Response to Reviewer #1

General comments:

In this paper the authors report further on a framework they have been developing to extend the utility of functional group analysis of organic aerosol to obtain more accurate information on possible missing chemical components, organic mass/organic carbon ratio, elemental composition (O/C ratio), and carbon oxidation state. A thorough discussion of the conceptual and analytical aspects of the methods are presented and discussed, and software is made available that can be downloaded by others wishing to use these methods. The methods were applied to a data set of SOA composition predicted for a-pinene photooxidation using the MCM coupled with gas-particle partitioning, and for which FTIR functional group data were also available. A variety of useful results are extracted from the analysis. The paper is technically very dense and beyond my ability to fully understand without a much greater investment of time and effort, and so I am unable to evaluate many of the details that are presented. Nonetheless, the approach seems reasonable to me.

Overall, I think this is likely to be a useful paper as more people begin to adapt these methods for analysis of functional group data. Not many group conduct functional group analysis, but there are reasons to think this could increase in the future because of its unique value compared with molecular and AMS analysis. Im very pleased to see that the authors have made the software needed to conduct the analysis freely available. I recommend the paper be published after these very minor comments are addressed.

We thank the reviewer for this encouraging assessment.

Specific comments:

1. Page 5, line 7: I have never heard it said that FTIR measures the total abundance of bonds. The spectrum depends on vibrations and bending of molecules, but I dont see how this translates directly into bond abundance. This should be clarified.

   The absorbances associated with stretching and bending modes of molecular bonds can be calibrated (according to Beer-Lambert law) to quantify their abundances, but it may have been confusing to include the word, "total", which was meant to indicate the independence from extraction or ionization efficiencies often used in other types of chemical analysis. Also, describing the method of quantification is not the main purpose of this Section 2.2. To remedy these issues, we have first simplified the expression to in Section 2.2:

   "The molar abundance of molecules $\boldsymbol{n}_{\text{molec}} = [n_i]$ in a mixture (consisting of a set of molecules denoted by $\mathcal{M}$) can be related to FG abundance $\boldsymbol{n}_{\text{group}} = [n_j]$ (for each FG in $\mathcal{J}$) obtained by FT-IR — or other means — by invoking a group composition matrix $\mathbf{X} = [x_{ij}]$, which describes the FG makeup of each molecule."

   We have also included the appropriate references for obtaining bond or FG abundances from FT-IR spectra and other methods in Section 1:

   "However, studies on this topic have thus far been very limited on account of challenges in quantitative characterization of FGs, which requires either advanced algorithms (e.g., Takahama et al., 2013; Ruthenburg et al., 2014; Takahama and Dillner, 2015) for spectral interpretation or derivitization steps (e.g., Dron et al., 2010; Aimanant and Ziemann, 2013) for chemical analysis."

Furthermore, we have edited Section 1 to make note of the fact that we are not considering less than "total recovery" of bonds as we do not chemically extract our samples:

"The benefit of developing a systematic approach is that we can precisely understand the achievable mass recovery, and biases incurred on the calculated O/C and OM/OC for a given set of molecules and FGs analyzed (when chemical extraction is not required, OM mass recovery is primarily dependent on the completeness of FG calibration models constructed)."

2. Page 13, lines 4–5: Baltensperger and co-workers (AMT) measured a peroxide lifetime of a few hours in chamber SOA generated from a-pinene + ozone.

   We thank the reviewer for pointing out this important reference (in *Cell*) — the text has been modified to include this citation:

   "The rate of transformation of these FGs remains uncertain — for instance, reported lifetimes of hydroperoxides range from less than an hour to many days (Epstein et al., 2014; Krapf et al., 2016); resolving their reaction pathways may play a critical role in understanding model-measurement discrepancies (McVay et al., 2016)."

Technical comments:

1. Page 1, line 17: Insert "a" before "framework".
2. Page 3, line 1: Should "metric" be inserted after "carbon-centric"?
3. Page 4, line 8: Delete "a" after "specific".
4. Page 5, line 10 or 11. Delete "weighted" on one line or the other. Page 6, line 1: Can probably delete "expressed".
5. Page 6, line 2: Should be "below" not "blow".
6. Page 7, line 4: Should "atoms" be inserted after "carbon"?
7. Page 7, line 11: Should a comma be added after "eCH"?

We thank the reviewer for these technical corrections — the changes have been made in the manuscript.

Response to Reviewer #2

The present paper is a technical note to already-published model-measurement comparison SOA studies (Ruggeri et al., ACP 2016). The general scientific objectives are already explained in the previous publications. Specifically, the Authors aim to exploit organic functional group distributions to constrain explicit-chemistry models of secondary organic aerosol (SOA) formation. The idea, although not new, is sound, because functional groups provide direct information about organic reactivity (like in the Carbon Bond Mechanism developed long ago for gas-phase reactions) and keep track of the chemical mechanisms that govern the enrichment of SVOC in the particulate phase. This technical note, in particular, focuses on the derivation of carbon-based metrics (such as oxidation state, and O/C ratios) from measurable functional groups distributions, with the aim to support and inform the comparison between FG-based techniques (such as FTIR) and more established mass spectrometric methods (AMS). The methodology is discussed in detail, and for the first time O/C ratios from FTIR measurements are reported taking into account the possible biases due to the selectivity of FTIR spectroscopy for specific FGs.

We thank the reviewer for this lucid assessment.

Specific comments:

1. Simple examples, like the one shown in Figure 1, are essential for a chemist who is not familiar with linear algebra. I invite the Authors to comment such examples in the text, or in the Supplementary Information. For instance, it is not straightforward why negative values for phi can be obtained. This is important also to understand why values for lambda lower than 1/3 (the theoretical value for a tri-substituted carbon atom) are found in Table 2.

   As the author correctly notes, lambda values should nominally take on values rational values of {1/4, 1/3, 1/2, 1} at the level of individual carbon atoms. However, the coefficients reported represent a single set of values to be used across all carbon types, which leads to irrational values. A value of lambda below 1/3 results from the empirical nature of regression methods, in which obtained coefficient values are insensitive to under-represented FGs (therefore could be replaced with 1/3 with little impact on results), or eliminated (set to zero) in the case of redundant FGs (i.e. strongly correlated to another FG). We have included the following statements in Section 2.4 to address potential questions from other readers:

   "Each of the solutions produces a series of irrational numbers (due to the multiplicitous configurations of FGs on carbon atoms) that may be overly precise for the data set used for estimation."

   and

   "Direct fitting methods, on the other hand, may lead to insignificant coefficients from under-represented or redundant FGs [...]."

   While lambda may have a physical interpretation in simple instances (not in complex mixtures with multiplicitous configurations of functional groups as noted above), negative *values for phi are permitted in that they are only required to satisfy the carbon type balance (equation 4). We have added the following statement to Section 2.4 to better connect readers with the illustration in Figure 1:*

"The elements of $\mathbf{\Phi}$ satisfy the carbon type balance (equation 4) but are not required to be non-negative, but their summation across rows (equation 6) yields values for $\boldsymbol{\lambda}_{\mathrm{C}}$ that corresponds to the number of carbon atoms per FG associated with them.

[...]

"$\boldsymbol{\lambda}_{\mathrm{C}}$ may also not correspond to a physically interpretable quantity in such instances, as a single set of coefficients are insufficient to estimate the exact abundances of carbon atoms under these circumstances."

2. The three methods for carbon abundance estimation discussed in Section 2.4 are compared for the examples of SOA formation considered in this study (Table 2, Figures 8, 9), but it is difficult to derive general conclusions on their applicability. How much details on the qualitative composition (in carbon types) must be known a priori? It would be important to expand the paragraph about the type and nature of the datasets which the three methods rely on.

   It is yet difficult to recommend a best method for estimation in a general sense, but we hope that this exploratory study can initiate further inquiry into this topic. However, as noted by the reviewer, it is worth considering the nature of underlying data sets in more detail to describe their tradeoffs. We had previously written in Section 2.4:

   "Numerical details aside, the main differences among the three are the data sets used for estimation. COUNT uses information from $\mathbf{\Theta}$ only (defined for the FGs in the APIN mechanism), COMPOUND uses carbon type abundances in compounds (limited to SVOCs in the APIN mechanism), and MIXTURE uses mixture information of the condensed-phase (from different periods in the APIN simulation)."

   However, the important difference which may have not been emphasized was the weighting of the estimates. We have added to Section 2.4 the following explanation:

   "The resulting differences in estimates of $\hat{\lambda}_{\mathrm{C}}$ are largely due to weighting of FGs associated with each carbon type: each type receiving equal weight (COUNT), by frequency of occurrence in SVOCs (COMPOUND), and by abundance in SOA formed in the APIN simulation (MIXTURE). While the COUNT method is physically significant at the level of individual carbon atoms, the representativeness of estimated values for use in mixtures can vary according to composition. Direct fitting methods, on the other hand, may suffer from errant coefficients from redundant or under-represented FGs, or be overly specific such that they cannot be generalized to other systems. Therefore, the results from all three methods are evaluated to explore the range of plausible values."

   Regarding the COUNT method, we have also added the following statement to aid the physical significance of this estimate:

   "The main premise of this approach is to apportion fractional units of carbon to each measured FG such that their sum equals unity."

3. The paper makes use of the notion of carbon types. These are exemplified for simple molecules in Figure 1, and are otherwise listed as numbers in the other Tables and Figures. I suggest to explicit the full list in the Supplementary Information. It will be important to

understand how many carbon types contain heteroatoms in functional groups (alcohols, carboxylic acids, nitro groups etc.), which seems to be the focus of the paper, instead of being included in the skeleton of the molecules (ethers, esters, etc.).

We thank the reviewer for this comment. We had discussed the case of anhydrides, esters, and organic peroxides in the supporting information but had not referred to it in the main text (ethers are currently out of the scope of our analysis as they are not included in mechanisms which we have studied). We have modified Section 2.3 to include the following statements:

"All elements in equation 3 can be known precisely for any set of molecules $\mathcal{M}$ from the chemometric patterns and atom-level validation described by Ruggeri and Takahama (2016), and is summarized in Section S1. Furthermore, the FGs included in the APIN system are all those which are defined by association only to single carbon atoms (e.g., alcohol, carboxylic, methylene groups). Methods for extending this analysis to FGs containing multiple carbon atoms (e.g., anhydride, ester, and organic peroxide groups) are described in Section S2."

While the APIN system does not include these multi-carbon functional groups, we have included additional analyses (Supporting Information, Section S2, and Tables S1–S4, and online code repository) to conclude that at least in terms of frequency of occurrence in α-pinene and 1,3,5-trimethylbenzene degradation systems studied by Ruggeri et al. (2016), our 41 carbon types presented in the main text encompasses 92% of the 2867 carbon atoms in the 441 molecules. The four tables can be viewed in the Supporing Information, but the text is copied below:

"Tables S1–S3 show carbon atom types associated with single-carbon FGs (conversely stated, each FG is uniquely associated with one carbon atom), two-carbon FGs (carbon atoms in these FGs share some heteroatoms with other carbon atoms), and carbon-only structures present in the combined set of molecules from the α-pinene and 1,3,5-trimethylbenzene degradation schemes. In this set of 441 molecules, there are 2867 carbon atoms that can be classified into one of 60 types (labeled in order of frequency, X1–X60, prefixed by character 'X' to prevent confusion with carbon type labels used in the APIN simulation) that differ in their association with 30 unique FGs. 46 of these types contain unique FGs (2557 / 2867 carbon atoms belong in this category), 11 of these types share FGs (116 / 2867 carbon atoms belong in this category), and 3 are bonded only to other carbon atoms (194 / 2867 carbon atoms belong in this category). 92% of the carbon atoms in this superset belong to the 41 carbon types (which includes two of the tertiary and quaternary carbon types) from the APIN simulation discussed in the main body of this manuscript, though this relative abundance is reported on a frequency basis and does not consider molecular abundances that might be typical in a SOA mixture. The correspondence of labels used in the main document (numbered by abundance of total carbon during the APIN simulation) and Tables S1–S3 (numbered by frequency of occurrence of in the 441 molecules) are listed in Table S4."

The rest of Section S2 is dedicated to describing adjustments necessary to include carbon atoms associated with anhydrides, esters, and peroxides for estimation of carbon-centered metrics, and can be applied to any functional groups with which more than one carbon atom would be associated. The text has been modified to indicate that our framework is quite general in this sense (Section S2):

"The second generalization concerns FGs that contain skeletal heteroatoms. FGs of this type — specifically in this case, anhydride, ester, and (organic) peroxide — are present in photooxidation products of 1,3,5-trimethylbenzene included in the MCMv3.2 mechanism (Bloss et al., 2005; Ruggeri et al., 2016), and corresponding SMARTS patterns were developed by Ruggeri and Takahama (2016) to match these structures. Equation S3 should accordingly permit two carbon atoms to be associated with each of these exceptional FGs. To accommodate such groups (and other FGs defined by membership of multiple carbon atoms) in our framework, the carbon type formulation can be a) extended to "carbon units" consisting of one or more carbon atoms and their bonded heteroatoms, or b) modified by the introduction of a correction factor."

Technical Note: Relating functional group measurements to carbon types for improved model-measurement comparisons of organic aerosol composition

Satoshi Takahama[1] and Giulia Ruggeri[1]

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

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

Abstract. ~~Functional group (FG) analysis provides a means by which functionalization in organic aerosol can be attributed to the abundances of its underlying molecular structures. However, performing this attribution requires additional, unobserved details about the molecular mixture to provide constraints in the estimation process.~~ We present an approach for conceptualizing functional group (FG) measurements of organic aerosol in terms of its functionalized carbon atoms. This reformulation facilitates estimation of mass recovery and biases in popular carbon-centric metrics that describe the extent of functionalization (such as oxygen to carbon ratio, organic mass to organic carbon mass ratio, and mean carbon oxidation state) for any given set of molecules and FGs analyzed. Furthermore, this approach allows development of parameterizations to more precisely estimate the organic carbon content from measured FG abundance. We use simulated photooxidation products of α-pinene secondary organic aerosol previously reported by Ruggeri et al., (*Atmos. Chem. Phys.*, 16, 4401–4422, 2016) and FG measurements by Fourier Transform Infrared (FT-IR) spectroscopy in chamber experiments by Sax et al. (*Aerosol Sci. Tech.*, 39, 822–830, 2005) to infer the relationships among molecular composition, FG composition, and metrics of organic aerosol functionalization. We find that for this simulated system, $\sim$80% of the carbon atoms should be detected by FGs for which calibration models are commonly developed, and $\sim$7% of the carbon atoms are undetectable by FT-IR analysis because they are not associated with vibrational modes in the infrared. Estimated biases due to undetected carbon fraction for these simulations are used to make adjustments in these carbon-centric metrics such that model-measurement differences are framed in terms of unmeasured heteroatoms (e.g., in hydroperoxide and nitrate groups for the case studied in this demonstration). The formality of this method provides framework for extending FG analysis to not only model-measurement but also instrument intercomparisons in other chemical systems.

1 Introduction

Organic aerosols are complex mixtures of thousands of different types of compounds that vary in structure and physicochemical properties. This diversity poses challenges for comprehensive characterization, even while estimates of overall mass abundance and its contributing factors are still desirable. Functional group (FG) analysis is an approach that presents a level of characterization that provides a bridge between full molecular speciation, which is useful for precisely tracking specific classes of physical and chemical transformations, and elemental composition, which is useful for mass closure analysis. FGs are structural units in molecules that describe important condensed-phase interactions that contribute to properties like volatility and hygroscopicity, and FG analysis provides information useful for overall organic mass quantification and its apportionment by source class in past studies (e.g., Russell et al., 2011). FGs are also central to understanding reactivity and resulting chemical transformations, and their characterization by measurement and in model simulation can provide a method of evaluating our understanding of functionalization (i.e., through bonding with heteroatoms) in organic aerosol mixtures. However, studies on this topic have thus far been very limited on account of challenges in quantitative characterization of FGs, which requires either advanced algorithms (e.g., Takahama et al., 2013; Ruthenburg et al., 2014; Takahama and Dillner, 2015) for spectral interpretation or derivitization steps (e.g., Dron et al., 2010; Aimanant and Ziemann, 2013) for chemical analysis. In anticipation of continued progress in analytical technology, Ruggeri and Takahama (2016) and Ruggeri et al. (2016) introduced a method for harvesting FG information from molecularly speciated measurements (e.g., gas chromatography-mass spectrometry, GC-MS; Rogge et al., 1993) and chemically explicit model simulation (e.g., Master Chemical Mechanism, MCMv3.2; Jenkin et al., 1997, 2003; Saunders et al., 2003).

In this study, we build upon the work by Ruggeri et al. (2016) to further improve our capability for model-measurement intercomparison using FG analysis. Ruggeri et al. (2016) compared changes in relative molar abundances of FGs in chamber experiments measured by Fourier Transform Infrared (FT-IR) spectroscopy against composition simulated with a chemically explicit gas-phase reaction mechanism coupled to a gas/particle (G/P) partitioning module. As molar FG composition is directly obtained from measured FT-IR absorbances, this is a sensible metric used to track changes in chemical composition and has been used in other studies (e.g., Camredon et al., 2007). However, estimating FG contributions to carbon-centric metrics more commonly used to characterize organic aerosol oxidation or mass yields, such as organic carbon (OC) and organic matter (OM) mass, OM/OC mass ratios, atomic ratios, and mean carbon oxidation state (Russell, 2003; Aiken et al., 2008; Kroll et al., 2011, 2015) is not straightforward. Central to this task is understanding which fraction of carbon atoms are “detected” by measurement of any given set of FGs, and estimating the overall carbon abundance from FGs without multiply counting the polyfunctional carbon atoms.

Some of these metrics have been calculated from FT-IR measurements by previous researchers based on assumptions regarding the underlying molecular structure (e.g., Allen et al., 1994; Maria et al., 2003; Reff et al., 2007; Russell et al., 2009; Chhabra et al., 2011). For instance, Chhabra et al. (2011) assumed bonding configurations in secondary organic aerosol (SOA) products to be consistent to the parent volatile organic compound (VOC) to estimate the carbon content from measured FG abundance. Ranney and Ziemann (2016) also use the number of carbon atoms in the parent VOC to normalize FG concentrations reported for SOA mixtures. Russell (2003) introduced a functional group index (FGI) to conceptualize how OM/OC ratios varies according to chain length and functionalization for specific sets of compound classes, and provided an evaluation from mass spectrometry measurements that comprised up to 10% of the total OM mass. Using results from numerical simulation of SOA formation, we now describe methods for estimating carbon content based on molecular parameters that describe the underlying mixture composition consisting of a diverse set of polyfunctional compounds, and a means of examining dependence of carbon-centric metrics on composition without invoking knowledge about molecular chain lengths, which is not well characterized by FG analysis. The benefit of developing a systematic approach is that we can precisely understand the achievable mass recovery, and biases incurred on the calculated O/C and OM/OC for a given set of molecules and FGs analyzed (when ~~extraction efficiencies are not invoked~~ chemical extraction is not required, OM mass recovery is primarily dependent on the completeness of FG calibration models constructed). These estimates may then be used to propose mixture-specific adjustments to facilitate more direct intercomparisons with other data. This work will focus on FG abundances obtained by FT-IR measurements, but many aspects are generalizable to other types of FG analysis (e.g., Dron et al., 2010; Ranney and Ziemann, 2016).

The objective described above is addressed in this work by 1) conceptualizing SOA as a collection of carbon atoms that are functionalized in different ways, and 2) the FT-IR as a tool that measures some subset of such functionalized carbon structures. These "carbon types" can be used to calculate the OM properties described above, and gives rise to observed FGs in measurement. Carbon type representation of complex mixtures has a strong precedent in the study of organic chemistry in the atmosphere. For example, the Carbon Bond Mechanism (Whitten et al., 1980) defines chemical reaction schemes according to reactivity of carbon atoms classified according to functionality, without regard to membership in a molecule. The "carbon vector" in GECKO-A (Aumont et al., 2005) is a description of functionalized carbon types and retains information regarding transformations in functionalization (while a separate connectivity matrix tracks transformation in the carbon skeleton upon accretion or fragmentation). In the commonly used volatility basis set (VBS), changes in carbon mass are conserved according to functionalization by oxygen, nitrogen, or overall carbon oxidation state (Kroll et al., 2011, 2015; Donahue et al., 2012; Chuang and Donahue, 2016). Quantitative analysis of additional "groups" that describe the underlying skeletal (e.g., ring, aromatic, or unsaturated) structures that change with fragmentation and accretion reactions (Kroll et al., 2011) have not been sufficiently advanced by FG analysis to provide complete estimates of mean molecular size and other aerosol properties that govern volatility and solubility (Zuend et al., 2008). However, past precedents mentioned above indicate that classification of carbon atoms according to extent of functionalization may have merit in harmonizing observations with model representations for calculating common mixture characteristics of OM.

In this work, we illustrate how measured FGs can be related to properties of various carbon types comprising a diverse set of polyfunctional molecules. We use the proposed relationships to determine which carbon types are measured according to FGs included in calibration models, and biases resulting from partial analysis of the different carbon types in the mixture. For illustration, α-pinene gas-phase photooxidation simulation in the presence of $\mathrm{NO_x}$ with G/P partitioning is analyzed and compared against chamber experiments upon which the simulations were based. We will assume a perfect calibration where we assume flawless knowledge of the bond abundance to isolate biases due to measured and unmeasured carbon types. Such a scenario is obviously not physically achievable, but serves as a convenient reference by which we can proceed with a meaningful model-measurement comparison.

2 Methods

After describing our data set in Section 2.1, we introduce a few relationships among FG, atomic composition, and carbon types in Section 2.2. We then describe how we can estimate whether a particular carbon type is detected by FT-IR based on the set of FG calibrations used and properties that we calculate as a result in Section 2.3.We then present methods for actually estimating the number of polyfunctional carbon atoms from FG abundance to minimize multiple counting in Section 2.4. The code and software used in this and previous manuscripts are made available under the GNU Public License (Appendix A).

2.1 Data set

We focus this analysis on a specific simulation scenario of Ruggeri et al. (2016) in which comparison of model results to reference measurements had the smallest discrepancy according to relative molar abundance of FGs, until model-measurement agreement diverged on what was attributed to the role of heterogeneous chemistry and aging not implemented in the model. To briefly describe the simulation, the MCMv3.2 gas-phase chemistry module generated by the Kinetic Pre-Processor (Sandu and Sander, 2006; Henderson, 2015) was coupled with a gas/particle organic absorptive partitioning scheme via operator splitting (Yanenko, 1971). The SIMPOL.1 group contribution model (Pankow and Asher, 2008) was used to estimate the equilibrium vapor pressure for individual molecules, and the dynamics of mass transfer to a monodisperse particle population were simulated using LSODE (Livermore Solver for Ordinary Differential Equations; Radhakrishnan and Hindmarsh, 1993). Wall losses of particles and semivolatile volatile organic compounds (SVOCs) were neglected. The scenario we further analyze for this study was defined by initial α-pinene and NO_x concentrations of 300 and 240 ppb, respectively. The relative humidity was fixed at 61%, which influenced the rate of HO_2 radical self reaction to form hydrogen peroxide, but water uptake and influence on G/P partitioning was not considered. The light intensity was fixed (Saunders et al., 2003) to be consistent with experimental conditions. This scenario was labeled the “APIN-lNOx” simulation. In this work, we will refer to this as the APIN simulation, as we discuss none of the other scenarios and thus eliminate the need for an additional modifier to the label. To focus on a particular mixture, we select a reference period as the apex in SOA concentration occurring at 9.3 hours (labeled as $t_{\max \mathrm{SOA}}$) of the 22 hour simulation as used by Ruggeri et al. (2016) to examine molecular contributions to overall SOA mass and FG abundance. With detailed knowledge of molecular structure and composition in this simulation, we apply the analysis described in Sections 2.2–2.4.

The conditions for the simulations described above were selected to mimic chamber experiments in which FG composition was measured by Sax et al. (2005). Sax et al. (2005) collected particles between 86 and 343 nm onto (infrared-transparent) zinc selenide crystals by impaction, and samples were analyzed immediately afterward to minimize storage artifacts. Samples were scanned rapidly to minimize evaporative losses in the FT-IR sample compartment. Sax et al. (2005) report that repeated analysis of the same samples by FT-IR yielded consistent results, suggesting robustness in reported values. Samples collected during 3.1–4.2 hours and 17.6–21.6 hours (which we label as “4h” and “21h”, respectively) were selected by Ruggeri et al. (2016) for comparison against model simulation for the corresponding periods, and we will follow this convention here.

Only relative metrics are used as Sax et al. (2005) reported measurements in mole fractions of FGs, and the simulations do not include wall losses of particles and SVOCs that affect overall estimates of yield. Neglecting compound-specific SVOC deposition to walls may further incur biases in relative compositions as raised by Ruggeri et al. (2016), but for this conceptual study we neglect its effect as its parameters are not precisely known.

2.2 Definitions

The molar abundance of molecules $\boldsymbol{n}_{\text{molec}} = [n_i]$ in a mixture (consisting of a set of molecules denoted by $\mathcal{M}$) can be related to FG abundance $\boldsymbol{n}_{\text{group}} = [n_j]$ (for each FG in $\mathcal{J}$) obtained by FT-IR — or other means — by invoking a group composition matrix $\mathbf{X} = [x_{ij}]$, which describes the FG makeup of each molecule. Using scalar notation, we write:

$$n_j = \sum_{i \in \mathcal{M}} n_i x_{ij} \quad \forall\, j \in \mathcal{J}\ . \tag{1}$$

n_j is the observed quantity from measurement, and represents the sum of FG composition of molecules weighted by their molar abundance.

A statement of atom balance is enabled by the group-atom matrix $\boldsymbol{\Lambda} = [\lambda_{aj}]$ (Takahama et al., 2013) by relating n_j to the atomic abundance $\boldsymbol{n}_{\text{atom}} = [n_a]$ in the mixture:

$$n_a = \sum_{j \in \mathcal{J}} \lambda_{aj} n_j\ , \tag{2}$$

However, the fact that the same polyfunctional carbon atom can be associated with several FGs poses challenges for reasoning out $\lambda_{\text{C},j}$ for carbon. Therefore, we introduce a carbon type matrix $\mathbf{Y} = [y_{ik}]$ that enumerates the composition of each molecule in terms of specific number of carbon types, and a carbon-group matrix $\boldsymbol{\Theta} = [\theta_{kj}]$ that relates each carbon type to its unique structure of functionalization.

A statement of FG balance can be constructed from the carbon type matrix, carbon-group matrix, and group composition matrix:

$$\sum_{k \in \mathcal{C}} y_{ik} \theta_{kj} = x_{ij} \quad \forall\, i \in \mathcal{M}, j \in \mathcal{J}\ . \tag{3}$$

Conversely, a statement of carbon type balance can be made by introducing a matrix, $\boldsymbol{\Phi} = [\phi_{jk}]$ from which carbon type abundance can be obtained with FG abundance to construct a statement of carbon type balance:

$$y_{ik} = \sum_{j \in \mathcal{J}} x_{ij} \phi_{jk} \quad \forall\, i \in \mathcal{M}, j \in \mathcal{J}\ . \tag{4}$$

A minimal illustration for two simple molecules, ethane and ethanol, are shown in Fig. 1. Symbols are tabulated in Table B1. Explanation of additional arrays $\boldsymbol{\Lambda}$ (atom-group matrix), $\boldsymbol{\zeta}$ (carbon oxidation state vector), and $\boldsymbol{z}$ (oxidation state contribution vector) completing the atom and oxidation state balance follow below. In contrast to concise expressions used in Figure 1, we continue with use of scalar notation below to more conveniently invoke element-wise, row-wise, and column-wise summations, but will return to array notation for describing solutions to system of equations (Section 2.4).

In our APIN mechanism, there are 327 molecules, 22 FGs, and 41 carbon types (Figure 2), though several are associated with radical structures or unusual structures that are not found in the most abundant compounds. These do not contribute to the organic aerosol mass, but is included for a complete description of the APIN mechanism. Furthermore, while the equalities introduced in Figure 1 are formulated to hold at the level of individual molecules, we demonstrate their application in describing the underlying relationships in molecular mixtures.

The carbon type matrix provides a conceptual relationship for relating FGs to number of carbon atoms in a mixture (equation 2 for carbon is also restated on the right hand side):

$$n_{\mathrm{C}} = \sum_{i \in \mathcal{M}} \sum_{k \in \mathcal{C}} n_i y_{ik} = \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{J}} n_i \lambda_{\mathrm{C},j} x_{ij} \tag{5}$$

and we can see from equations 4 and 5 that $\lambda_{\mathrm{C},j}$ is equivalent to the column-wise summation of ϕ_{jk}.

$$\lambda_{\mathrm{C},j} = \sum_{k \in \mathcal{C}} \phi_{jk} \quad \forall\, j \in \mathcal{J}. \tag{6}$$

Previous values for λ_{C} are shown in Table 1. The atomic abundance for each carbon type k is calculated as $n_{ka} = \sum_{j \in \mathcal{J}} \lambda_{aj} \theta_{kj}$, as follows from equation 3 and 2.

The mean carbon oxidation state can be estimated from: 1) y_{ik} through the oxidation state $\boldsymbol{\zeta} = [\zeta_k]$ specific to carbon type, and 2) x_{ij} and individual FG contributions $\boldsymbol{z} = [z_j]$ to carbon oxidation state:

$$\overline{\mathrm{OS}}_{\mathrm{C}} = \frac{1}{n_{\mathrm{C}}} \sum_{i \in \mathcal{M}} \sum_{k \in \mathcal{C}} n_i y_{ik} \zeta_k = \frac{1}{n_{\mathrm{C}}} \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{J}} n_i x_{ij} z_j \tag{7}$$

From equation 3, we can see that ζ_k and z_j are related through the following equality:

$$\zeta_k = \sum_{j \in \mathcal{J}} \theta_{kj} z_j \quad \forall\, k \in \mathcal{C} \tag{8}$$

All elements in equation 3 can be known precisely for any set of molecules $\mathcal{M}$ from the chemometric patterns and atom-level validation described by Ruggeri and Takahama (2016), and is summarized in Section S1. Furthermore, the FGs included in the APIN system are all those which are defined by association only to single carbon atoms (e.g., alcohol, carboxylic, methylene groups). Methods for extending this analysis to FGs containing multiple carbon atoms (e.g., anhydride, ester, and organic peroxide groups) are described in Section S2. Solution methods for ϕ_{jk} and $\lambda_{\mathrm{C},j}$ are presented in 2.4.

2.3 Theoretical mass recovery and estimated properties

This section describes methods for determining if the carbon type is detected by FT-IR and how relationships introduced in Section 2.2 can be modified for a more direct comparison with measurements. The main idea is to consider only the subset of carbon atoms which is bonded to any of the FGs measured in a given experiment, and analyze properties only for those carbon atoms as to what is the achievable degree of characterization of the SOA.

Given a set of FG which are measured $\mathcal{J}^* \subseteq \mathcal{J}$ and the corresponding subset of carbon atoms $\mathcal{C}^* \subseteq \mathcal{C}$ which only contain these FGs, we can estimate the number of carbon atoms measured from a modification of equation 5:

$$n_{\mathrm{C}}^* = \sum_{i \in \mathcal{M}} \sum_{k \in \mathcal{C}^*} n_i y_{ik} = \sum_{i \in \mathcal{M}} \sum_{k \in \mathcal{C}} n_i y_{ik} \cdot \mathrm{sgn}\left(\sum_{j \in \mathcal{J}^*} \theta_{kj}\right) . \tag{9}$$

sgn is the signum function, which will return 0 when its argument is 0 (no FGs associated with carbon type k are in the measured set) and 1 when its argument is positive (one or more FGs belong to the measured set). The total carbon recovery is calculated as $n_{\mathrm{C}}^*/n_{\mathrm{C}}$.

We consider three sets of FGs for $\mathcal{J}^*$. Set1 = {aCH, aCOH, COOH, ketone and aldehyde carbonyl, $\mathrm{CONO_2}$}, and comprises FGs reported by Sax et al. (2005) and many others (e.g., Maria et al., 2003; Coury and Dillner, 2008; Russell et al., 2009; Day et al., 2010). Set2 = Set1 + {eCH, hydroperoxide, peroxyacyl nitrate}, and comprises Set1 and three additional FGs that are not commonly reported for OM characterization but have medium to strong absorption bands in the mid-infrared wavelengths (Appendix C) (not inclusive) and relevant for this system. The set labeled as Full comprises all groups present in OM, including quaternary and tertiary sp^2 carbon (carbon atoms that are only bonded to other carbon atoms) that accounts for 7% of the mass in the APIN simulation at $t_{\max \mathrm{SOA}}$, and also the remaining groups (Figure 2) that accounts for <1% of the remaining mass.

We can estimate OM as the sum of elements multiplied by their respective molecular weights using equation 2. Atomic ratios are calculated as n_a/n_{C} for all heteroatoms $a = \{\mathrm{H}, \mathrm{N}, \mathrm{O}\}$ (S is not included in this chemical mechanism, but this principle can be extended for mechanisms that include it):

$$n_a^* = \sum_{j \in \mathcal{J}*} \lambda_{aj} n_j \ . \tag{10}$$

Atomic ratios are calculated as n_a^*/n_{C}^*

To estimate the mean carbon oxidation state, we can replace n_{C} with n_{C}^* and sum over $\mathcal{J}^*$ instead of $\mathcal{J}$ in equation 7 by corollary with equation 9:

$$\overline{\mathrm{OS}}_{\mathrm{C}} \approx \frac{1}{n_{\mathrm{C}}^*} \sum_{j \in \mathcal{J}^*} z_j n_j \ . \tag{11}$$

2.4 Estimation of carbon abundance

In this section, we describe methods for estimating n_{C} from measured abundance of FGs. The main objective is to arrive at a set of coefficients $\hat{\lambda}_{\mathrm{C}}$ that, when multiplied by FG abundance n_j for measured FGs $\mathcal{J}^*$, provides an estimate $\hat{n}_{\mathrm{C}}^*$ that does not count multiples of the same carbon atoms which are attached to the suite of FGs analyzed:

$$\hat{n}_{\mathrm{C}}^* = \sum_{j \in \mathcal{J}^*} \hat{\lambda}_{\mathrm{C},j} n_j \ . \tag{12}$$

The use of the hat over a symbol denotes a statistically estimated quantity.

It is convenient to continue discussion of solutions to a system of equations in array notation (similar to what is used in Figure 1). Let $\mathbf{Y} = [n_i y_{ik}]$, $\mathbf{X} = [n_i x_{ij}]$, $\mathbf{\Theta} = [\theta_{kj}]$, $\mathbf{\Phi} = [\phi_{jk}]$, $\lambda_\mathrm{C} = [\sum_{k \in \mathcal{C}} \phi_{jk}]$, and $\boldsymbol{n}_\mathrm{C} = [\sum_{k \in \mathcal{C}} y_{ik}]$. The FGs and carbon type abundances can be written as $\mathbf{Y\Theta} = \mathbf{X}$. The most obvious solution is to take the generalized or Moore-Penrose inverse, $\hat{\mathbf{\Phi}} = \mathbf{\Theta}^+$. In the example illustrated in Figure 1, the solution to $\mathbf{\Phi} = \mathbf{\Theta}^{-1}$ and $\boldsymbol{\lambda}_\mathrm{C}$ (a row of $\mathbf{\Lambda}^T$) using such an approach is provided. The elements of $\mathbf{\Phi}$ satisfy the carbon type balance (equation 4) but are not required to be non-negative, but their summation across rows (equation 6) yields values for $\boldsymbol{\lambda}_\mathrm{C}$ that corresponds to the number of carbon atoms per FG associated with them. While exact solutions can be found for this illustration because $\mathbf{\Theta}$ is square (i.e., the number of carbon types equals the number of types of FGs), the pseudo-inverse solution will not be meaningful in a more general case as the number of ways in which FGs are arranged on carbon atoms exceeds the number of measured FG used for discrimination. $\boldsymbol{\lambda}_\mathrm{C}$ may also not correspond to a physically interpretable quantity in such instances, as a single set of coefficients are insufficient to estimate the exact abundances of carbon atoms under these circumstances.

Therefore, while carbon types are a useful concept to describe the underlying representation of functionalized organic compounds, it is generally not possible to retrieve the exact abundance of each carbon type from FG measurements. To arrive at an approximate solution for estimation of the total carbon atoms without discrimination of carbon types, we consider the three approaches described below.

First, we consider each carbon type in isolation ("COUNT" method) and average the reciprocal of measured FGs per carbon enumerated for each carbon type.

$$\hat{\lambda}_{\mathrm{C},j} = \frac{1}{|\mathcal{C}_j|} \sum_{k \in \mathcal{C}_j} \frac{1}{\sum_{j' \in \mathcal{J}^*} \theta_{kj'}} \tag{13}$$

$|\cdot|$ denotes the cardinality of (i.e., number of elements in) a set and $\mathcal{C}_j$ is the set of carbon types in which FG j appears, and is the origin of the dependence of λ_C on j. The main premise of this approach is to apportion fractional units of carbon to each measured FG such that their sum equals unity. The rationale can be supported by the illustration (Figure 1) in which 1/3 for $\boldsymbol{\lambda}_\mathrm{C}$ reflects the number of measured FGs attached to each carbon atom.

In the second approach ("COMPOUND" method), we find $\mathbf{\Phi}$ that corresponds to the least squares solution to the following equation:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{\Phi}} \,. \tag{14}$$

$\hat{\boldsymbol{\lambda}}_\mathrm{C}$ is found by row-wise summation of $\hat{\mathbf{\Phi}}$ (equation 6) (which is also equivalent to solving for $\hat{\boldsymbol{\lambda}}_\mathrm{C}$ directly in the reduced expression, $\hat{\boldsymbol{n}}_\mathrm{C} = \boldsymbol{X}\hat{\boldsymbol{\lambda}}_\mathrm{C}$). Given the wide range of possibilities in composition, we set molar abundances to unity such that each compound within each group (SVOC) is uniformly weighted. We average over carbon types present in molecules relevant to certain mixture classes with uniform weighting such that the derived coefficients are not overly specific to any particular mixture.

In the third approach ("MIXTURE" method), we reformulate $\mathbf{Y} = [n_{mi} y_{ij}]$ and $\mathbf{X} = [n_{mi} x_{ij}]$ such that its rows contain the FG abundance of the mixture of each time step t_m of the APIN simulation, and $\boldsymbol{\lambda}_{\mathrm{C}}$ is found by fitting $\mathbf{X}$ to $\boldsymbol{n}^*_{\mathrm{C}}$, the time series of carbon atom concentration in the condensed-phase at each time step. For MIXTURE, we use a constrained least squares approach where the values of the regression coefficients are bounded between 0 and 1 as the coefficients for FGs with low abundance (e.g., eCH and CONO2) are not well constrained (the solution is insensitive to their values).

Numerical details aside, the main differences among the three are the data sets used for estimation. COUNT uses information from $\boldsymbol{\Theta}$ only (defined for the FGs in the APIN mechanism), COMPOUND uses carbon type abundances in compounds (limited to SVOCs in the APIN mechanism), and MIXTURE uses mixture information of the condensed-phase (from different periods in the APIN simulation). The resulting differences in estimates of $\hat{\lambda}_{\mathrm{C}}$ are largely due to weighting of FGs associated with each carbon type: each type receiving equal weight (COUNT), by frequency of occurrence in SVOCs (COMPOUND), and by abundance in SOA formed in the APIN simulation (MIXTURE). While the COUNT method is physically significant at the level of individual carbon atoms, the representativeness of estimated values for use in mixtures can vary according to composition. Direct fitting methods, on the other hand, may lead to insignificant coefficients from under-represented or redundant FGs, or be overly specific such that they cannot be generalized to other systems. Therefore, the results from all three methods are evaluated to explore the range of plausible values.

Each of the solutions produces a series of irrational numbers (due to the multiplicitous configurations of FGs on carbon atoms) that may be overly precise for the data set used for estimation. As later shown, we will also adjust the COUNT solutions to rational values of {1/4, 1/3, 1/2, 1} (with exception for $\lambda_{\mathrm{C,aCH}}$ which we fix to a value of 0.45 as explained in Section 3.3), and we will refer to this as the "NOMINAL" solution. For the COMPOUND and MIXTURE methods, FGs and carbon types with a unique (one-to-one) correspondence (e.g., carbon atoms associated with carboxylic acid and ketonic and aldehydic carbonyl groups) are excluded from the fitting, as their coefficients are known unambiguously. Evaluations of estimates are expressed as a ratio of the estimate over the reference value: $\hat{n}^*_{\mathrm{C}}/n^*_{\mathrm{C}}$. We remark that we focus on harvesting information from the APIN simulation results only, but these methods can (and should) be applied to study abundances in molecular speciation data from chamber experiments under different oxidation and environmental conditions (e.g., Yu et al., 1999; Glasius et al., 2000) in future work.

3 Results

We first describe the APIN simulation results of Ruggeri et al. (2016) recast in terms of abundance of carbon types in Section 3.1. We then describe mass recovery and biases in property estimates due solely to unmeasured carbon atoms in Section 3.2. In Section 3.3, we describe results from applying different methods for estimating carbon abundance from measured FGs. Finally, in Section 3.4, we present estimates of properties from FG measurements and compare to model simulations.

3.1 Evolution of carbon types

The time series of carbon type abundance is shown by its contribution fraction for each time period in Figure 3, and the carbon type composition of the most abundant molecules at $t_{\max\mathrm{SOA}}$ is depicted in Figure 4. Descriptions for the carbon types found in $t_{\max\mathrm{SOA}}$ are shown in Figure 2. We observe that changes in carbon type composition is rapid within the first four hours, but generally changes much more slowly after this period. Many of the dominant carbon types are generally similar between the gas and aerosol phases and include: methyl ($\mathrm{CH_3}$), methylene ($\mathrm{CH_2}$), ketone, primary alcohol, and secondary alcohols, acid (COOH), hydroperoxides, and peroxyacyl nitrate groups. However, the order of abundance is different between phases — for instance, the peroxyacyl nitrate is more abundant in the gas phase (carbon type 10; Figure 2). As visualized in Figure 4 and described by Ruggeri et al. (2016), the molecular abundance is dominated by a small number of polyfunctional compounds (out of the [200] compounds in the mechanism), so their carbon types are weighted heavily in the overall carbon type composition.

3.2 Theoretical mass recovery and property estimation

The ordered contribution to mass recoveries of OC and OM for the most dominant carbon types at $t_{\max\mathrm{SOA}}$ are displayed in Figure 5. Greater than 99.9% of the OC and OM mass is accounted for by 15 carbon types during this period, while more than 20 compounds are required to reconstruct aerosol OC mass with $>$99.9% recovery (Figure 4)). Mass recovery with Set1 is on the order of 80%. The fraction of OC estimated by FT-IR relative to OC measured by thermal optical methods are often within a similar range (e.g., Maria et al., 2003; Ruthenburg et al., 2014). With additional bonds in Set2, 93% carbon recovery is achieved. The unmeasured carbon types are quaternary and tertiary sp^2 carbon that are bonded to C-bonds only, and together comprise 7% of the OC (Full case).

Going from Set1 to Set2, the increase in fraction of recovered OM is greater than recovered OC because of the hydroperoxide and peroxyacyl nitrate mass is much greater than the mass of carbon bearing these FGs. The resulting effect on estimated properties is shown in Figure 6. H/C recovery is high for Set1 already, but we are missing the oxygen from hydroperoxide and peroxyacyl nitrate. eCH is small. N/C is very small (low $\mathrm{NO_x}$ conditions). OM/OC can be off by 0.2. Even with nearly full mass recovery, ratios are often inflated by a small amount on account of the unmeasured carbon (i.e., $n_\mathrm{C}^* \leq n_\mathrm{C}$).

The carbon oxidation state distribution and recoverable portions for $t_{\max\mathrm{SOA}}$ are shown in Figure 7a. This figure visually reinforces the abundance of methyl carbons ($\mathrm{CH_3}$, $\mathrm{OS_C}$ = -3), methylene carbons ($\mathrm{CH_2}$, $\mathrm{OS_C}$ = -2) discussed above, though there are other carbon types contributing to the $\mathrm{OS_C}$ = -2 category (Figure 2). The unmeasurable carbon types with FT-IR are those with $\mathrm{OS_C}$ = 0, which are the quaternary and tertiary sp2 carbon (carbon types which are measurable in the $\mathrm{OS_C}$ = 0 category have a balance of negative and positive values from aCH and electronegative heteroatoms). The value of the additional FGs in Set2 are for characterization of oxidizing FGs (hydroperoxide and peroxyacyl nitrate) that on carbon atoms with overall oxidation states of 1 and 3. Estimates of the mean $\overline{\mathrm{OS}}_\mathrm{C}$ is shown in Figure 7, panel (b). We can see that the bias in estimation for neglecting hydroperoxide and peroxyacyl nitrate is not as great as for the O/C ratio, since the $\mathrm{OS_C}$ is determined by the atom and bond connected to the carbon atom directly, and the rest of the multiple oxygen atoms in the FG are not considered. The 2O/C-H/C estimate commonly used with elemental analysis will lead to a slight overestimation of the $\overline{\mathrm{OS}}_{\mathrm{C}}$ in the event that oxygen single-bonded to carbon (hydroxyl and hydroperoxide groups) exist in large abundance proportionally to double-bonded carbonyl groups (Kroll et al., 2011).

3.3 Estimation of carbon abundance

Table 2 summarizes the new values for $\hat{\lambda}_{\mathrm{C}}$ obtained by the different estimation methods described in Section 2.4. Comparison of $\hat{n}^*_{\mathrm{C}}$ estimated using these values against n^*_{C} in individual compounds is shown in Figure 8, and the comparison of $\hat{n}^*_{\mathrm{C}}$ and n^*_{C} in overall aerosol mixtures at different time periods in the APIN simulation is shown in Figure 9.

Values for $\hat{\lambda}_{\mathrm{C}}$ are roughly similar among estimation methods, with exception to the MIXTURE estimate. Overall, we find that the coefficient for aCH is close to but less than the often assumed value of 0.5 (Table 1), which can play an important role on account of the abundance of aCH bonds and carbon types associated with aCH. For the MIXTURE estimate, $\hat{\lambda}_{\mathrm{C,aCH}} = 0.5$ but is balanced by exceptionally small coefficients for aCOH and hydroperoxide. This combination of coefficients essentially downweights the contributions from carbon types associated with aCH and hydroperoxide, which we know to be present in abundance (within top 6 for the APIN simulation at $t_{\mathrm{max\,SOA}}$, but remains significant throughout the simulation as seen in Figure 3). Therefore, we conclude that the estimates obtained for this fit are statistically convenient but less physically relevant than the other estimates. For the NOMINAL case, we fix the aCH to $\lambda_{\mathrm{C,aCH}} = 0.45$ and the rest to the nearest rational numbers.

For individual compounds, we note that using either Set1 and Set2 reproduce n^*_{C} with similar biases on average: 11% for COUNT and within 4% for the others. COUNT underestimates n^*_{C} in large compounds with lower oxidation states containing many aCH groups, because of the low estimate of $\hat{\lambda}_{\mathrm{C,aCH}}$. COMPOUND reproduces n^*_{C} well because this is the data set COMPOUND was fit to, but MIXTURE also does well. The NOMINAL solution also does well, but largely owing to the $\lambda_{\mathrm{C,aCH}}$ adjustment.

For reproducing mixture composition, trends in biases are similar to individual compounds, with underestimation by as much as 18% for COUNT and within 7% for the other estimation methods. MIXTURE performs the best because this is the data set it was fitted to, but we see that the COMPOUND and NOMINAL are also acceptable. There is generally a trend toward increasing $\hat{n}^*_{\mathrm{C}}/n^*_{\mathrm{C}}$ over the duration of the simulation, which indicates an evolving relationship between FGs and carbon abundance with mixture composition. Time-dependent (i.e., mixture-specific) estimates of λ_{C} may be warranted when the change in composition becomes more significant.

We therefore conclude that errors for estimation of n^*_{C} can be quite low and are well below 10% according to our evaluation. Even a 10% error in estimation of n^*_{C} will lead to a 9% error in the estimation of any individual atomic ratio, and 5% estimation in the OM/OC ratio (Appendix D). Therefore, in applying the NOMINAL coefficients to measured values of FGs under conditions upon which the APIN simulations were based (Section 3.4), we discuss deterministic explanations for model-measurement discrepances with less consideration toward statistical estimation error of n^*_{C}.

3.4 Comparison with measurements

In this section, we discuss O/C, OM/OC, and $\overline{\mathrm{OS}}_\mathrm{C}$ estimated from measurements ending at hours 4 and 21 and APIN simulation results integrated over the same periods (Figure 10). We label the interpretation of measurements with previous estimates of λ_C (Table 1) as "MEAS-PREV", measurements with revised estimates of λ_C (Table 2) as "MEAS-NOM", simulation results using FGs from Set1 as "SIM-SET1", and full simulation results as "SIM-FULL"; further adjustments are made for the last three estimates as justified next. In Section 3.2, we presented an estimate of mass recovery ($n^*_\mathrm{C}/n_\mathrm{C}$) and how this led to biased estimates of atomic ratios and OM/OC ratio. In Section 3.3, we also showed that we can derive estimates of λ_C such that errors in estimation of n^*_C was small (i.e., $\hat{n}^*_\mathrm{C}/n^*_\mathrm{C}$ near unity). Therefore, for the following comparisons, we neglect the latter error and correct biases due to carbon mass recovery by using our best estimate of n_C, rather than n^*_C, as the normalization factor. The proportion of detected carbon to make this correction is obtained from SIM-SET1 in which the same FGs as measurements are used. While the adjustment is only approximate on account of differences in the real experimental system and model simulation, it reduces systematic biases in carbon-centric metrics as described in Section 3.2 such that deviations from true ratios can be largely attributed to the unmeasured heteroatoms. For MEAS-NOM, the atomic ratio is then estimated as $n^*_\mathrm{a}/n_\mathrm{C} = n^*_\mathrm{a}/n^*_\mathrm{C} \times (n^*_\mathrm{C}/n_\mathrm{C})_{\text{SIM-SET1}}$ and the OM/OC and $\overline{\mathrm{OS}}_\mathrm{C}$ by similar adjustment. MEAS-PREV remains unadjusted to be used as a reference estimated without prior knowledge about the underlying molecular structures of the SOA products.

First, we remark on differences for estimated metrics from two sets of coefficients applied to the same FG measurements. MEAS-PREV overestimates the n^*_C compared to MEAS-NOM by 21-28% on account of higher λ_C coefficients used in the former. However, the uncorrected bias due to lower mass recovery of carbon is approximately the same magnitude, and ultimately leads to ratioed values (O/C, H/C, OM/OC, $\overline{\mathrm{OS}}_\mathrm{C}$) similar to MEAS-NOM. While it is not clear that λ_C derived in this work accurately represents the true mixture, we posit that the degree of functionalization characterized by the new estimate is likely to be more representative for the product mixture after successive oxidation of the APIN, rather than APIN itself (as assumed by MEAS-PREV). Chhabra et al. (2011) report O/C and H/C estimates from FT-IR using coefficients of MEAS-PREV and found that they were within range of AMS values; this is possibly due to the offsetting of errors as demonstrated here. In further discussion, we will discuss the interpretation of observations based on MEAS-NOM.

MEAS-NOM and SIM-SET1 are the two estimates intended to provide the most direct comparison between experiment and numerical simulation. While the discrepancy in carbonyl and carboxyl groups at 4 hours is only 2% and 3% in mole fraction, respectively (Ruggeri et al., 2016), this leads to an overall discrepancy of 0.16 for O/C and 0.2 for OM/OC. Since aCOH, carbonyl, and COOH groups are a larger contributor to the mass relative to the aCH group, discrepancies in molar abundance of oxygenated FGs are magnified when represented in OM/OC ratios and can have a non-negligible influence on interpretation of mass yields. After 21 hours, the difference is 0.38 in O/C and 0.48 in OM/OC. Ruggeri et al. (2016) attributed the apparent divergence to mechanisms not included in the model. Oligomerization was not considered a likely candidate as this process not expected to contribute to increased oxygenation reported by FT-IR. Condensed-phase photolysis can lead to conversion of hydroperoxides to carbonyls (some of which are lost to the vapor phase as more volatile molecules) (Epstein et al., 2014), but even a hypothetical full molar conversion is insufficient to explain the model-measurement differences in carbonyl groups (Ruggeri et al., 2016). Other missing mechanisms may include autoxidation (Crounse et al., 2013) which can produce extremely low volatility (ELVOC; Ehn et al., 2014) or highly oxygenated molecules (HOM; Tröstl et al., 2016) in the gas phase, or radical reactions in the condensed phase that lead to highly oxidized products (Lim et al., 2010) containing these measured FGs. In these comparisons, we cannot rule out that some biases in measurement may originate from molar absorption coefficients estimated for each FG in FT-IR. The absorption intensity is determined by a change in the magnitude of the dipole moment and can vary according to molecule or mixture environment; the representativeness of applied absorption coefficients in these SOA mixtures is a possible area for future inquiry. However, Takahama et al. (2013) cite variations on the order of 20% for oxygenated FGs in several carboxylic acid, and ketone species, which provide some constraints on this uncertainty for the range of compound classes evaluated in their study.

As reported by Ruggeri et al. (2016), SIM-FULL has similar O/C of observations in similar chamber studies where Aerosol Mass Spectrometer (AMS) measurements were available (Chen et al., 2011; Zhang et al., 2015). OM in MEAS-NOM is less functionalized than in SIM-FULL at hour 4, but the opposite is true at hour 21 even while hydroperoxide and peroxyacyl nitrate is not included. The rate of transformation of these FGs remains uncertain — for instance, reported lifetimes of hydroperoxides range from less than an hour to many days (Epstein et al., 2014; Krapf et al., 2016); resolving their reaction pathways may play a critical role in understanding model-measurement discrepancies (McVay et al., 2016). Using the estimates of MEAS-NOM, the additional oxidation and aging process between 4 and 21 hours leads to an increase in O/C of about 0.24, including a 0.09 difference in O/C from carbonyl (a product of hydroperoxide photolysis). If we extrapolate the O/C of MEAS-NOM to that which includes hydroperoxide and peroxyacyl nitrate groups by assuming the same hydroperoxide and peroxyacyl nitrate contributions from SIM-FULL, we would obtain an overall O/C ratio of 0.7 at hour 4 and 0.9 at hour 21. The latter value is at the higher end of O/C values by reported by AMS (e.g., Aiken et al., 2008; Jimenez et al., 2009; Canagaratna et al., 2015; Lambe et al., 2015). A concurrent measurement of overall O/C and O/C partitioned by measured FG may provide better constraints on our understanding of OM transformations.

As with O/C and OM/OC, $\overline{\mathrm{OS}}_\mathrm{C}$ also highlights the greater extent of functionalization in observations than in simulations between hours 4 and 21. $\overline{\mathrm{OS}}_\mathrm{C}$ estimated from MEAS-NOM is in the range of low-volatility oxygenated organic aerosol (LV-OOA) (Donahue et al., 2012), while they are in the range of semi-volatile oxygenated organic aerosol (SV-OOA) in the simulations as consistent with the species included in the MCMv3.2 mechanism. In simulation, the products found in the aerosol phase are contain more than six carbon atoms, and the smaller, highly oxidized molecules remain in the gas phase (Section S3, Figure S1) As discussed in Section 3.2 and shown in comparison between SIM-MEAS1 and SIM-FULL (Figure 10c), the missing contributions from hydroperoxide and peroxyacyl to $\overline{\mathrm{OS}}_\mathrm{C}$ are likely to be small as only the valence of the bonded atoms, and not the total atomic count of the FGs, contribute to the carbon oxidation state.

4 Conclusions

This study extends the work of Ruggeri and Takahama (2016) and Ruggeri et al. (2016) to demonstrate how molecular structure — specifically, functionalization — can inform comparisons between model and measurement through knowledge of the underlying carbon type abundances. For a measured subset of molar FG abundances, we estimate the expected mass recovery of simulated OC and OM, and how this impacts reported properties such as atomic ratios (O/C, H/C) and OM/OC mass ratios that are of interest to the atmospheric aerosol community. Furthermore, we show how information regarding the underlying molecular structure can be used to better constrain the abundance of polyfunctional carbon that can be estimated from measurements of FGs.

For the α-pinene photooxidation simulation analyzed, we find that 80% of the carbon is detectable by the set of commonly measured FGs, and 7% is unmeasurable on account of having only carbon-carbon bonds. The problem of multiply enumerating polyfunctional carbon atoms using FG abundances for types in this simulated mixture introduces a smaller error, typically less than 10%. The coefficients required to map FG abundance to carbon abundance varies slightly from what has been assumed for ambient samples; until more studies are conducted there may be reason to continue using previous coefficients for consistency. Comparison of simulation results to measured O/C, OM/OC, and carbon oxidation state partitioned by FG contributions elucidated the magnitude of missing LV-OOA (among other classes of molecules) in our model on these widely use metrics. Our current model only includes gas-phase chemistry prescribed by MCMv3.2 combined with gas-particle partitioning at present time, but such comparisons can be extended as additional mechanisms are added. Within the context of this framework, the value of improving our knowledge of SOA formation and aging, investigating measurement artifacts, and developing calibration models for additional FGs for improved comparison with models can be better evaluated.

In that FG analysis measures characteristics of carbon types present in molecules of complex SOA mixtures, it can bridge our understanding of the atomic composition (e.g., measured via AMS) and constituent molecules identified by the growing number of emerging analytical methods (e.g., Kalberer et al., 2006; Altieri et al., 2008; Jokinen et al., 2012; Chan et al., 2013; Chhabra et al., 2015; Lopez-Hilfiker et al., 2015; Nozière et al., 2015) to place their contributions in perspective. With regards to numerical simulation, model-measurement integration using FGs can further guide development of chemical mechanism generators (e.g., Aumont et al., 2005; Fooshee et al., 2012; Gao et al., 2016) and detailed benchmark models (e.g., Saunders et al., 2003), upon which reduced chemical reaction schemes are based (e.g., Dawson et al., 2016). We anticipate that the work expounded in this series of manuscripts will strengthen the ensemble of tools available to study the complex phenomena of organic aerosol formation and aging.

Appendix A: Code and software

Code and software associated with Ruggeri and Takahama (2016), Ruggeri et al. (2016), and this work are released under the GNU Public License (GPLv3) and listed in Table A1. The code can be downloaded as a zipped file from the listed repositories,

Table A1. Code.

Name	Description	Repository
Substructure Search Program	Enumerates FGs in molecules.	https://github.com/stakahama/aprl-ssp
KPP with G/P Partitioning	Generates model for gas phase chemistry with partitioning based on MCM mechanism.	https://github.com/stakahama/aprl-kpp-gp
Carbon type analysis	Maps to FGs to carbon types. Reproduces analysis and figures in this manuscript.	https://github.com/stakahama/aprl-carbontypes

or via command line by the syntax `git clone https://github.com/stakahama/{reponame}`. Instructions are included in the README.md file in each repository. The corresponding author can be contacted for more information.

Appendix B: Notation

Symbols used throughout this manuscript are summarized in B1. Indices are written in lower case, vectors (single-column matrix) in bold italic, matrices in bold, and sets in calligraphy font. A hat over a variable indicates its statistically estimated value. A starred symbol indicates the detectable value corresponding to any given set of FGs.

Appendix C: Vibrational modes

Absorption bands for additional FGs in Set2 (Section 2.3) are shown in Table C1. Hydroperoxide in the condensed phase has been measured using FT-IR (e.g., Shreve et al., 1951; van de Voort et al., 1994), but peroxyacyl nitrate analysis has mostly been limited to the gas phase (e.g., Gaffney et al., 1984; Monedero et al., 2008).

Appendix D: Error estimation

In this section, relative uncertainties arising from the deviation between $\hat{n}^*_{\mathrm{C}}$ and n^*_{C} are translated into uncertainties of atomic ratios and OM/OC. As abundances of heteroatoms are determined from FG measurement do not suffer from multiple counting, uncertainties in their abundances are not considered.

Any of the estimation methods for n^*_{C} incurs a deviation from its true value by ϵ, which we write as $\hat{n}^*_{\mathrm{C}} = n^*_{\mathrm{C}} + \epsilon$. We can recast this deviation as a relative error $\delta_{[n^*_{\mathrm{C}}]}$ with respect to n^*_{C} such that $\epsilon = \delta_{[n^*_{\mathrm{C}}]} n^*_{\mathrm{C}}$. The magnitude of $\delta_{[n^*_{\mathrm{C}}]}$ can be associated with

Table B1. Mathematical symbols used in the manuscript and their descriptions.

Category	Symbol	Description
Indices	i	compound or molecule index
	k	carbon type index
	j	FG index
	a	atom index
Variables	n	number of moles of a substance (atom, compound, or FG)
	$\mathbf{X} = [x_{ij}]$	group composition matrix
	$\mathbf{Y} = [y_{ik}]$	carbon type matrix
	$\mathbf{\Theta} = [\theta_{kj}]$	carbon-group matrix
	$\mathbf{\Phi} = [\phi_{jk}]$	group-carbon matrix
	$\boldsymbol{\zeta} = [\zeta_k]$	carbon type oxidation state vector
	$\boldsymbol{z} = [z_j]$	oxidation state contribution vector
	$\mathbf{\Lambda} = [\lambda_{aj}]$	atom-group matrix
	$\boldsymbol{\lambda}_{\mathrm{C}} = [\hat{\lambda}_{\mathrm{C},j}]$	carbon atom-group vector
	$\mathrm{OS_C}$	carbon oxidation state
	$\overline{\mathrm{OS}}_{\mathrm{C}}$	mean carbon oxidation state
Sets	$\mathcal{A}$	set of atoms
	$\mathcal{M}$	set of molecule types
	$\mathcal{J}$	set of FGs
	$\mathcal{C}$	set of carbon types

Table C1. Absorption bands in the mid-infrared for vibrational modes present in FGs proposed for Set2 (Section 2.3).

FG	$\tilde{\nu}$ ($\mathrm{cm^{-1}}$)	description
eCH[1]	3005–2980	C-H stretch
hydroperoxide[2]	3300–3400	OO-H stretch (strong)
	860–840	O–OH stretch (weak)
peroxyacyl nitrate[2,3]	760–849	NO scissoring
	1340–1223	NO_2 symmetric stretch
	1777–1700	NO_2 anti-symmetric stretch
	1880–1777	C=O stretch

[1] Maria et al. (2003); [2] Shurvell (2006); [3] Monedero et al. (2008)

the ratio $\hat{n}^*_{\mathrm{C}}/n^*_{\mathrm{C}}$ shown in Figures 8 and 9 by the relation: $\delta_{[n^*_{\mathrm{C}}]} = 1 - \hat{n}^*_{\mathrm{C}}/n^*_{\mathrm{C}}$. The resulting expression $\hat{n}^*_{\mathrm{C}} = n^*_{\mathrm{C}}(1+\delta_{[n^*_{\mathrm{C}}]})$ is then used to anticipate relative errors on the actual atomic ratios and OM/OC ratio as follows:

$$\delta_{[n^*_a/n^*_{\mathrm{C}}]} = 1 - \frac{[n^*_a/n^*_{\mathrm{C}}]\,/\,\left(1+\delta_{[n^*_{\mathrm{C}}]}\right)}{[n^*_a/n^*_{\mathrm{C}}]} = 1 - \frac{1}{1+\delta_{[n^*_{\mathrm{C}}]}} \tag{D1}$$

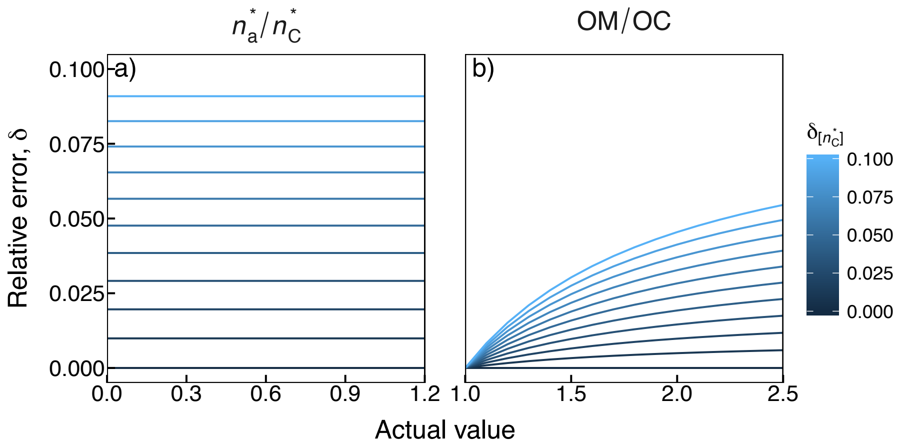


Figure D1. Magnitude of relative errors in atomic ratios ($\delta_{[n_a^*/n_C^*]}$) and OM/OC mass ratios ($\delta_{[OM/OC]}$) due to relative errors ($\delta_{[n_C^*]}$) in the estimation of number of carbon atoms n_C^*. Ten colored lines shown in each panel correspond to values of $\delta_{[n_C^*]} = \{0.0, 0.01, 0.02, \ldots, 0.1\}$.

$$\delta_{[OM/OC]} = 1 - \frac{1 + \left([OM/OC] - 1\right) / \left(1 + \delta_{[n_C^*]}\right)}{[OM/OC]} = 1 - \left(\frac{1}{1 + \delta_{[n_C^*]}} + \frac{1}{[OM/OC]} - \frac{1}{[OM/OC]\left(1 + \delta_{[n_C^*]}\right)} \right) \tag{D2}$$

Author contributions. S. Takahama and G. Ruggeri designed and performed the analysis. S. Takahama wrote the manuscript.

Acknowledgements. Funding was provided by the Swiss National Science Foundation (200021_143298).

References

Aiken, A. C., Decarlo, P. F., Kroll, J. H., Worsnop, D. R., Huffman, J. A., Docherty, K. S., Ulbrich, I. M., Mohr, C., Kimmel, J. R., Sueper, D., Sun, Y., Zhang, Q., Trimborn, A., Northway, M., Ziemann, P. J., Canagaratna, M. R., Onasch, T. B., Alfarra, M. R., Prevot, A. S. H., Dommen, J., Duplissy, J., Metzger, A., Baltensperger, U., and Jimenez, J. L.: O/C and OM/OC ratios of primary, secondary, and ambient organic aerosols with high-resolution time-of-flight aerosol mass spectrometry, *Environmental Science & Technology*, 42, 4478–4485, doi:10.1021/es703009q, 2008.

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, 2013.

Allen, D. T., Palen, E. J., Haimov, M. I., Hering, S. V., and Young, J. R.: Fourier-transform Infrared-spectroscopy of Aerosol Collected In A Low-pressure Impactor (LPI/FTIR) - Method Development and Field Calibration, *Aerosol Science and Technology*, 21, 325–342, doi:10.1080/02786829408959719, 1994.

Altieri, K. E., Seitzinger, S. P., Carlton, A. G., Turpin, B. J., Klein, G. C., and Marshall, A. G.: Oligomers formed through in-cloud methylglyoxal reactions: Chemical composition, properties, and mechanisms investigated by ultra-high resolution FT-ICR mass spectrometry RID A-7867-2011, *Atmospheric Environment*, 42, 1476–1490, doi:10.1016/j.atmosenv.2007.11.015, 2008.

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

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.

Camredon, M., Aumont, B., Lee-Taylor, J., and Madronich, S.: The SOA/VOC/NOx system: an explicit model of secondary organic aerosol formation, *Atmospheric Chemistry and Physics*, 7, 5599–5610, 2007.

Canagaratna, M. R., Jimenez, J. L., Kroll, J. H., Chen, Q., Kessler, S. H., Massoli, P., Hildebrandt Ruiz, L., Fortner, E., Williams, L. R., Wilson, K. R., Surratt, J. D., Donahue, N. M., Jayne, J. T., and Worsnop, D. R.: Elemental ratio measurements of organic compounds using aerosol mass spectrometry: characterization, improved calibration, and implications, *Atmos. Chem. Phys.*, 15, 253–272, 2015.

Chan, A. W. H., Isaacman, G., Wilson, K. R., Worton, D. R., Ruehl, C. R., Nah, T., Gentner, D. R., Dallmann, T. R., Kirchstetter, T. W., Harley, R. A., Gilman, J. B., Kuster, W. C., deGouw, J. A., Offenberg, J. H., Kleindienst, T. E., Lin, Y. H., Rubitschun, C. L., Surratt, J. D., Hayes, P. L., Jimenez, J. L., and Goldstein, A. H.: Detailed chemical characterization of unresolved complex mixtures in atmospheric organics: Insights into emission sources, atmospheric processing, and secondary organic aerosol formation, *Journal of Geophysical Research-atmospheres*, 118, 6783–6796, doi:10.1002/jgrd.50533, 2013.

Chen, Q., Liu, Y., Donahue, N. M., Shilling, J. E., and Martin, S. T.: Particle-Phase Chemistry of Secondary Organic Material: Modeled Compared to Measured O:C and H:C Elemental Ratios Provide Constraints, *Environmental Science & Technology*, 45, 4763–4770, doi:10.1021/es104398s, 2011.

Chhabra, P. S., Ng, N. L., Canagaratna, M. R., Corrigan, A. L., Russell, L. M., Worsnop, D. R., Flagan, R. C., and Seinfeld, J. H.: Elemental composition and oxidation of chamber organic aerosol, *Atmospheric Chemistry and Physics*, 11, 8827–8845, doi:10.5194/acp-11-8827-2011, 2011.

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

Chuang, W. K. and Donahue, N. M.: A two-dimensional volatility basis set — Part 3: Prognostic modeling and NO_x dependence, *Atmospheric Chemistry and Physics*, 16, 123–134, doi:10.5194/acp-16-123-2016, 2016.

Coury, C. and Dillner, A. M.: A method to quantify organic functional groups and inorganic compounds in ambient aerosols using attenuated total reflectance FTIR spectroscopy and multivariate chemometric techniques, *Atmospheric Environment*, 42, 5923–5932, doi:10.1016/j.atmosenv.2008.03.026, 2008.

Crounse, J. D., Nielsen, L. B., Jørgensen, S., Kjaergaard, H. G., and Wennberg, P. O.: Autoxidation of Organic Compounds in the Atmosphere, *The Journal of Physical Chemistry Letters*, 4, 3513–3520, doi:10.1021/jz4019207, 2013.

Dawson, M. L., Xu, J., Griffin, R. J., and Dabdub, D.: Development of aroCACM/MPMPO 1.0: a model to simulate secondary organic aerosol from aromatic precursors in regional models, *Geoscientific Model Development*, 9, 2143–2151, doi:10.5194/gmd-9-2143-2016, 2016.

Day, D. A., Liu, S., Russell, L. M., and Ziemann, P. J.: Organonitrate group concentrations in submicron particles with high nitrate and organic fractions in coastal southern California, *Atmospheric Environment*, 44, 1970–1979, doi:10.1016/j.atmosenv.2010.02.045, 2010.

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., Sipila, 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., Kurten, T., Nielsen, L. B., Jorgensen, S., Kjaergaard, H. G., Canagaratna, M., Maso, M. D., Berndt, T., Petaja, 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, 2014.

Epstein, S. A., Blair, S. L., and Nizkorodov, S. A.: Direct Photolysis of a-Pinene Ozonolysis Secondary Organic Aerosol: Effect on Particle Mass and Peroxide Content, *Environmental Science & Technology*, 48, 11 251–11 258, doi:10.1021/es502350u, 2014.

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

Gaffney, J., Fajer, R., and Senum, G.: An improved procedure for high purity gaseous peroxyacyl nitrate production: Use of heavy lipid solvents, *Atmospheric Environment (1967)*, 18, 215 – 218, doi:http://dx.doi.org/10.1016/0004-6981(84)90245-2, 1984.

Gao, C. W., Allen, J. W., Green, W. H., and West, R. H.: Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms, *Computer Physics Communications*, 203, 212 – 225, doi:http://dx.doi.org/10.1016/j.cpc.2016.02.013, 2016.

Glasius, M., Lahaniati, M., Calogirou, A., Bella, D. D., Jensen, N. R., Hjorth, J., Kotzias, D., , and Larsen, B. R.: Carboxylic Acids in Secondary Aerosols from Oxidation of Cyclic Monoterpenes by Ozone, *Environmental Science & Technology*, 34, 1001–1010, doi:10.1021/es990445r, 2000.

Henderson, B. H.: Python-based Environment for Reaction Mechanisms/Mathematics (PERMM), doi:dx.doi.org/10.5281/zenodo.44396, https://github.com/barronh/permm/, 2015.

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.

Jimenez, J. L., Canagaratna, M. R., Donahue, N. M., Prevot, A. S. H., Zhang, Q., Kroll, J. H., DeCarlo, P. F., Allan, J. D., Coe, H., Ng, N. L., Aiken, A. C., Docherty, K. S., Ulbrich, I. M., Grieshop, A. P., Robinson, A. L., Duplissy, J., Smith, J. D., Wilson, K. R., Lanz, V. A., Hueglin, C., Sun, Y. L., Tian, J., Laaksonen, A., Raatikainen, T., Rautiainen, J., Vaattovaara, P., Ehn, M., Kulmala, M., Tomlinson, J. M., Collins, D. R., Cubison, M. J., Dunlea, E. J., Huffman, J. A., Onasch, T. B., Alfarra, M. R., Williams, P. I., Bower, K., Kondo, Y., Schneider, J., Drewnick, F., Borrmann, S., Weimer, S., Demerjian, K., Salcedo, D., Cottrell, L., Griffin, R., Takami, A., Miyoshi, T., Hatakeyama, S., Shimono, A., Sun, J. Y., Zhang, Y. M., Dzepina, K., Kimmel, J. R., Sueper, D., Jayne, J. T., Herndon, S. C., Trimborn, A. M., Williams, L. R., Wood, E. C., Middlebrook, A. M., Kolb, C. E., Baltensperger, U., and Worsnop, D. R.: Evolution of Organic Aerosols in the Atmosphere, *Science*, 326, 1525–1529, doi:10.1126/science.1180353, 2009.

Jokinen, T., Sipilä, M., Junninen, H., Ehn, M., Lönn, G., Hakala, J., Petäjä, T., Mauldin III, R. L., Kulmala, M., and Worsnop, D. R.: Atmospheric sulphuric acid and neutral cluster measurements using CI-APi-TOF, *Atmospheric Chemistry and Physics*, 12, 4117–4125, doi:10.5194/acp-12-4117-2012, 2012.

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.

Krapf, M., El Haddad, I., Bruns, E., Molteni, U., Daellenbach, K., Prévôt, A. H., Baltensperger, U., and Dommen, J.: Labile Peroxides in Secondary Organic Aerosol, *Chem*, 1, 603–616, 2016.

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, *The Journal of Physical Chemistry A*, 119, 10 767–10 783, doi:10.1021/acs.jpca.5b06946, 2015.

Lambe, A. T., Chhabra, P. S., Onasch, T. B., Brune, W. H., Hunter, J. F., Kroll, J. H., Cummings, M. J., Brogan, J. F., Parmar, Y., Worsnop, D. R., Kolb, C. E., and Davidovits, P.: Effect of oxidant concentration, exposure time, and seed particles on secondary organic aerosol chemical composition and yield, *Atmospheric Chemistry and Physics*, 15, 3063–3075, doi:10.5194/acp-15-3063-2015, 2015.

Lim, Y. B., Tan, Y., Perri, M. J., Seitzinger, S. P., and Turpin, B. J.: Aqueous chemistry and its role in secondary organic aerosol (SOA) formation, *Atmospheric Chemistry and Physics*, 10, 10 521–10 539, doi:10.5194/acp-10-10521-2010, 2010.

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.

Lopez-Hilfiker, F. D., Mohr, C., Ehn, M., Rubach, F., Kleist, E., Wildt, J., Mentel, T. F., Carrasquillo, A. J., Daumit, K. E., Hunter, J. F., Kroll, J. H., Worsnop, D. R., and Thornton, J. A.: Phase partitioning and volatility of secondary organic aerosol components formed from α-pinene ozonolysis and OH oxidation: the importance of accretion products and other low volatility compounds, *Atmospheric Chemistry and Physics*, 15, 7765–7776, doi:10.5194/acp-15-7765-2015, 2015.

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

McVay, R. C., Zhang, X., Aumont, B., Valorso, R., Camredon, M., La, Y. S., Wennberg, P. O., and Seinfeld, J. H.: SOA formation from the photooxidation of α-pinene: systematic exploration of the simulation of chamber data, *Atmospheric Chemistry and Physics*, 16, 2785–2802, doi:10.5194/acp-16-2785-2016, 2016.

Monedero, E., Salgado, M., Villanueva, F., Martín, P., Barnes, I., and Cabañas, B.: Infrared absorption cross-sections for peroxyacyl nitrates (nPANs), *Chemical Physics Letters*, 465, 207 – 211, doi:http://dx.doi.org/10.1016/j.cplett.2008.10.020, 2008.

Nozière, B., Kalberer, M., Claeys, M., Allan, J., D'Anna, B., Decesari, S., Finessi, E., Glasius, M., Grgić, I., Hamilton, J. F., Hoffmann, T., Iinuma, Y., Jaoui, M., Kahnt, A., Kampf, C. J., Kourtchev, I., Maenhaut, W., Marsden, N., Saarikoski, S., Schnelle-Kreis, J., Surratt, J. D., Szidat, S., Szmigielski, R., and Wisthaler, A.: The Molecular Identification of Organic Compounds in the Atmosphere: State of the Art and Challenges, *Chem. Rev.*, 115, 3919–3983, doi:10.1021/cr5003485, 2015.

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

Radhakrishnan, K. and Hindmarsh, A. C.: Description and use of LSODE, the Livermore solver for ordinary differential equations, Tech. Rep. UCRL-ID-113855, Lawrence Livermore National Laboratory, nASA Reference Publication 1327, 1993.

Ranney, A. P. and Ziemann, P. J.: Microscale spectrophotometric methods for quantification of functional groups in oxidized organic aerosol, *Aerosol Science and Technology*, 50, 881–892, doi:10.1080/02786826.2016.1201197, 2016.

Reff, A., Turpin, B. J., Offenberg, J. H., Weisel, C. P., Zhang, J., Morandi, M., Stock, T., Colome, S., and Winer, A.: A functional group characterization of organic PM2.5 exposure: Results from the RIOPA study RID C-3787-2009, *Atmospheric Environment*, 41, 4585–4598, doi:10.1016/j.atmosenv.2007.03.054, 2007.

Rogge, W. F., Hildemann, L. M., Mazurek, M. A., Cass, G. R., and Simoneit, B. R. T.: Sources of Fine Organic Aerosol .2. Non-catalyst and Catalyst-equipped Automobiles and Heavy-duty Diesel Trucks, *Environmental Science & Technology*, 27, 636–651, doi:10.1021/es00041a007, 1993.

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

Ruggeri, G., Bernhard, F. A., Henderson, B. H., and Takahama, S.: Model-measurement comparison of functional group abundance in α-pinene and 1,3,5-trimethylbenzene secondary organic aerosol formation, *Atmospheric Chemistry and Physics*, 16, 8729–8747, doi:10.5194/acp-16-8729-2016, 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.: 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.

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.

Ruthenburg, T. C., Perlin, P. C., Liu, V., McDade, C. E., and Dillner, A. M.: Determination of organic matter and organic matter to organic carbon ratios by infrared spectroscopy with application to selected sites in the IMPROVE network, *Atmospheric Environment*, 86, 47–57, doi:10.1016/j.atmosenv.2013.12.034, 2014.

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

Sax, M., Zenobi, R., Baltensperger, U., and Kalberer, M.: Time resolved infrared spectroscopic analysis of aerosol formed by photo-oxidation of 1,3,5-trimethylbenzene and alpha-pinene, *Aerosol Science and Technology*, 39, 822–830, doi:10.1080/02786820500257859, 2005.

Shreve, O. D., Heether, M. R., Knight, H. B., and Swern, D.: Infrared Absorption Spectra of Some Hydroperoxides, Peroxides, and Related Compounds, *Analytical Chemistry*, 23, 282–285, doi:10.1021/ac60050a015, 1951.

Shurvell, H.: Spectra–Structure Correlations in the Mid- and Far-Infrared, John Wiley & Sons, Ltd, doi:10.1002/0470027320.s4101, 2006.

Takahama, S. and Dillner, A. M.: Model selection for partial least squares calibration and implications for analysis of atmospheric organic aerosol samples with mid-infrared spectroscopy, *Journal of Chemometrics*, 29, 659–668, doi:10.1002/cem.2761, 2015.

Takahama, S., Johnson, A., and Russell, L. M.: Quantification of Carboxylic and Carbonyl Functional Groups in Organic Aerosol Infrared Absorbance Spectra, *Aerosol Science and Technology*, 47, 310–325, doi:10.1080/02786826.2012.752065, 2013.

Tröstl, J., Chuang, W. K., Gordon, H., Heinritzi, M., Yan, C., Molteni, U., Ahlm, L., Frege, C., Bianchi, F., Wagner, R., Simon, M., Lehtipalo, K., Williamson, C., Craven, J. S., Duplissy, J., Adamov, A., Almeida, J., Bernhammer, A.-K., Breitenlechner, M., Brilke, S., Dias, A., Ehrhart, S., Flagan, R. C., Franchin, A., Fuchs, C., Guida, R., Gysel, M., Hansel, A., Hoyle, C. R., Jokinen, T., Junninen, H., Kangasluoma, J., Keskinen, H., Kim, J., Krapf, M., Kürten, A., Laaksonen, A., Lawler, M., Leiminger, M., Mathot, S., Möhler, O., Nieminen, T., Onnela, A., Petäjä, T., Piel, F. M., Miettinen, P., Rissanen, M. P., Rondo, L., Sarnela, N., Schobesberger, S., Sengupta, K., Sipilä, M., Smith, J. N., Steiner, G., Tomè, A., Virtanen, A., Wagner, A. C., Weingartner, E., Wimmer, D., Winkler, P. M., Ye, P., Carslaw, K. S., Curtius, J., Dommen, J., Kirkby, J., Kulmala, M., Riipinen, I., Worsnop, D. R., Donahue, N. M., and Baltensperger, U.: The role of low-volatility organic compounds in initial particle growth in the atmosphere, *Nature*, 533, 527–531, 2016.

van de Voort, F. R., Ismail, A. A., Sedman, J., Dubois, J., and Nicodemo, T.: The determination of peroxide value by fourier transform infrared spectroscopy, *Journal of the American Oil Chemists' Society*, 71, 921–926, doi:10.1007/BF02542254, 1994.

Whitten, G. Z., Hogo, H., and Killus, J. P.: The carbon-bond mechanism: a condensed kinetic mechanism for photochemical smog, *Environ. Sci. Technol.*, 14, 690–700, doi:10.1021/es60166a008, 1980.

Yanenko, N. N.: The Method of Fractional Steps: The Solution of Problems of Mathematical Physics in Several Variables, Springer, 1 edition edn., 1971.

Yu, J. Z., Cocker, D. R., Griffin, R. J., Flagan, R. C., and Seinfeld, J. H.: Gas-phase ozone oxidation of monoterpenes: Gaseous and particulate products, *Journal of Atmospheric Chemistry*, 34, 207–258, doi:10.1023/A:1006254930583, 1999.

Zhang, X., McVay, R. C., Huang, D. D., Dalleska, N. F., Aumont, B., Flagan, R. C., and Seinfeld, J. H.: Formation and evolution of molecular products in α-pinene secondary organic aerosol, *Proceedings of the National Academy of Sciences*, 112, 14 168–14 173, doi:10.1073/pnas.1517742112, 2015.

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.

Tables

Table 1. Average number of atoms attached to each type of bond assumed for various types of mixtures. $\lambda_{\mathrm{C,COOH}} = \lambda_{\mathrm{C,carbonyl}} = 1$.

Study	Mixture type	$\lambda_{\mathrm{C,CH}}$	$\lambda_{\mathrm{C,COH}}$	$\lambda_{\mathrm{C,CONO2}}$
Allen et al. (1994)	ambient	0.5		1
Russell (2003)	ambient	0.5	1	
Reff et al. (2007)	indoor/ambient	0.48		
Chhabra et al. (2011)	α-pinene SOA	0.63	0.63	0.63
	guaiacol SOA	0.88	0.88	0.88
Several*	ambient	0.5	0.5	0.25
Ruthenburg et al. (2014)	ambient	0.5	0	

*reflects assumptions by Russell et al. (2009), Liu et al. (2009), and Day et al. (2010).

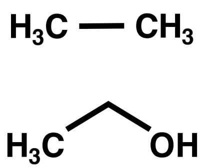
Table 2. Values for λ_{C} with standard errors in parentheses where available (uncertainties were not calculated for the constrained optimization algorithm in the MIXTURE estimation method). Values for $\lambda_{\mathrm{C,COOH}} = \lambda_{\mathrm{C,carbonyl}} = 1$ are fixed and therefore not included in the table.

Set	Method	aCH	aCOH	$CONO_2$	eCH	hydroperoxide
Set1	COUNT	0.39 (0.04)	0.52 (0.17)	0.52 (0.17)		
Set1	COMPOUND	0.47 (0.01)	0.31 (0.06)	0.64 (0.11)		
Set1	MIXTURE	0.45	0.09	1.00		
Set1	NOMINAL	0.45	0.50	0.50		
Set2	COUNT	0.39 (0.04)	0.52 (0.17)	0.52 (0.17)	0.75 (0.25)	0.52 (0.17)
Set2	COMPOUND	0.48 (0.01)	0.26 (0.05)	0.54 (0.09)	1.08 (0.20)	0.35 (0.07)
Set2	MIXTURE	0.50	0.16	0.41	1.00	0.00
Set2	NOMINAL	0.45	0.50	0.50	1.00	0.50

Figures

ethane:

ethanol:



$$\mathbf{Y\Theta} = \mathbf{X}$$ FG balance

$$\mathbf{Y} = \mathbf{X\Phi}$$ carbon type balance

$$\mathbf{Y\Theta\Lambda}^T = \mathbf{X\Lambda}^T = \mathbf{N}_{atom}$$ atom balance

$$\mathbf{Y}\zeta = \mathbf{X}z$$ oxidation state balance

$\mathbf{X}$ =

	CH	COH
ethane	6	0
ethanol	5	1

$\mathbf{Y}$ =

	CH_3	H_2COH
ethane	2	0
ethanol	1	1

$\mathbf{N}_{atom}$ =

	C	H	O
ethane	2	6	0
ethanol	2	5	1

z =

	OS
CH	-1
COH	+1

$\mathbf{\Theta}$ =

	CH	COH
CH_3	3	0
H_2COH	2	1

$\mathbf{\Phi}$ =

	CH_3	H_2COH
CH	1/3	0
COH	-1/6	1

$\mathbf{\Lambda}^T$ =

	C	H	O
CH	1/3	1	0
COH	1/3	1	1

ζ =

	OS_C
CH_3	-3
H_2COH	-1

Figure 1. Illustration of carbon type and FG relationships for ethane and ethanol. The FG composition matrix ($\mathbf{X}$), carbon type matrix ($\mathbf{Y}$), and atom composition matrix ($\mathbf{A}$) describe properties of the compounds, and the remaining arrays — oxidation state contribution vector ($\boldsymbol{z}$), carbon-FG matrix ($\mathbf{\Theta}$), FG-carbon matrix ($\mathbf{\Phi}$), atom-FG matrix ($\mathbf{\Lambda}$), and carbon oxidation state vector ($\boldsymbol{\zeta}$) — establish their inter-relationships.

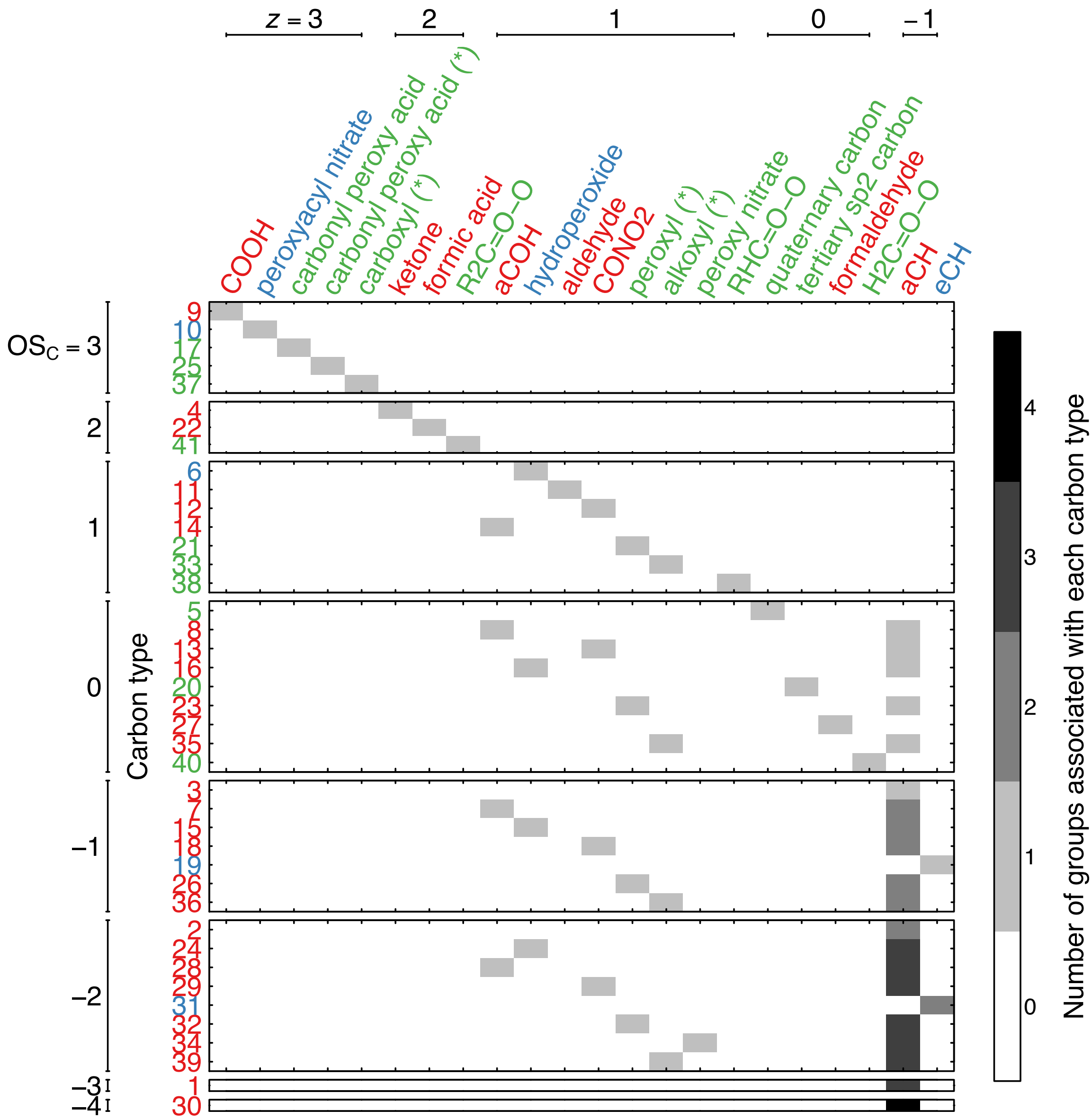


Figure 2. Visualization of the carbon type matrix $\boldsymbol{\Theta}$ for the APIN mechanism. Radical groups are denoted with (*). Carbon types and FGs are ordered by their aerosol abundance (in decreasing order) in the APIN simulation at $t_{\mathrm{max\,SOA}}$ (Section 2.1) with each value of $\mathrm{OS_C}$ and z, respectively. The numeric label for carbon types indicates the overall rank (without regard for its $\mathrm{OS_C}$) in the APIN simulation at $t_{\mathrm{max\,SOA}}$. Formaldehyde and formic acid are subclasses of aldehyde and COOH, respectively, but are defined separately to fulfill the conditions described in Appendix S1. Further details regarding the FG definitions are provided by Ruggeri and Takahama (2016). FGs belonging to measured subset $\mathcal{J}^*$ = Set1 (Section 2.3) is colored in red; additional FGs belonging to Set2 and Full are colored in blue and green, respectively. Corresponding carbon atoms $\mathcal{C}^*$ that are associated with (i.e., detectable by) $\mathcal{J}^*$ are shown in the same colors.

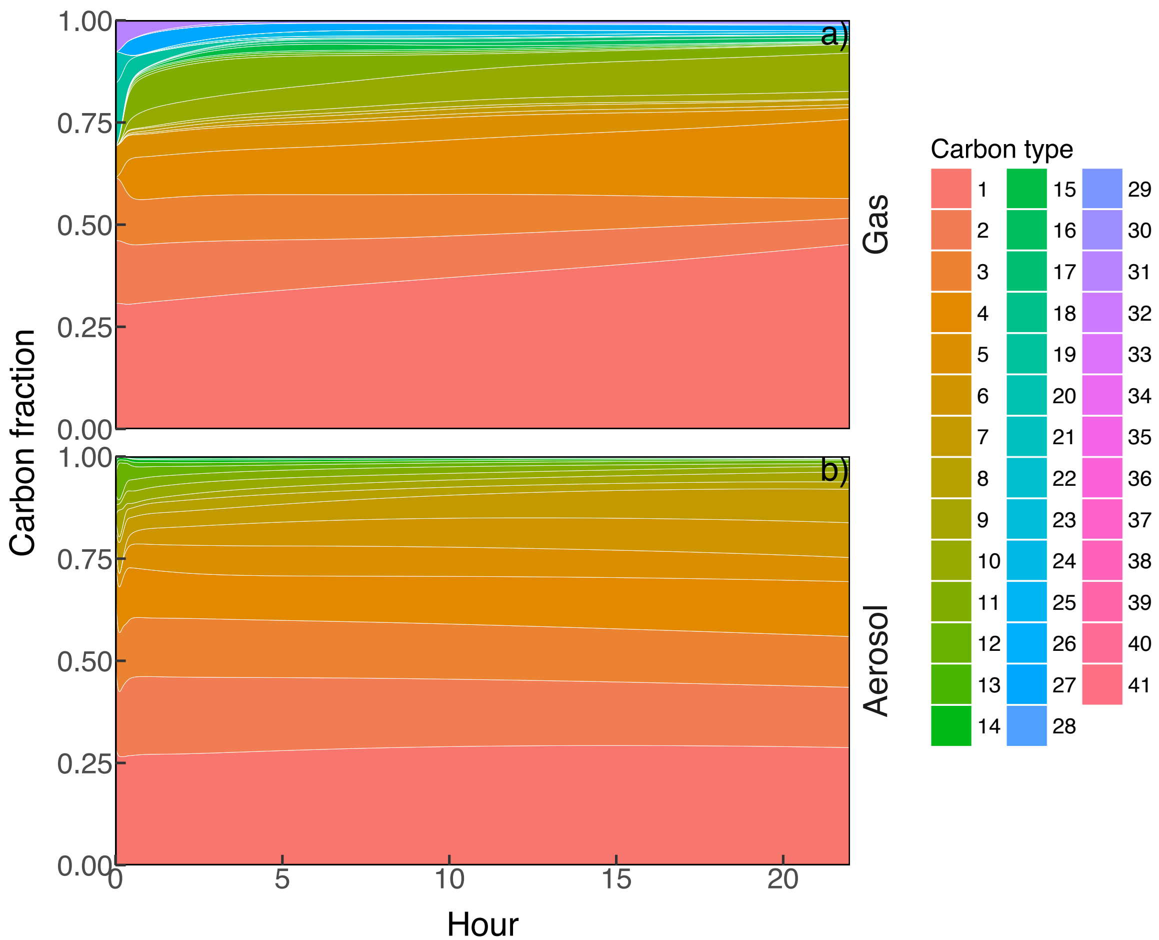


Figure 3. Time series of carbon type abundances for the APIN simulation described in Section 2.1. The carbon types are defined in Figure 2.

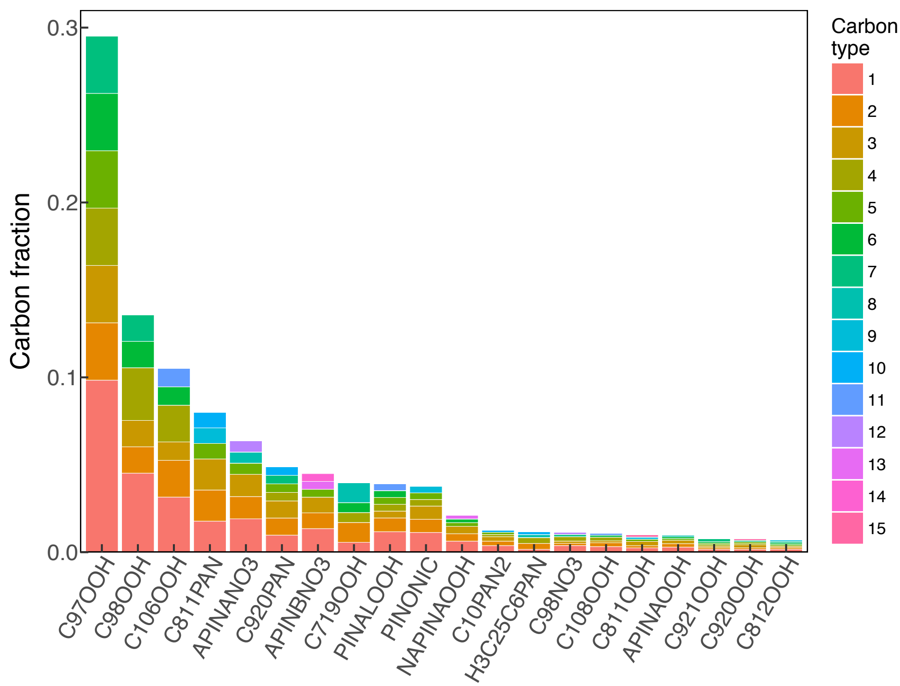


Figure 4. Compound and carbon type abundance for APIN simulation at $t_{\max\mathrm{SOA}}$. C97OOH and C98OOH are large, polyfunctional compounds containing ketone and hydroperoxide groups. The carbon types are defined in Figure 2.

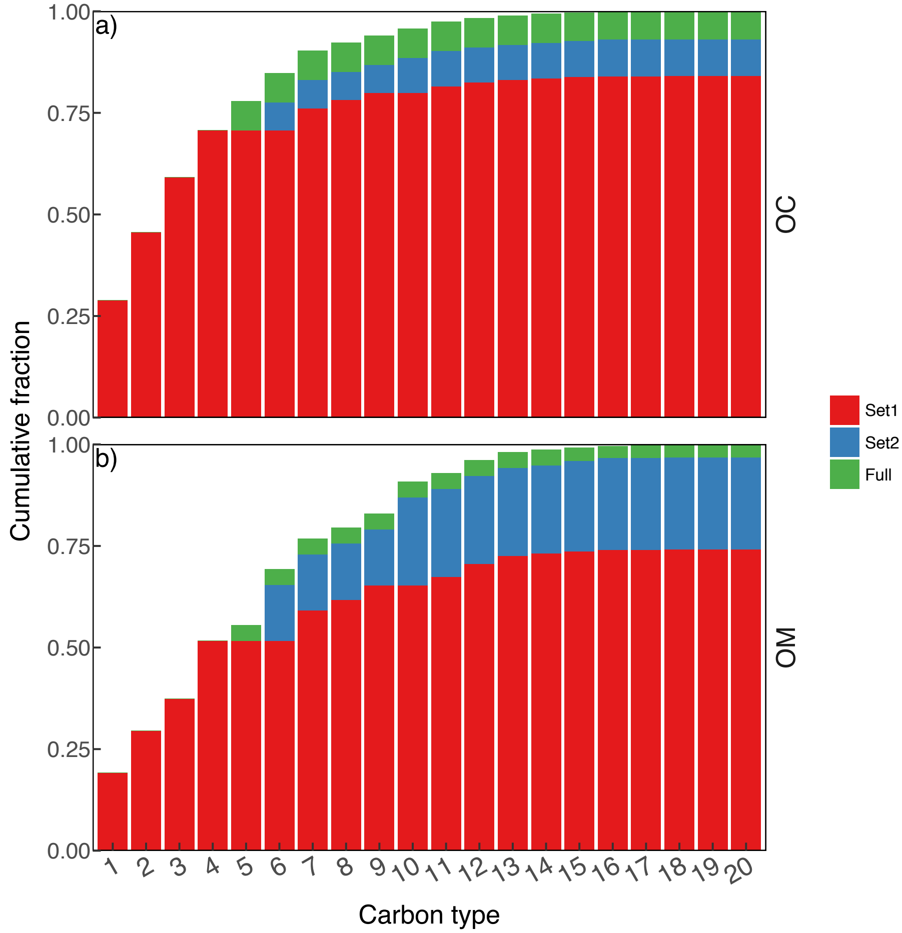


Figure 5. Cumulative carbon fraction for APIN simulation at $t_{\mathrm{max\,SOA}}$. Colors show carbon atoms measurable by different sets of FGs (Section 2.3). The carbon types are defined in Figure 2.

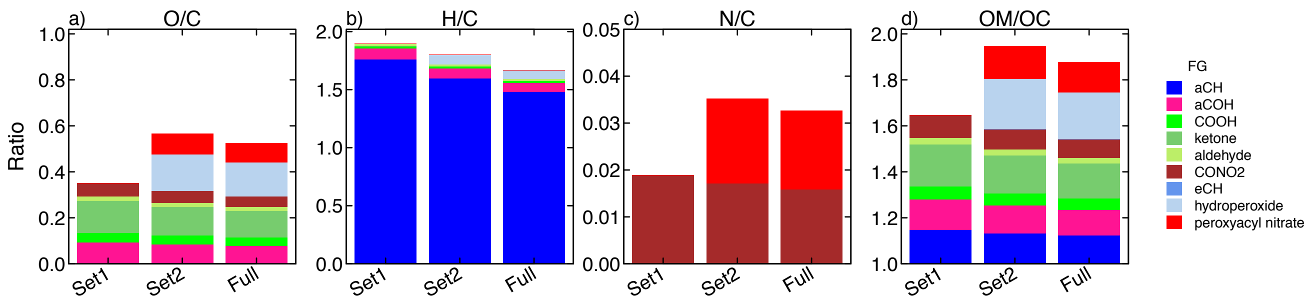


Figure 6. SOA properties for APIN simulation at $t_{\max\mathrm{SOA}}$. Atomic ratios (n_a^*/n_C^*) shown in panels (a)–(c) are in molar units, and OM/OC ratios shown in panel (d) are in mass units. The abundance of carbon used for normalization is defined by the detectable carbon for each set of FGs (Section 2.3), which can lead to estimated ratios with Set1 or Set2 to exceed the Full.

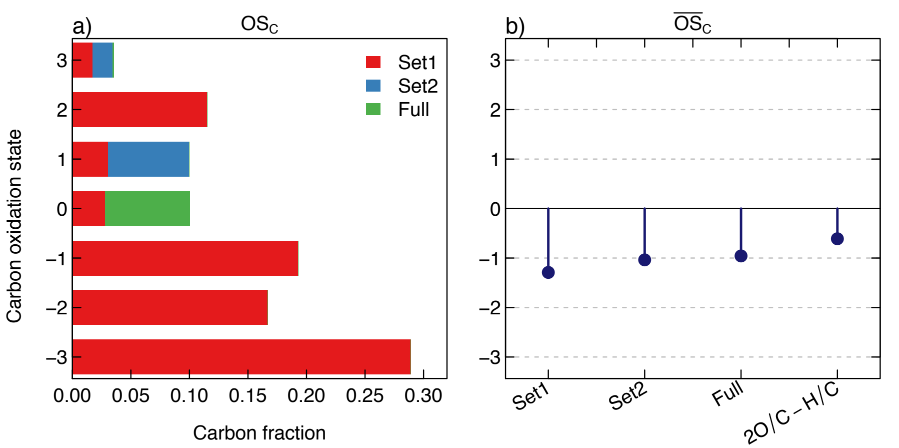


Figure 7. Distribution of carbon oxidation states and their ensemble estimate APIN simulation at $t_{\max \mathrm{SOA}}$. Panel a) shows distribution and measurable carbon atoms with same color scheme 5. Panel b) shows various estimates of $\mathrm{OS_C}$ (b) for the mixture using different FG sets (Section 2.3). $2\mathrm{O}/\mathrm{C} - \mathrm{H}/\mathrm{C}$ is a common approximation used by elemental analysis and is included for reference.

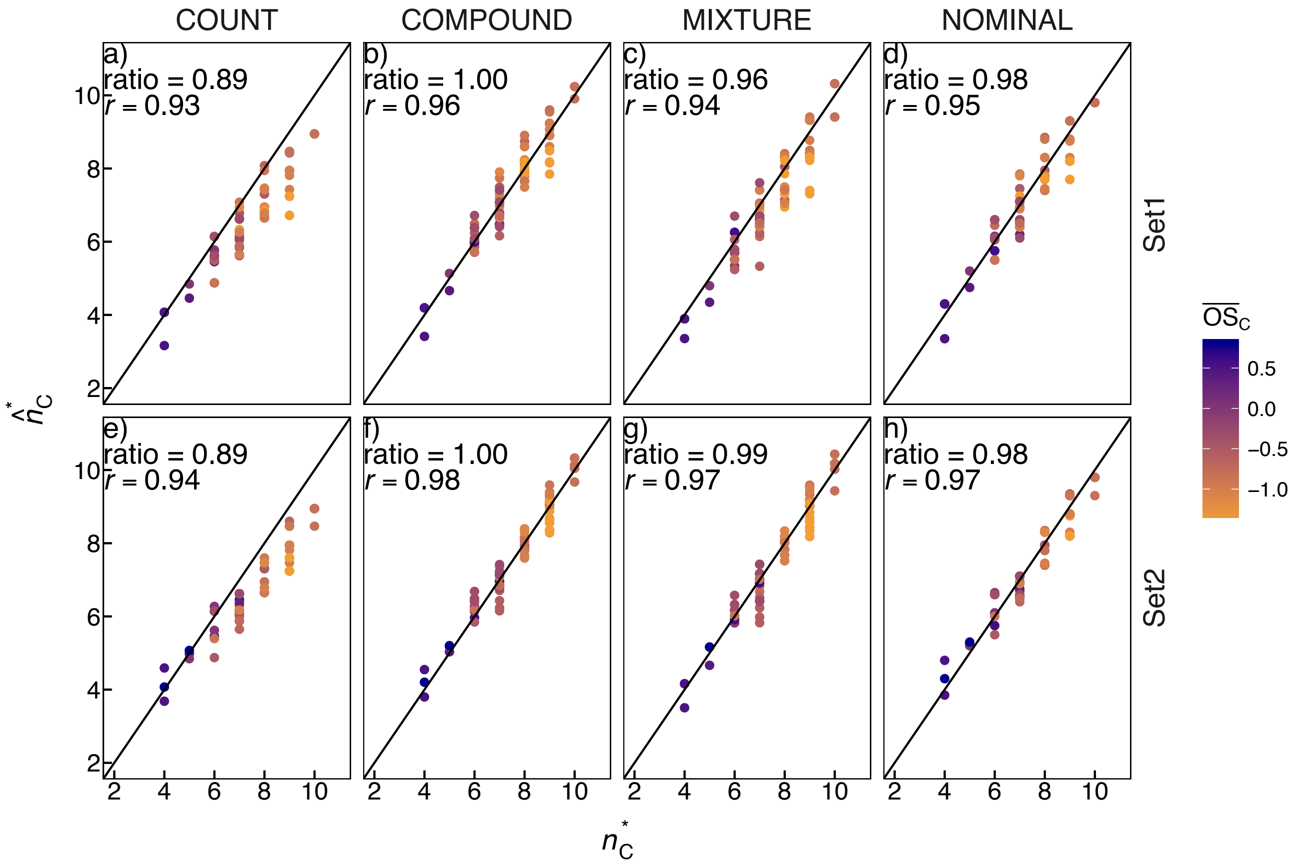


Figure 8. Comparison of estimated ($\hat{n}^*_{\mathrm{C}}$) and actual (n^*_{C}) number of measurable carbon atoms in different SVOC compounds (colored by their compound-averaged oxidation states, $\overline{\mathrm{OS}}_{\mathrm{C}}$) using estimates of $\hat{\boldsymbol{\lambda}}_{\mathrm{C}}$ for various FG sets and solution methods. The diagonal line is the $x = y$ line provided for visual reference. The ratio is defined as $\hat{n}^*_{\mathrm{C}}/n^*_{\mathrm{C}}$ and estimated as the slope (not drawn) of $\hat{n}^*_{\mathrm{C}}$ regressed on n^*_{C}. r is the Pearson's correlation coefficient.

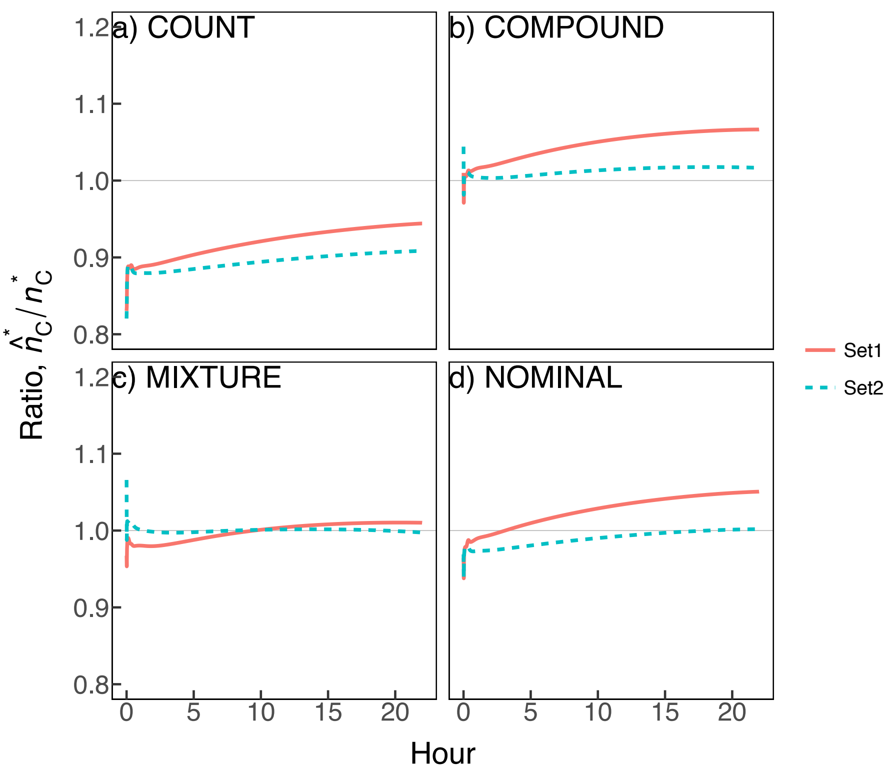


Figure 9. Ratios of estimated ($\hat{n}_{\mathrm{C}}^*$) and actual (n_{C}^*) number of measurable carbon atoms in the APIN simulated aerosol mixture using estimates of $\hat{\lambda}_{\mathrm{C}}$ for various FG sets and solution methods. The gray horizontal line corresponds to $y = 1.0$ (perfect estimate).

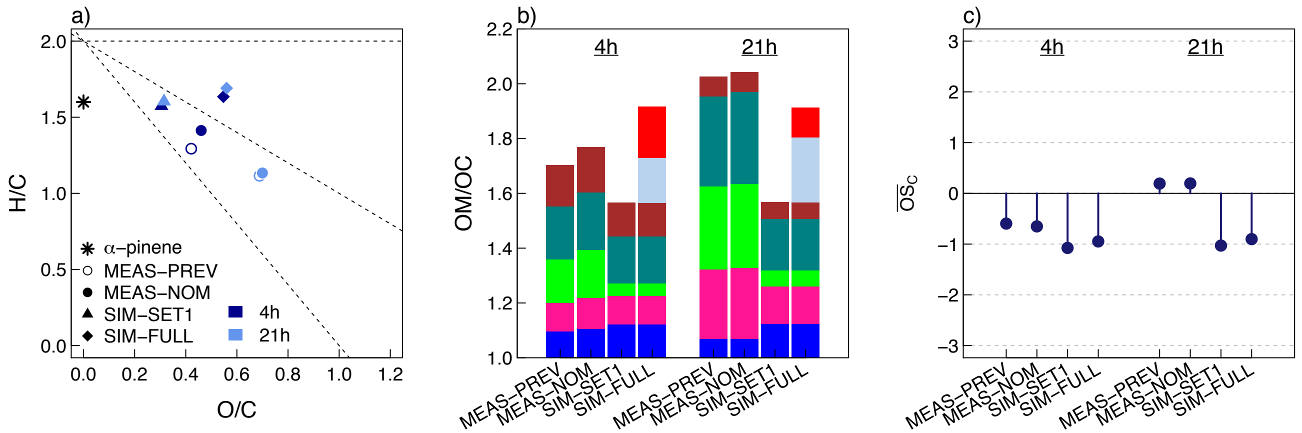


Figure 10. Comparison of measurement (MEAS) and simulations (SIM) for samples ending approximately at 4 and 21 hours (time-integrated over 3.1 to 4.2 hours and 17.6 and 21.6 hours, respectively) after initiation of photochemistry (Sax et al., 2005; Ruggeri et al., 2016). Further details on labels for estimates are defined in Section 3.4. Colors for (b) are the same as for Figure 6, except that ketone and aldehyde has been combined into a single color (teal) because the reported measurements do not differentiate between the two types of carbonyl.

Supporting Information for Technical Note: Relating functional group measurements to carbon types for improved model-measurement comparisons of organic aerosol composition

Satoshi Takahama[1] and Giulia Ruggeri[1]

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

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

S1 Chemoinformatic tools

To construct $\mathbf{Y}$, $\mathbf{\Theta}$, and $\mathbf{X}$ (Section 2.2), we use the APRL-SSP program with a minor modification. Ruggeri and Takahama (2016) showed that there can be a one-to-one correspondence between functional groups (FGs) and non-carbon atoms such that the mapping is unambiguous. Past constraints specified that all atoms must be accounted for by at least one FG:

$$\bigcup_{j \in \mathcal{J}} \{a : a \in \mathcal{A}_i, a \in \mathcal{A}_j\} = \mathcal{A}_i \quad \forall\, i \in \mathcal{M}\ , \tag{S1}$$

and that all non-carbon atoms $\mathcal{A}_i \setminus \mathcal{C}_i$ cannot be matched by more than one group:

$$\bigcap_{j \in \mathcal{J}} \{a : a \in \mathcal{A}_i \setminus \mathcal{C}_i, a \in \mathcal{A}_j\} = \varnothing \quad \forall\, i \in \mathcal{M}\ . \tag{S2}$$

Polyfunctional carbon atoms were not included in their validation at the time, but we now impose an additional condition that each FG includes the definition for only one carbon atom (with exceptions noted in Section S2):

$$\left| \{a : a \in \mathcal{C}_i, a \in \mathcal{A}_j\} \right| = 1 \qquad \forall\, i \in \mathcal{M}, j \in \mathcal{J}\ . \tag{S3}$$

$|\cdot|$ denotes the cardinality of the set. To satisfy this new condition, we split the `C=O-O` group into three separate groups (`R2C=O-O`, `RHC=O-O`, `H2C=O-O`) as carbon was double counted otherwise. This step is inconsequential from a mass perspective, but important for fulfilling the relationship (equation 3) for the complete APIN MCM mechanism. The corresponding patterns have been updated in the APRL-SSP repository.

S2 Generalization of carbon types

We note two generalizations to the carbon type descriptions introduced in Section 2.2 that can be considered. First, these carbon types focus on the functionality of each carbon, but do not consider its complete bonding environment (e.g., configuration to other carbon atoms). For instance, carbon atoms defined by functionalization only by hydroperoxide, alkoxy radicals, and peroxy radicals can differ according to whether the carbon is sp^3 or sp^2-bonded to other carbon atoms. Hydroxyl groups in phenols are differentiated from alcohols in similar instances, but we have not made this distinction for these three groups as nomenclature for them are not common and also does not affect our analysis. It is possible to define SMARTS patterns to make the differentiation in other applications where carbon type representations are useful.

The second generalization concerns FGs that contain skeletal heteroatoms. FGs of this type — specifically in this case, anhydride, ester, and (organic) peroxide — are present in photooxidation products of 1,3,5-trimethylbenzene included in the MCMv3.2 mechanism (Bloss et al., 2005; Ruggeri et al., 2016), and corresponding SMARTS patterns were developed by Ruggeri and Takahama (2016) to match these structures. Equation S3 should accordingly permit two carbon atoms to be associated with each of these exceptional FGs. To accommodate such groups (and other FGs defined by membership of multiple carbon atoms) in our framework, the carbon type formulation can be a) extended to "carbon units" consisting of one or more carbon atoms and their bonded heteroatoms, or b) modified by the introduction of a correction factor. In the latter approach, the carbon-group matrix θ_{kj} can be replaced by $\tilde{\theta}_{kj} = \theta_{kj}\gamma_j$ and group-carbon matrix ϕ_{jk} replaced by $\tilde{\phi}_{kj} = \phi_{kj}\gamma_j^{-1}$, where γ is a coefficient is a correction factor to complete the FG and carbon balances of equations 3 and 4, respectively. $\gamma_j = 0.5$ for these two-carbon FGs, and $\gamma_j = 1$ for the rest (single-carbon FGs). All equalities expressed in this manuscript would hold exactly, except for carbon type oxidation state (equation 8) that will be only approximately true for ester groups (since one carbon atom is double-bonded to oxygen while the other is only singly bonded to another oxygen atom). However, the overall oxidation state estimate (equation 7) still holds when summed over each molecule that contains both carbon atoms of the ester group.

Tables S1–S3 show carbon atom types associated with single-carbon FGs (conversely stated, each FG is uniquely associated with one carbon atom), two-carbon FGs (carbon atoms in these FGs share some heteroatoms with other carbon atoms), and carbon-only structures present in the combined set of molecules from the α-pinene and 1,3,5-trimethylbenzene degradation schemes. In this set of 441 molecules, there are 2867 carbon atoms that can be classified into one of 60 types (labeled in order of frequency, X1–X60, prefixed by character "X" to prevent confusion with carbon type labels used in the APIN simulation) that differ in their association with 30 unique FGs. 46 of these types contain unique FGs (2557 / 2867 carbon atoms belong in this category), 11 of these types share FGs (116 / 2867 carbon atoms belong in this category), and 3 are bonded only to other carbon atoms (194 / 2867 carbon atoms belong in this category). 92% of the carbon atoms in this superset belong to the 41 carbon types (which includes two of the tertiary and quaternary carbon types) from the APIN simulation discussed in the main body of this manuscript, though this relative abundance is reported on a frequency basis and does not consider molecular abundances that might be typical in a SOA mixture. The correspondence of labels used in the main document (numbered by abundance of total carbon during the APIN simulation) and Tables S1–S3 (numbered by frequency of occurrence of in the 441 molecules) are listed in Table S4.

S3 Supporting interpretations for chemical basis sets

When combined with information regarding the carbon skeleton, carbon types presented in this work can provide another origin for derivation of chemical basis sets. In Figure S1, molecular abundances for gas and aerosol phases in the APIN simulation are depicted using carbon types and n_{C}. Together with Figure 2 and definitions in Section 2.2, each of the common basis set dimensions (O, C, H, OS_{C}) used in the aerosol community can be derived. When neighboring interactions among groups are desired, these carbon types can form the basis of multi-carbon unit representations as hinted above.

References

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.

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

Ruggeri, G., Bernhard, F. A., Henderson, B. H., and Takahama, S.: Model-measurement comparison of functional group abundance in α-pinene and 1,3,5-trimethylbenzene secondary organic aerosol formation, *Atmospheric Chemistry and Physics*, 16, 8729–8747, doi:10.5194/acp-16-8729-2016, 2016.

Tables and Figures

Table S1: Functionalized carbon atoms that do not contain multi-carbon FGs (anhydride, ester, and peroxide). Rows are ordered according to OS_C (ascending) and number of occurrences n (out of 2867) (descending). z indicates the FG contribution to oxidation state (equation 7).

		$z \rightarrow$	-1			0		1										2			3				
Label	OS_C	n	aCH	eCH	rCH	formaldehyde	H2C=O-O	aCOH	aldehyde	alkoxyl (*)	CONO2	hydroperoxide	nitro	peroxy nitrate	peroxyl (*)	phenol	RHC=O-O	formic acid	ketone	R2C=O-O	carbonyl peroxy acid	carbonyl peroxy acid (*)	carboxyl (*)	COOH	peroxyacyl nitrate
X58	-4	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X1	-3	720	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X3	-2	366	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X49	-2	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X53	-2	1	3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
X54	-2	1	3	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
X55	-2	1	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X56	-2	1	3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X57	-2	1	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
X60	-2	1	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X4	-1	208	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X8	-1	78	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X10	-1	48	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X11	-1	48	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X28	-1	18	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
X29	-1	18	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X30	-1	18	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
X35	-1	9	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X7	0	96	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X23	0	26	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X24	0	22	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X25	0	21	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
X26	0	21	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
X44	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X48	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X5	1	159	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X15	1	30	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X16	1	29	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X20	1	27	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X21	1	26	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
X22	1	26	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
X36	1	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Table S1: Functionalized carbon atoms that do not contain multi-carbon FGs (anhydride, ester, and peroxide). Rows are ordered according to OS_C (ascending) and number of occurrences n (out of 2867) (descending). z indicates the FG contribution to oxidation state (equation 7).

		$z \rightarrow$	-1			0		1										2			3				
Label	OS_C	n	aCH	eCH	rCH	formaldehyde	H2C=O-O	aCOH	aldehyde	alkoxyl (*)	CONO2	hydroperoxide	nitro	peroxy nitrate	peroxyl (*)	phenol	RHC=O-O	formic acid	ketone	R2C=O-O	carbonyl peroxy acid	carbonyl peroxy acid (*)	carboxyl (*)	COOH	peroxyacyl nitrate
X42	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
X46	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
X2	2	368	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
X38	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
X39	2	3	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
X40	2	3	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X41	2	3	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
X45	2	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X47	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
X9	3	54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
X17	3	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
X18	3	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
X19	3	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
X43	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

Table S2. Functionalized carbon atoms that contain multi-carbon FGs (anhydride, ester, and peroxide). Format follows that of Table S1.

		$z \rightarrow$	-1	1					2		4	6
Label	OS_C	n	aCH	aCOH	alkoxyl (*)	hydroperoxide	nitro	peroxyl (*)	ether	peroxide	ester	anhydride
X33	1	15	1	0	0	0	0	0	1	0	0	0
X14	2	31	0	0	0	0	0	0	0	1	0	0
X32	2	15	0	0	0	0	0	0	1	0	0	0
X34	3*	13	1	0	0	0	0	0	0	0	1	0
X37	3	3	0	0	0	0	1	0	0	1	0	0
X31	4*	17	0	0	0	0	0	0	0	0	1	0
X50	4*	1	1	0	0	0	0	1	0	0	1	0
X51	4*	1	1	0	1	0	0	0	0	0	1	0
X52	4*	1	1	0	0	1	0	0	0	0	1	0
X59	4*	1	1	1	0	0	0	0	0	0	1	0
X27	6	18	0	0	0	0	0	0	0	0	0	1

* Value is approximate since the current SMARTS pattern for ester groups defines memebership of carbon atoms to the entire $-CO_2C-$ substructure, which does not differentiate between the carbon functionalized by ester carbonyl and the other which is not.

Table S3. Carbon atoms not functionalized by any heteroatoms (i.e., only bonded to other carbon atoms). Format follows that of Table S1.

		$z \rightarrow$	0		
Label	OS_C	n	aromatic sp2 carbon	quaternary carbon	tertiary sp2 carbon
X6	0	100	0	1	0
X12	0	47	1	0	0
X13	0	47	0	0	1

Table S4. Correspondence of carbon type labels used in the main document (molecules from APIN simulation, Figure 2) and Tables S1–S3 (molecules combined from α-pinene and 1,3,5-trimethylbenzene degradation schemes). Numbers indicate their relative abundance (in descending order) in APIN simulation (first column) and in combined set of molecules (right column). The character "X" is prepended to labels in the combined set only to prevent confusion with labels used for the APIN simulation.

Carbon type (α-pinene set)	Label (combined set)
1	X1
2	X3
3	X4
4	X2
5	X6
6	X22
7	X8
8	X7
9	X9
10	X17
11	X5
12	X15
13	X23
14	X16
15	X30
16	X26
17	X19
18	X35
19	X11
20	X13
21	X21
22	X47
23	X25
24	X57
25	X18
26	X28
27	X48
28	X60
29	X56
30	X58
31	X49
32	X53
33	X20
34	X54
35	X24
36	X29
37	X43
38	X36
39	X55
40	X44
41	X38

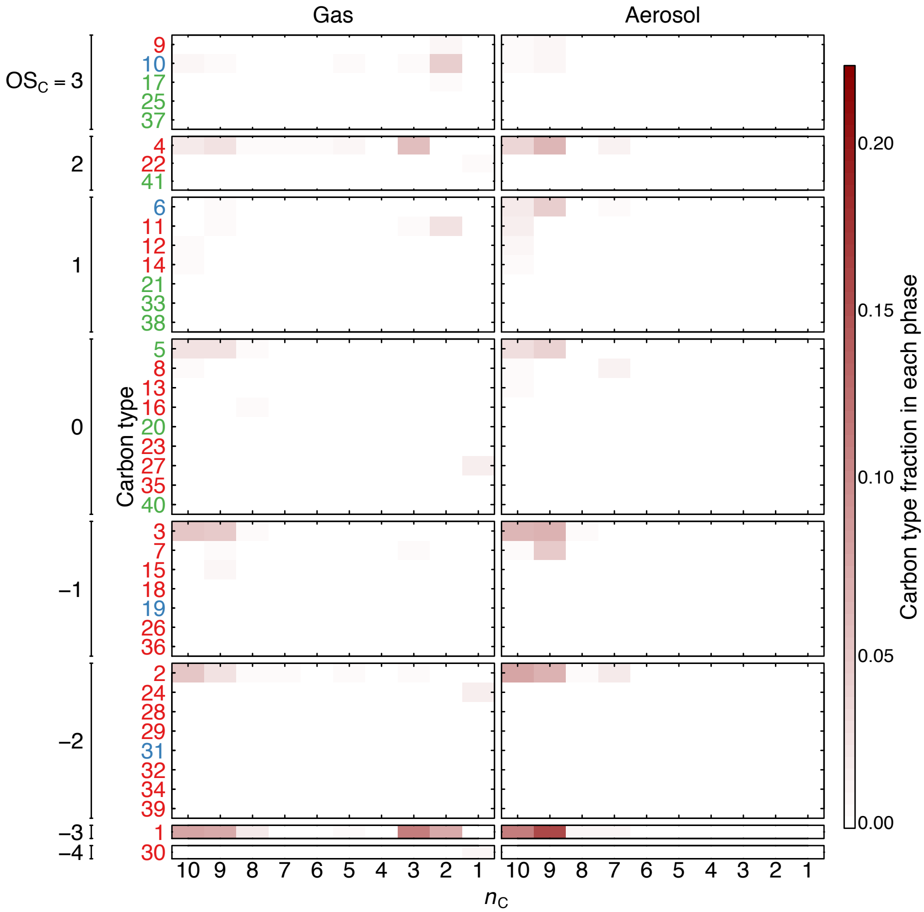


Figure S1. Molecular abundance at $t_{\max\mathrm{SOA}}$ described in terms of their carbon types and number of carbon atoms. The carbon abundance in each grid cell is normalized by the total molar abundance of its phase (gas or aerosol). The colors for the carbon types are the same as in Figure 2.