc1)])] ([c])(=[O;!$([O][O])])[c]`	Ar-C-Ar / O
46	Conjugated ester C=O and α,β C=C	An ester C=O conjugated with alkene C=C in α and β positions .	`[C;!$(Cc)]=[C] [CX3;$([C]([O][C]) (=[O])[C]=[C])]([O][C]) (=[O;!$([O][O])])`	C=C-C-O-C / 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;!$([O][O])])`	Ar-C-O-C / O
48	Conjugated ester C-O with C=C or phenyl	An ester C=O conjugated with alkene C=C in α and β positions and a phenyl group.	`[CX3;$([C]([#6])(=[O])[O] [C]=[C]),$([C]([#6])(=[O]) [O][c;$(c1cc[c]cc1)])] (=[O;!$([O][O])])[O] [#6;$(C=C),$(c1cc[c]cc1)]`	C-C-O-C=C / O and C-C-O-Ar / O
49	Nonacid carbonyl	Carbonyl group in ketones and aldehydes.	`[CX3;$(C([#6,#1])(=[O]) [#6,#1])](=[O;!$([O][O])])`	C(H) C=O C(H)
50	Acyl Chloride	An acyl group bonded to a chloride atom.	`[C,$([C]([#6])(=[O]))] (=O)[Cl]`	O C C Cl
51	Isocyanate	An -N=C=O group.	`[N;$([N]([#6])=[C]=[O])] =[C]=[O]`	C-N=C=O
52	Isothiocyanate	An -N=C=S group.	`[N;$([N]([#6])=[C]=[S])] =[C]=[S]`	C-N=C=S
53	Imine	A carbon- nitrogen double bond, R$_2$C=NR.	`[C;$(C([#6,#1])([#6,#1])=[N])] =[N][#1,#6]`	C(H) C=N C(H) C(H)
54	Oxime	A carbon-nitrogen double bond, R$_2$C=NOH.	`[C;$(C([#6,#1])([#6,#1]) =[N][O][H])]=[N][O][H]`	C(H) C=N OH C(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), $([NX3](=O)[O-])]+ (~[O])(~[O])`	C-N O O
56	Aromatic nitro	Compounds having the nitro group, -NO$_2$ (free valence on nitrogen), which is attached to an aromatic carbon.	`[c][$([NX3](=O)=O), $([NX3](=O)[O-])]+ (~[O])(~[O])`	Ar-N O O
57	Nitrile	A carbon atom bonded to a nitrogen atom with a triple bond.	`[C;$([C]#[N])]#[N]`	C-C≡N

Table 2: Chemical substructures required by SIMPOL.1 model (Pankow and Asher, 2008). The column denoted by k corresponds to the group number of 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*†	`{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‡	`count_aromatic_rings(molecule)`	3
Non aromatic ring‡	`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;!$([O][O])])`	8
Ketone	Table 1, number 8	9
Carboxylic acid	`[CX3](=O)[OX2H][H]`	10
Ester (SIMPOL.1)†	`{Ester}-{Nitroester}`	11
Ether (SIMPOL.1)	`[OD2]([C;!R;!$(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
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

27

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[‡]	`count_nitrophenols(molecule,'{phenol},'{nitro})`	29
Nitroester[*]	`[#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))]`	

[*]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.

[†]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.

[‡]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. List of SMARTS patterns and coefficients associated with each bond type, used to calculate the carbon oxidation state as described in the Section 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 4: 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)
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)

Figures
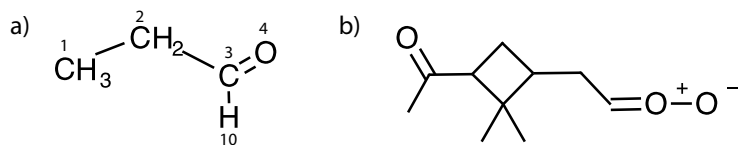
Figure 1. Propionaldehyde (a, SMILES code CCC=O) and compound named APINOOB in MCMv3.2 scheme (b, SMILES code [O-][O+]=CCC1CC(C(=O)C)C1(C)C). The carbon and oxygen atoms are enumerated, together with the hydrogen of the aldehyde group in compound a.
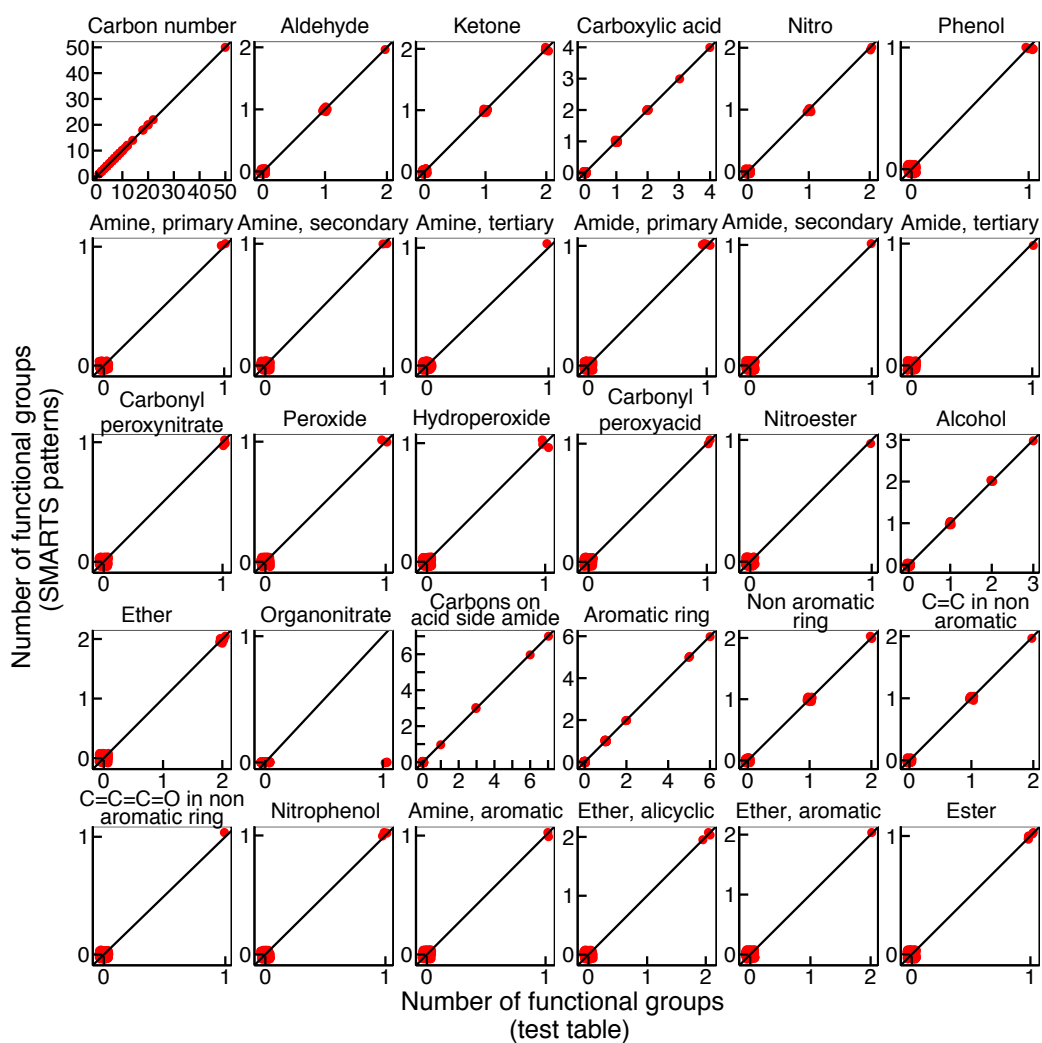
Figure 2. Validation of the developed chemoinformatic patterns for the chemical substructures required in the SIMPOL.1 model (Pankow and Asher, 2008). This validation set includes 99 compounds as described in Section 2.
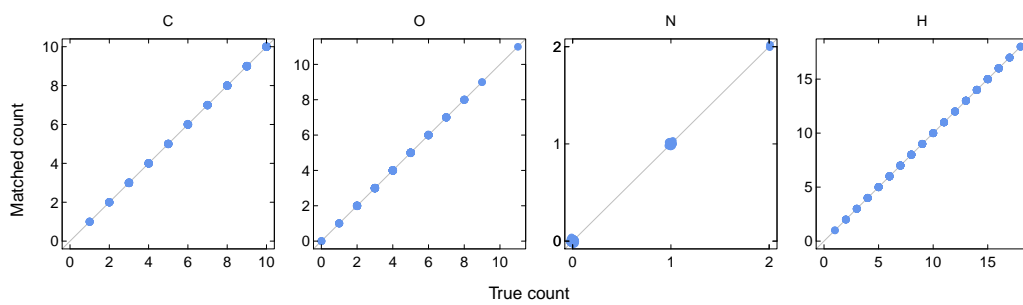
Figure 3. Test of the completeness of matching of all the atoms in the α-pinene and 1,3,5-trimethylbenzene degradation scheme in MCMv3.2 by the SMARTS patterns in Table 1, substructures 1-33.
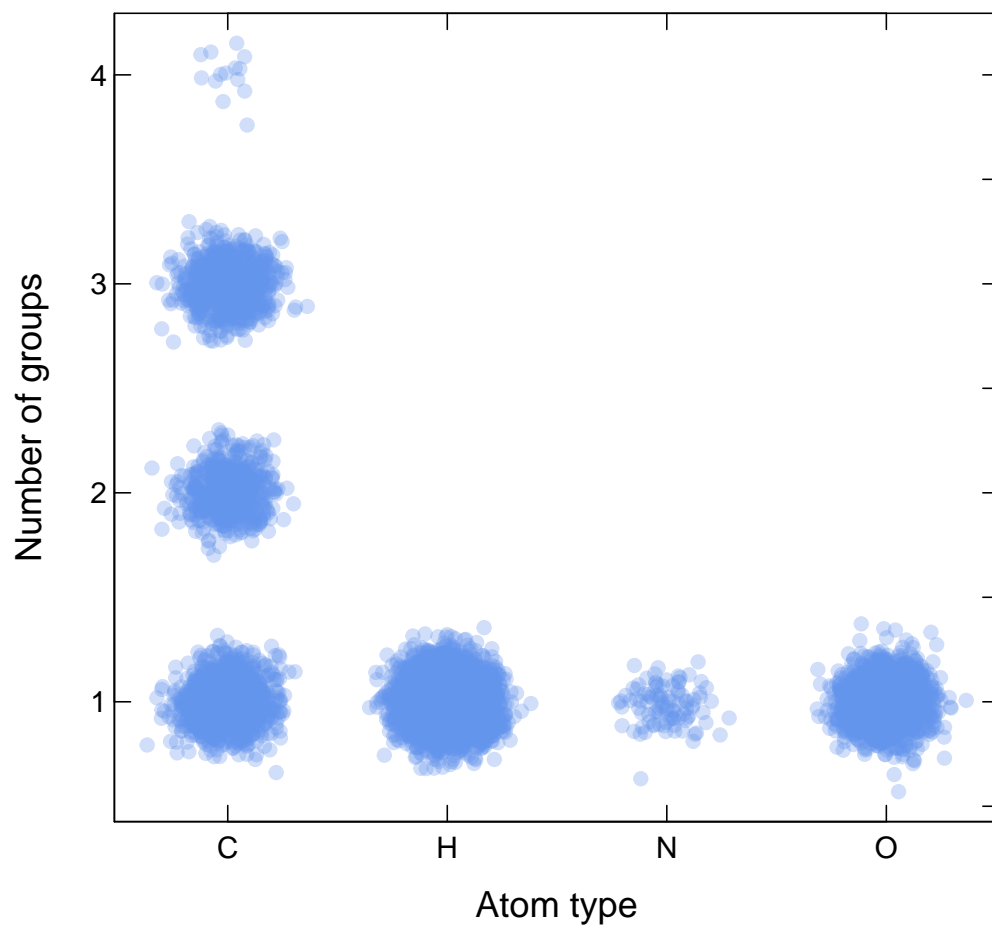
Figure 4. Test for the uniqueness of matching for each atom. Number of times a specific atom has been matched, in the α-pinene and 1,3,5-trimethylbenzene degradation scheme in MCMv3.2 by the SMARTS patterns in Table 1, substructures 1-33. Oxygen, nitrogen and hydrogen atoms are matched only once. The carbon atoms are matched multiple times when multifunctional.
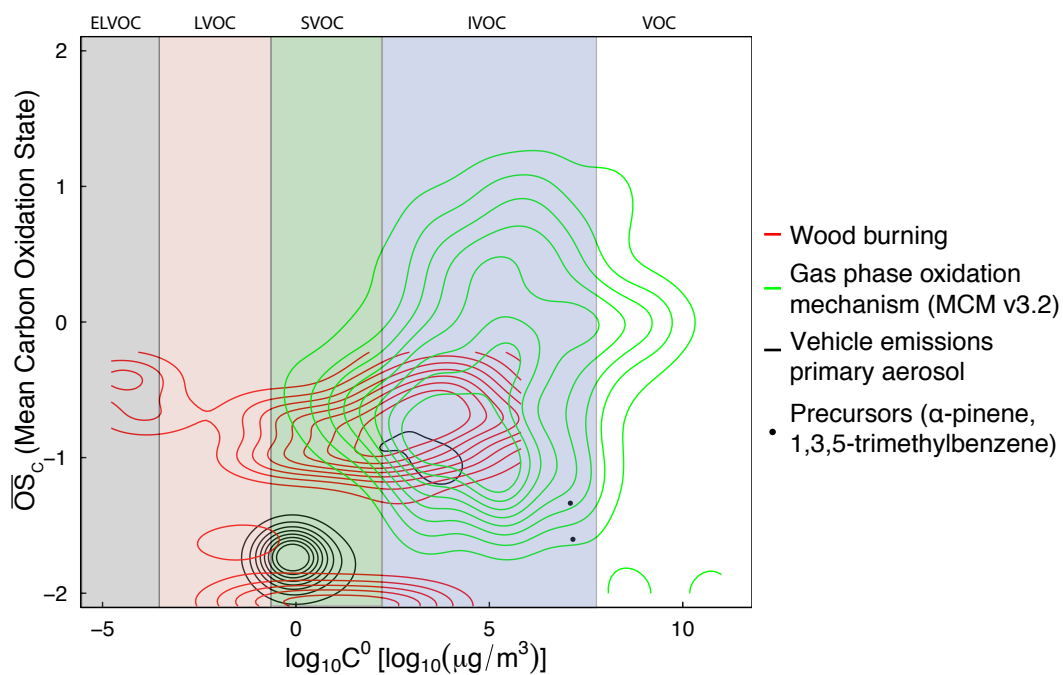
Figure 5. Logarithm of the pure component saturation concentration ($\log_{10}C^0$) and mean carbon oxidation state of each compound ($\overline{OS}_C$) measured by Rogge et al. (1993) and Rogge et al. (1998) for biomass burning and vehicle emissions sources (green and blue lines), and of each molecule constituting the MCMv3.2 gas phase oxidation mechanism of α-pinene and 1,3,5-trimethylbenzene. The lines in the plot denote isolines $(0, 0.1, ..., 0.9)$ 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).
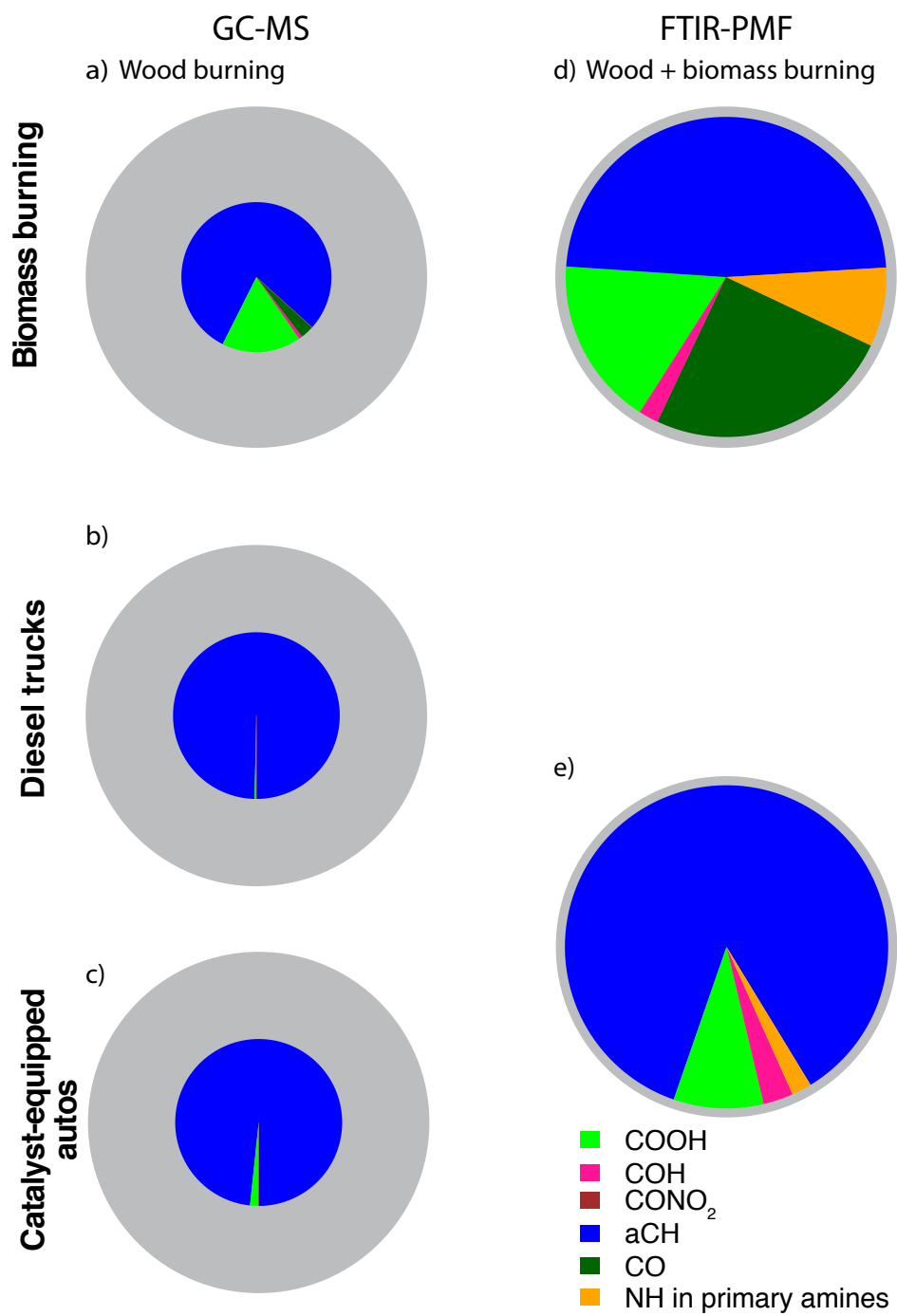
Figure 6. Comparison of the FG distribution of the quantified fraction measured by GC-MS (a,b and c; Rogge et al., 1998; Rogge et al., 1993) and FTIR-PMF (d and e; Hawkins and Russell, 2010) in aerosol emitted by biomass burning (a and d) and vehicle emission (b,c and e) sources. The gray area is the unresolved OA fraction by the two different analytical techniques used (around 80% for GC-MS and around 10% for FTIR). The type of biomass burning is specified in the pie charts a and d.
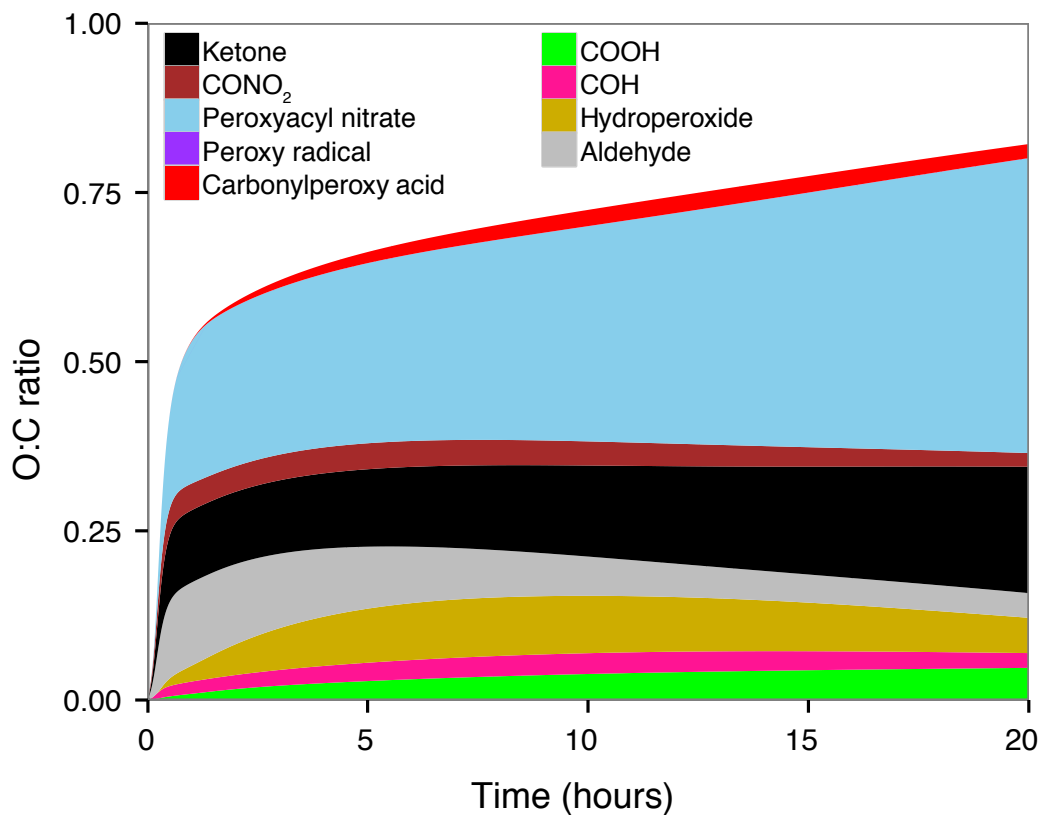
Figure 7. Time series of FG contributions to the total O:C of the gas phase generated by photooxidation of α-pinene in low-NO$_x$ regime, simulated using the MCMv3.2 degradation scheme.