UCB-GLOBES: An open-access mass spectral database of identified and unidentified atmospheric organic compounds

Lindsay D. Yee,[1] Emily B. Franklin,[2,a] Robin J. Weber,[1] Jessica Zhao,[3] Tiger Zhang,[4] Stephanie Xu,[4] Isaac Santillan,[4] Fangyuan Li,[5] Coty N. Jen,[1,b] Haofei Zhang,[1,c] Yutong Liang,[1,d] Gabriel Isaacman Van-Wertz,[1,e] Rebecca A. Wernis,[2] John Offenberg,[6,f] Michael Lewandowski,[6☆] Taekyu Joo,[7,g] Masayuki Takeuchi,[8] Gamze Eris,[8] Weiqi Xu,[9,h] Nga L. Ng,[7,8,9,i] Yuzhi Chen,[10] John E. Shilling,[10] Mary Alice Upshur,[11] Ariana Gray Bé,[11] Regan J. Thomson,[11] Franz M. Geiger,[11] and Allen H. Goldstein[1,2]

[1] Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, 94720, USA
[2] Department of Civil and Environmental Engineering, University of California, Berkeley, Berkeley, 94720, USA
[3] Department of Industrial Engineering & Operations Research, University of California, Berkeley, Berkeley, 94720, USA
[4] Department of Electrical Engineering & Computer Sciences, University of California, Berkeley, Berkeley, 94720, USA
[5] Department of Statistics, University of California, Berkeley, Berkeley, 94720, USA
[6] Office of Research and Development, Center for Environmental Measurement and Modeling, Atmospheric and Environmental Systems Modeling Division, US Environmental Protection Agency, Research Triangle Park, 27711, USA
[7] School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, 30332, USA
[8] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, 30332, USA
[9] School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta, 30332, USA
[10] Atmospheric, Climate, and Earth Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA
[11] Department of Chemistry, Northwestern University, Evanston, 60208, USA

[a] Now at CSIRO Environment, Aspendale, Victoria 3195, Australia
[b] Now at Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, 15213, USA
[c] Now at Department of Chemistry, University of California, Riverside, Riverside, 92507, USA
[d] Now at Thrust of Sustainable Energy and Environment, The Hong Kong University of Science and Technology (Guangzhou), Guangdong, 511453, China
[e] Now at Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, 24060, USA
[f] Present affiliation: Office of Applied Science and Environmental Solutions, Applied Science and Environmental Methods Division, US Environmental Protection Agency, Research Triangle Park, 27711, USA
[g] Now at Department of Earth and Environmental Sciences, Korea University, Seoul, 02841, South Korea
[h] Now at State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, 100864, China
[i] Now at Division of Engineering and Applied Science, California Institute of Technology, Pasadena, 91125, USA

☆Retired

Correspondence to: Lindsay D. Yee (lindsay.yee@berkeley.edu)

Abstract. Chemical characterization of atmospheric organic aerosols using gas chromatography with 70 eV electron ionization mass spectrometry (GC/EI-MS) has been used for decades in advancing molecular marker detection and identification, though primarily through suspect screening and/or targeted analyses. To advance non-targeted analyses of environmental samples, we have catalogued approximately 27,000 mass spectra (MS) of the trimethylsilyl derivatives of semi-volatile organic aerosol

(OA) analytes in the open-access University of California Berkeley Goldstein Library of Organic Biogenic Environmental Spectra (UCB-GLOBES). Analytes were observed in ambient samples from the U.S. and the Central Amazon and/or laboratory simulations of secondary OA (SOA) formation. These samples are representative of OA under urban and biomass burning influences as well as SOA derived from biogenic precursors (e.g., isoprene, monoterpenes, sesquiterpenes) and biomass burning intermediates. MS are documented in UCB-GLOBES without regard to known chemical identity, annotated with extensive metadata such as sample source/experimental conditions, any structural information gained from MS analyses, and predicted chemical properties such as average carbon oxidation state and carbon number. UCB-GLOBES MS are compatible for importing into NIST MS Search program, and we have also provided a Jupyter Notebook for MS visualization and comparisons. We demonstrate the utility of UCB-GLOBES through MS reanalyses of prior analytes observed in ambient data, finding a 20% reduction in the number of analytes assigned to OA source categories reliant solely on time series correlation and an overall 11% increase in new MS-based OA source categorization for the Southeast U.S. For 1,513 analytes observed previously in the Central Amazon, we found 375 MS matches using UCB-GLOBES vs. 136 MS matches during prior analyses, representing a 14% gain in newly confirmed or newly categorized OA species. While OA from laboratory oxidation experiments in UCB-GLOBES are highly diverse chemically, on average only 29% of UCB-GLOBES MS have a mass spectral match to another MS entry in UCB-GLOBES and/or in databases of known compounds (i.e. NIST MS Database, Adams Essential Oil, MANE Flavor and Fragrance Company). This indicates that roughly 70% of UCB-GLOBES MS are unique thus far, not observed more than once among the laboratory oxidation samples and ambient data in UCB-GLOBES MS. Further, only 18% can be positively identified using these databases or known authentic standards. This points to a large gap between these laboratory simulations and ambient OA. Overall, the UCB-GLOBES database can be utilized for improving confidence in OA source categorization and/or identification, novel chemical marker discovery, tracking chemical diversity, *de novo* structure and properties prediction, and improving MS search and matching algorithms. This can ultimately inform future research priorities for the chemical characterization of atmospheric organic samples.

1 Introduction

Atmospheric organic aerosols are highly complex chemical mixtures, of which most compounds have yet to be chemically identified (Goldstein and Galbally, 2007; Hallquist et al., 2009; Nozière et al., 2015). This is in large part due to the plethora of volatile organic compounds (VOCs) emitted to or formed in the atmosphere that participate in numerous chemical reactions to create many more unique oxidation products (Aumont et al., 2005; C. Ditto et al., 2020; Kanakidou et al., 2005). The multitude of chemicals generated in the gas phase and those that partition into the particle phase, forming secondary organic aerosol (SOA), creates a challenge for probing the chemical makeup of OA. Various analytical techniques have been employed to interrogate the chemical identities of organic compounds within the particle phase with previous observations finding hundreds to thousands of individual compounds within aerosol samples (Franklin et al., 2023; Hamilton et al., 2004; Isaacman-Vanwertz et al., 2017; Isaacman et al., 2014; Kourtchev et al., 2014; Nizkorodov et al., 2011; Wang et

al., 2017; Zhang et al., 2024). Bulk characterization methods through the Aerodyne Aerosol Mass Spectrometer (AMS) can include reporting of elemental ratios (e.g. H:C, O:C) and statistical factor analysis of mass spectra to infer aerosol source-types and origins (Aiken et al., 2007, 2008; Canagaratna et al., 2015; Ulbrich et al., 2009). In addition, resources of characteristic AMS mass spectra available to the community aid in potential chemical classification of OA sources (Jeon et al., 2023). Several molecular-level high-resolution mass spectrometry techniques are employed to obtain molecular formula of analytes (Brophy and Farmer, 2015; Crounse et al., 2006; Lee et al., 2014; Lopez-Hilfiker et al., 2019; Roach et al., 2010a, 2010b), though definitive chemical structure and isomer-specific level identification can only be gained by adaptive configurations and separation (Johnston and Kerecman, 2019; Krechmer et al., 2016; Nozière et al., 2015; Wen et al., 2023). Despite advances in mass spectrometry as well as organic syntheses targeted for atmospheric chemical markers (Gagan et al., 2023), there remains a large fraction of measurable yet chemically unidentified organic compounds in the atmosphere. This problem is chronically plagued by the lack of commercially available standards as well as the extreme difficulty of synthesizing standards of interest. The sheer challenge of measuring (being sensitive to all compound types), much less identifying and quantifying these compounds, typically results in targeted analyses (reporting a specific list of analytes) vs untargeted analyses (reporting all measurable analytes without regard to known identity). Understandably, targeted analyses lessen the burden of increasingly complex data analysis from measurement techniques with increasing capacities. However, this can also lead to a loss in the utilization of available data and unintended potential biases in scientific foci. For example, Franklin et al., (2023) estimated that by restricting analyses to the top hundred analytes by mass observed in the Central Amazon, only ~70% of OA mass associated with either urban influences or biomass burning would be considered with a skew towards biomass burning products.

Discoveries of specific chemical markers in the atmosphere have been born from using gas chromatography with electron ionization mass spectrometry (GC/EI-MS), in which compounds are separated (typically by boiling point) while mass spectra of eluents are also measured to aid in structural identification (Claeys et al., 2004; van Eijck et al., 2013; Lin et al., 2012; Surratt et al., 2007; Szmigielski et al., 2007). In traditional one-dimensional GC, however, separations can be limited by co-elution of compounds with similar properties (e.g., boiling point), leading to convoluted signals and difficulty in acquiring pure compound MS. The application of two-dimensional gas chromatography (GCxGC) has drastically improved the resolving peak capacity compared to that of traditional one-dimensional GC, by performing a second separation (e.g., by polarity) of first-column eluents (Alam et al., 2013; Alam and Harrison, 2016; Goldstein et al., 2008; Hamilton et al., 2006; He et al., 2022; Song et al., 2022; Worton et al., 2012). Additional separation can result in greater chromatographic peak separation and thereby purer MS, allowing for enhanced acquisition of isomer-specific data. For example, Hamilton et al., (2004) isolated over 10,000 individual organic species from a single sample of collected urban aerosol.

Given the wide use of GC/EI-MS for many environmental applications, mass spectral libraries exist for comparing measured spectra to those in open-access and for purchase formats (MassBank, MassBank of North America, Golm Metabolome Database, Adams Essential Oil, MANE2010 flavor and fragrance mass, Wiley MS Library, NIST/EPA/NIH MS Database) to aid in determination of sample compounds' chemical identities (Stein, 2012). Still, most of these resources only

document compounds of known chemical identity or MS generated from authentic standards readily available, with some exceptions of MS data for unidentified compounds in biological samples (Kopka et al., 2005; Mallard et al., 2014). For unidentified compounds or those identified but not provided in state-of-the-art MS databases, it can be challenging for analysers of GC/EI-MS spectra to get a rigorous quantitative comparison such as cosine angle similarity analysis between their query mass spectra and novel mass spectra that are, for example, published as an image in the journal article. An analyst would greatly benefit from having the MS in a data format that allows for automated and computed MS searching in applications like the NIST MS Search program with quantitative calculations that compare the query spectrum against available MS, eliminating the uncertainty in qualitative identification by visual inspection. Qualitative MS comparison can lead to pitfalls in false identification of compounds of similar structure such as isomers or compound families with very similar MS fragmentation patterns (e.g., alkanes), without context of additional data such as Kováts chromatographic retention index, a proxy for elution time of compounds relative to a standard series such as *n*-alkanes that can be compared across instruments (Kováts, 1958).

Thus, as many atmospheric organic compounds have yet to be identified and limited authentic standards exist, we have established an open-access database of 70 eV electron ionization MS acquired with GCxGC from compounds in ambient atmospheric samples and laboratory oxidation experiments of atmospheric relevance without necessitating that chemical identity be known. Initial datasets included those developed in Worton et al. (2017) and Yee et al. (2018). The growing collection was then formally named the University of California, Berkeley Goldstein Library of Biogenic and Environmental Spectra (UCB-GLOBES) in Jen et al. (2019) with ~4,800 MS and this work extends the database to include a total of ~27,000 MS. UCB-GLOBES includes single instance and replicate MS, multiple instances of MS representing the same compound, that spans a collection of ambient samples analysed over the past decade (e.g., urban, rural, remote, biomass burning influenced) as well as laboratory samples simulating oxidation of isoprene, terpenes, and biomass burning emissions under varied chemical and environmental conditions.

MS data along with useful metadata (e.g., chromatographic retention index, origin of MS sample including field campaign/laboratory source and conditions, predicted chemical properties, inferred structural characteristics, predicted MW/chemical formula) are provided without regard to known chemical identity. The overall goal of UCB-GLOBES is to: 1) advance discovery of novel atmospheric and environmental chemical markers of importance, 2) improve the community's awareness and knowledge of compounds that have been regularly measured but lack identification, 3) encourage more open sharing of annotated MS data across the GC/EI-MS community, 4) increase quantitative confidence in identification by MS matching or narrowing down information about an unknown MS (structural similarity, potential chemical origins), and 5) enable further experimental and modelling studies employing mass spectral data. In this work we describe and expand on the MS datasets in UCB-GLOBES curated and released thus far, demonstrate how they can be used for more efficient analysis of sample MS through reanalysis of past ambient datasets, and conclude with discussions on potential avenues for extending UCB-GLOBES and its use in future studies.

2 UCB-GLOBES dataset description

2.1 Method of aerosol samples collection

Pre-baked (550° C for 12 hours) quartz filters were used for collecting various aerosol sample types during several field and laboratory measurement campaigns (Table 1), including the 2013 Southern Oxidant and Aerosol Study (SOAS), the 2014 Green ocean Amazon (GoAmazon) study, western US wildland fuel burns at the U.S. Forest Service Fire Science Laboratory as part of the Fire Influence on Regional and Global Environments Experiment (FIREX2016), and smoke transported to Berkeley, CA during the 2017 Napa, CA wildfires. Additionally, aerosols were also collected from seeded laboratory oxidation experiments of single hydrocarbon precursors including several sesquiterpenes (α-cedrene, α-copaene, β-caryophyllene, α-humulene and α-farnesene) and an essential oil rich in terpenes (copaiba essential oil, Amazon origin) under varied oxidant conditions (H_2O_2 as OH precursor with added NO_x or ozonolysis) in the U.S. Environmental Protection Agency National Exposure Research Laboratory as described in previous work (Jaoui et al., 2013; Yee et al., 2018; Zhang et al., 2018). Ozonolysis of monoterpene species such as α-pinene, d-limonene, and myrcene in the absence of seed aerosol was also tested in an oxidation flow reactor at UC Berkeley (Zhang et al., 2018), while reaction of α-pinene with nitrate radical ($NO_3^{\bullet}$) in the presence of seed aerosol was performed in the Georgia Tech Environmental Chamber (GTEC) facility (Walters et al., 2025). Oxidation of furan biomass burning intermediates, 3-methylfuran with OH and furfural with OH (HONO as OH precursor in both cases), was also conducted in GTEC with seed aerosol (Joo et al., 2024). Lastly, isoprene oxidation via reaction with OH radical (H_2O_2 as precursor, in the absence of added NO_x) was conducted with seed aerosol in the Pacific Northwest National Laboratory chamber facility according to previous methods (Chen et al., 2023). Of note is that the temperature (T) and relative humidity (RH) conditions tested are varied across the laboratory experiments included here, so experiments and accompanying data should not be considered perfect analogues between hydrocarbon precursor systems. A selection of UCB-GLOBES libraries newly added and/or updated in this work along with the number of records in each is summarized in Figure 1.

Table 1: MS datasets included in UCB-GLOBES with associated publications, number of MS, and brief description

MS Library (File) Name and *Associated/Related Publication(s)*	**# of MS**	**Dataset Description**
Field Samples		
Worton_VUV_2017.MSP *Worton et al. (2017)*	201	Compounds observed during the Biosphere Effects on AeRosols and Photochemistry Experiment (BEARPEX) 2009, local biogenic emissions from mixed conifer forests dominated by ponderosa pine trees, aged biogenic emissions from oak trees located several hours upwind and urban outflow from the Sacramento region; underivatized MS with vacuum ultra violet (VUV) ionization

MS Library (File) Name and *Associated/Related Publication(s)*	**# of MS**	**Dataset Description**
GoAmazon2014_terpox_v2.MSP *Yee et al. (2018)*	70	Compounds observed during GoAmazon 2014 wet season confirmed to originate from oxidation of isoprene, a few monoterpenes, and a few sesquiterpenes, "T3" Manacapuru site, AM, Brazil, cleared pasture site located 70 km downwind of urban Manaus and immediately surrounded by secondary forest with clear differentiation between clean conditions versus those with Manaus plume outflow
Soasox *Zhang et al. (2018)*	829	Compounds observed during the Southern Oxidant and Aerosol Study (SOAS) 2013, summertime, Centreville, AL, rural, primarily forested with a mixture of deciduous and coniferous trees with relatively little local influence and occasional urban influence from Tuscaloosa, AL
NAPA_2017_YL *Liang et al. (2021)*	573	Compounds observed during October-November 2017 at UC Berkeley during period of heavy wildfires in Napa Valley, CA
Goamz *Franklin et al. (2023)*	1498	Compounds observed during GoAmazon 2014 wet and dry seasons at "T3" Manacapuru site, AM, Brazil, cleared pasture site located 70 km downwind of urban Manaus and immediately surrounded by secondary forest
<u>Laboratory Burns of Biomass</u>		
FSL_FIREX2016_v2.MSP *Jen et al.* (Jen et al., 2019)	4828	Compounds observed from laboratory burning of common western US wildland fuels at the U.S. Forest Service Fire Science Laboratory as part of the Fire Influence on Regional and Global Environments Experiment (FIREX2016)
FSL_FIREX2016_v2.2024.MSP *McGlynn et al.* (*2025)*	4828	Reanalysis of FSL_FIREX2016_v2 to subtract out PFMD contamination from ~1000 MS as well as new chemical identifications using NIST23.
<u>Surrogate Biomass Burning Laboratory Oxidation</u>		
Mf3ox *Joo et al. (2024)*	353	3-methylfuran + OH (HONO as OH source) + $(NH_4)_2SO_4$ seed, T =22°C, RH =50%
Furfox *Joo et al. (2024)*	233	furfural + OH (HONO as OH source) + $(NH_4)_2SO_4$ seed, T = 22°C, RH = 5%
<u>Sesquiterpene Laboratory Oxidation</u>		
Acopox *Yee et al. (2018); Experiment conducted under same methodology as in Jaoui et al.* (2013) *and Offenberg et al.* (2017)*; Zhang et al. (2018)*	1616	α-copaene + O_3, + $(NH_4)_2SO_4$ seed, T =22°C, RH = 32%

MS Library (File) Name and *Associated/Related Publication(s)*	# of MS	Dataset Description
Aromox *Jaoui et al. (2013); Offenberg et al. (2017); Zhang et al. (2018)*	1016	aromadendrene + O_3 + $(NH_4)_2SO_4$ seed, T = 22°C, RH = 32%
Aromox2 *Jaoui et al. (2013); Offenberg et al. (2017); Zhang et al. (2018)*	1073	aromadendrene + NO_x + hν + $(NH_4)_2SO_4$ seed, T = 25°C, RH = 31%
Ahumox *Jaoui et al. (2013); Offenberg et al. (2017); Zhang et al. (2018)*	1554	α-humulene + O_3 + $(NH_4)_2SO_4$ seed, T = 22°C, RH = 32%
Ahumox2 *Jaoui et al. (2013); Offenberg et al. (2017); Zhang et al. (2018)*	1633	α-humulene + NO_x + hν + $(NH_4)_2SO_4$ seed, T = 24°C, RH = 28%
Acedox *Jaoui et al. (2013); Offenberg et al. (2017); Zhang et al. (2018)*	1090	α-cedrene + O_3, + $(NH_4)_2SO_4$ seed, T = 22°C, RH = 32%
Acedox2 *Jaoui et al. (2013); Offenberg et al. (2017); Zhang et al. (2018)*	1019	α-cedrene + NO_x + hν + $(NH_4)_2SO_4$ seed, T = 24°C, RH = 28%
Bcpox *Yee et al. (2018)*	5	β-caryophyllene aldehyde, β-caryophyllinic acid, β -caryophyllonic acid, β-nocaryophyllone aldehyde, β-nocaryophyllonic acid; synthesized chemical standards by Regan Thomson and Franz Geiger Groups at Northwestern University
Bcpox2 *Jaoui et al. (2013); Offenberg et al. (2017); Zhang et al. (2018)*	1030	β-caryophyllene + NO_x + hν + $(NH_4)_2SO_4$ seed, T = 25°C, RH = 29%
Bcpox3 *Jaoui et al. (2013); Offenberg et al. (2017); Yee et al. (2018)*	1001	β-caryophyllene + NO_x + acidic, $(NH_4)_2SO_4$ and H_2SSO_4, seed + hν, T = 24°C, RH = 29%
Bfarnox *Jaoui et al. (2013); Offenberg et al. (2017); Yee et al. (2018)*	1624	β-farnesene + NO_x + hν + $(NH_4)_2SO_4$ seed, T = 25°C, RH = 31%
Copox *Yee et al. (2018); Experiment conducted under same methodology as in Jaoui et al.* (2013) *and Offenberg et al.* (2017)	1356	copaiba oil + O_3 + $(NH_4)_2SO_4$ seed, dynamic, T = 22°C, RH = 32%

MS Library (File) Name and *Associated/Related Publication(s)*	# of MS	Dataset Description
Monoterpene Laboratory Oxidation		
Apinox *Zhang et al. (2018)*	568	α-pinene + O_3 (high SOA loading), T = 23°C, RH = 30%
Apinox2 *Zhang et al. (2018)*	204	α-pinene + O_3 (low SOA loading), T = 23°C, RH = 30%
Apinox3 *Experiment conducted under same methodology as in Walters et al. (2025)*	1596	α-pinene + NO_3 + $(NH_4)_2SO_4$ seed, T = 22°C, RH = 30%
Limonox *Zhang et al. (2018)*	483	limonene + O_3, T = 23°C, RH = 30%
Myrcox *Zhang et al. (2018)*	447	myrcene + O_3, T = 23°C, RH = 30%
Isoprene Laboratory Oxidation		
Isopox *Experiment conducted under same methodology as in Chen et al. (2023).*	1265	isoprene + OH (H_2O_2 as OH source), T = 23°C, RH = 57%, wet $(NH_4)_2SO_4$ seed

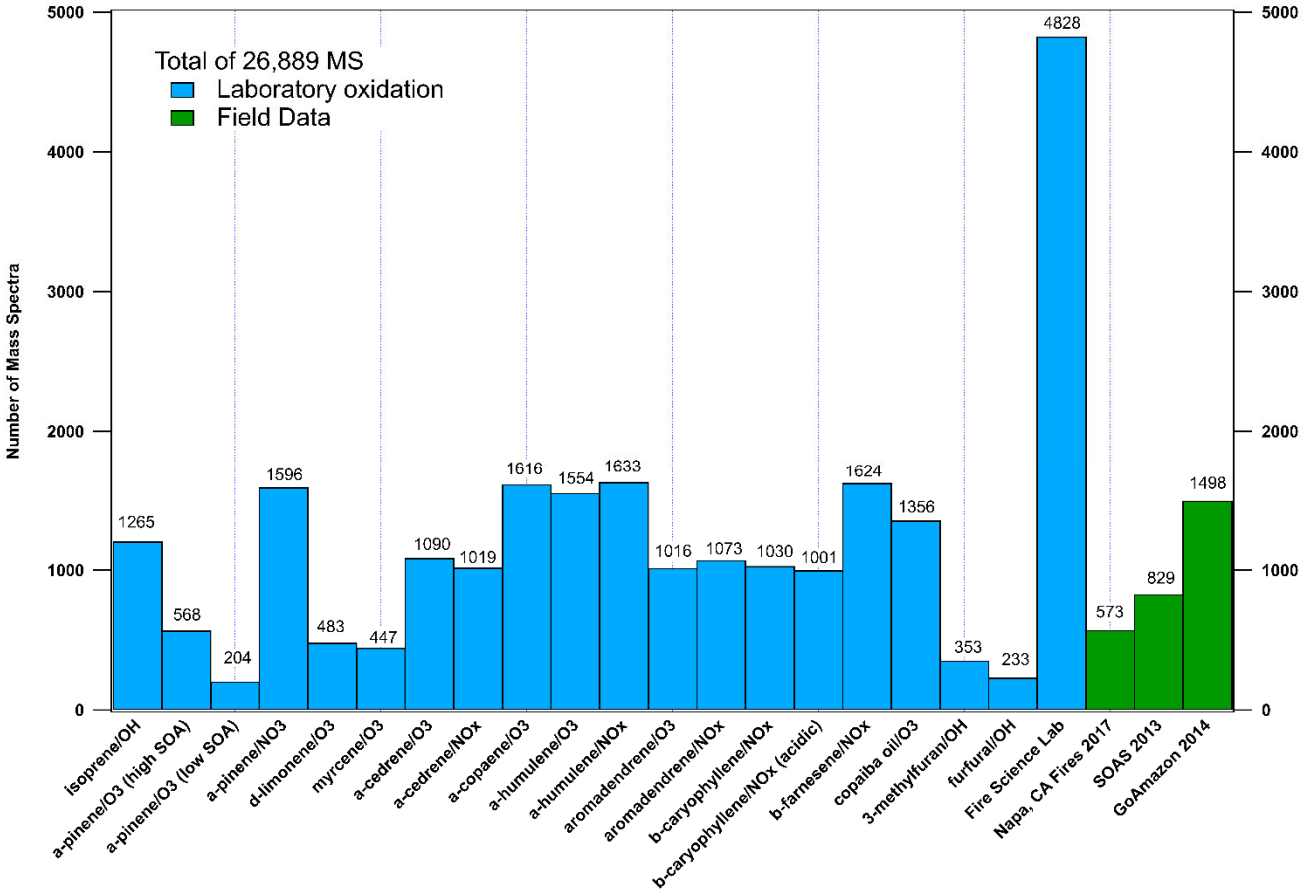


Figure 1: Number of mass spectra (MS) in each library of the UCB-GLOBES database used in this work. Blue bars indicate MS produced from laboratory oxidation experiments and green bars indicate field/ambient datasets.

2.2 GCxGC-HR-TOFMS analysis

All samples were analysed offline using two-dimensional Gas Chromatography with High-Resolution Time-Of-Flight Mass Spectrometry (GCxGC-HR-TOFMS) as employed in previous work (Franklin et al., 2023; Jen et al., 2019; Liang et al., 2021; Worton et al., 2017; Yee et al., 2018; Zhang et al., 2018). Briefly, one or multiple aliquots (0.41 cm^2) of each filter sample were prepared with an applied internal standard. The internal standard is a custom mixture of even carbon number perdeuterated *n*-alkanes spanning C_{12}-C_{38} (C/D/N Isotopes) and isotopically-labelled compounds with similar structures/functionality of atmospheric compounds previously encountered. Internal standards are used for establishing a basis for normalized chromatographic retention times for comparison across samples as well as for accounting for instrument changes in detector sensitivity and matrix effects during sample desorption.

Prepared sample punches were then desorbed into helium using an autosampler (Gerstel TDSA2) and temperature programmed thermal desorption unit (Gerstel TDS3), followed by focusing onto a glass wool liner held at 30 °C in the cooled inlet system (Gerstel CIS4). Contents were then desorbed from the cooled injection system (CIS) with in-situ gas-phase derivatization by MSTFA, *N-methyl-N-trimethylsilyl trifluoroacetamide* (Sigma-Aldrich), before GCxGC separation (Agilent GC 7890). Derivatization converts hydroxyl groups on compounds into trimethylsilyl (TMS) esters, allowing resolution of more polar and oxygenated organics that are not typically GC-amenable. The first GC column is a 60 m × 0.25 mm × 0.25 μm nonpolar capillary column (Restek Rxi-5Sil MS) which leads to separation by volatility. First column effluent is then cryogenically focused on a guard column (1 m × 0.25 mm Restek Rxi) using a dual-stage thermal modulator (Zoex) and injected every 2.3 s onto the second GC column, 1 m × 0.25 mm × 0.25 μm (Restek Rtx-200MS), which leads to separation by polarity. Helium was used as the carrier gas with a flowrate of 2 mL min^{-1} while the GC temperature program comprised of holding at 40 °C for 2 min, followed by ramping 3.5 °C min^{-1} until 320 °C is reached, and finally held there for 10 min. The second column was housed within a secondary oven that was maintained at 15 °C above the main oven temperature. Second column effluent was then transferred to the ionization region of a high-resolution ($m/\Delta m \sim 4000$) time-of-flight mass spectrometer (TofWerk HTOF), where analytes were ionized using 70 eV electron ionization before detection.

2.3 Creation of EI MS libraries

Within UCB-GLOBES are several EI MS libraries, each representing MS compiled from a particular field campaign or laboratory oxidation experiment. Following GCxGC-HR-TOFMS analysis, chromatographic and MS data were imported into the GCImage software (GCImage, LLC), in which initial processing steps included automated background subtraction, blob (3-D chromatographic peak) detection, and MS deconvolution. Chromatographic retention indices (RIs) were initially configured based on the perdeuterated alkane series present in the internal standard, and then mathematically translated via linear fit to the more conventional and widely-used *n*-alkanes based Kovats Retention index (Kováts, 1958), using the nearest in time GCxGC run with co-injections of perdeuterated and normal *n*-alkanes. Unit mass resolution MS at the peak of considered blobs were then extracted and searched within the NIST MS Search program v.2.4 (U.S. Department of Commerce) and exported from the Librarian tab to .MSP format or directly saved from GCImage. Unit mass resolution (UMR) spectra were selected for entry into UCB-GLOBES for computational simplification and are more widely useful to the wider GC/EI-MS community that does not necessarily acquire HRMS data. UMR data has not been shown to decrease the fidelity of MS identification through MS searching (Stein, 2012), although HRMS data is better suited for potential molecular formulae assignment of unknowns and determination of MW, which can be useful in hybrid searching to gain additional structural information. Exported peak table information (e.g. retention times, retention index) from GCImage were combined with MS data in the NIST MS Search program generated .MSP file within a Microsoft Excel worksheet with a custom-written macro for data merging. Additional custom metadata entry by analysts were added before reexport of the curated MS library as a .MSP file (to be compatible for reimport as a custom user library in the NIST MS Search program; see Supplementary Information: Using UCB-GLOBES data within NIST MS Search program for further details). Custom metadata entry fields

include those listed and described in Table 2. During import of UCB-GLOBES MS libraries into the NIST MS Search program, the metadata entries will appear in the comments section of MS records with fields formatted and labelled by "Tags" as according to the NIST MS Search program conventions. An example MS entry and its MS search results as viewed in the NIST MS Search program is shown in Figure S1. Further information on how to access UCB-GLOBES MS contents and version control is described under Data Availability. While creation of UCB-GLOBES MS entries here were generated with one specific instrument and method of sample preparation, the database could readily include deconvoluted MS from additional GC/EI-MS instruments with variations in sample prep, column loadings, and GC conditions. The importance of and quality of associated metadata with any MS entry is essential for helping an analyst discern the suitability and confidence in a MS match, just as similar data is included for known identity compounds within the NIST MS Database when available. For example, because analytes were derivatized here with MSTFA and separated on a nonpolar column, other GC/EI-MS analysts searching for MS matches among UCB-GLOBES should take this into account with their sample preparation method, interpretation of analyte MS, and comparison of retention indices. Contributions of new MS to UCB-GLOBES is welcome and further described in detail in Supporting Information.

Table 2: Metadata entries accompanying each MS entry along with descriptions of each metadata field

Metadata field/tag name	Description
Name	Compound name including trimethylsilyl groups if applicable (e.g., levoglucosan, 3TMS); otherwise, a placeholder name given by an analyst as an unknown compound (e.g. UNK_AHG_HZ_SOAS_394, following the format: UNKnown_PrincipalInvestigatorInitials_AnalystInitials_Dataset_PeakNumber)
Synonyms	Synonyms for compound name. Multiple names may be entered in this column, separated by semicolons (;) as delimiter
Campaign/Experimental Source	Field campaign/laboratory oxidation/standard used
Experimental_Conditions	Description of any experimental conditions that may provide scientific insight into the observation of this compound
UID	"Universal ID", a unique identifier for the MS within UCB-GLOBES (e.g., LDY-22110, following a format of CuratorInitials-MSNumber)
Contributor_ID	Original compound ID assigned by the contributor
Comments	Contributor's comments that may aid in interpretation of the compound MS or an unknown compound structure/source; Any comments on MS quality
Suspected_matches	List of UIDs that are suspected to be potential MS matches, but not necessarily confirmed by an analyzer/additional AI/ML MS match review

Metadata field/tag name	Description
Confirmed_matches	List of UIDs that are reasonably confirmed as MS match by an analyst
Predicted_Matches	List of UIDs that are predicted to be matched by automated MS review
Compound_Structural_Info	Contributor's description of compound structure (e.g. sugar, acid, unknown, etc.)
Formula	Elemental formula including silyl groups if applicable
MW	Molecular weight including silyl groups if applicable
ExactMass	Exact molecular weight including silyl groups if applicable
VUV_Exact_Mass	VUV measured exact mass if applicable
CAS#	IUPAC given CAS# if applicable (identity known)
Ch3MS-RF Cnum.p	Predicted carbon number of underivatized analyte
Ch3MS-RF OC.p	Predicted O:C ratio of underivatized analyte
Ch3MS-RF OSc.p	Predicted average carbon oxidation state of underivatized analyte
Ch3MS-RF VP.p	Predicted log(vapor pressure) atm of underivatized analyte
Column_Type	Class of GC column: polar, non-polar, etc.
Retention_index	*n*-alkane based retention index used in MS searching and RI comparison
d_alkane_RTI	Deuterated alkane-based retention index
n_alkane_RTI	*n*-alkane based retention index
Instrument	Instrument used to obtain MS
Ionization	Ionization energy
Injection_method	Method used for introducing sample into instrument
Derivatization_Agent	Derivatization agent used if applicable
GC_column	GC column description
Oven_temp	Oven program used
Publications	Associated publications that this MS library entry is used in or supporting publications on generation/measurement of this compound
Contributor	Name of person who made the MS entry
Date_of_Entry	Date the entry was made
Num Peaks	Number of m/z peaks for which MS measured
5_Highest_Intensity_MZ	*m/z* values of the top 5 highest intensity peaks in MS
Base_Peak	*m/z* of the highest intensity peak in MS
MS	Mass spectrum

2.3.1 Manual/automated MS Quality Checks

Additional manual/automated quality checks were performed in case of compromised MS quality due to potential missed 2nd dimension blob co-elutions by the software, filtering out known experimental or system artifacts, noisy MS due to insufficient signal, etc. Additional known contamination influencing MS quality included convolution of low polarity peaks coinciding with perfluoromethyldecalin (PFMD, an internal pulsed calibrant during chromatographic runs for MS calibration). Figure S2 shows that while the majority of MS were not flagged for failing any of these MS quality checks, a significant portion (~1,000 MS) within the FSL_FIREX2016_v2 dataset were found to have systematic PFMD contamination in the MS in those chromatographic runs. MS in FSL_FIREX2016_v2 have since been corrected and rereleased as FSL_FIREX2016_v2.2024 to subtract out PFMD influence in the MS and include updated identifications using the NIST MS 2024 database (pre-release) and additional analysis methods described in McGlynn et al. (2025).

While considerable effort has been made to identify and correct for any issues that may lead to decreased quality of the MS and its metadata for scientific use (including measures such as complete omission of bad MS entries), it is imperative that any user of UCB-GLOBES data be generally versed in EI-MS interpretation as would be the case in utilizing any MS databases for increasing confidence in identification. Additional training in environmental (atmospheric) chemistry concepts is useful to confidently use the data for scientific purposes (e.g. ascribing query MS matched with UCB-GLOBES entries to the same source/chemical conditions as those under which a UCB-GLOBES MS entry derived from). For example, as is the case for laboratory oxidation experiments, it is very feasible for lingering trace level contaminants to exist between experiments performed for different precursor systems at detectable levels in GCxGC analyses despite minimal/undetected impact on the main chemistry of focus from the perspective of other aerosol analyses. It is also possible for starting precursors though purchased from chemical suppliers to have small amounts of impurities. In these known cases, some MS were completely omitted based on the knowledge that certain compounds should not be generated during oxidation of a particular precursor or notes in the metadata fields of a MS entry highlight the authors' interpretation of a likely contaminant/artifact. It is the intent that should users of UCB-GLOBES database have additional MS interpretations or proposed corrections to the current entries, that the corresponding author be contacted to improve the database quality through informal peer scientific discourse. We expect UCB-GLOBES to be under continuous evolution including deletion/addition of MS records as well as new metadata for existing entries as parallel analyses provide new information (e.g. new annotated MS datasets that inform past MS data, predictive capabilities of compound properties/identification based on MS features evolve).

2.3.2 Automated metadata entries from MS featurization: MW prediction, base peak, five highest intensity *m/z* ions

An automated algorithm for picking out the potential molecular ion was generated on the basis that trimethylsilyl derivatives in EI MS tend to have a characteristic $[M - CH_3^{\bullet}]^+$ or $[M - 15]^+$ ion (Harvey and Vouros, 2020). For each MS, its peaks were first found by simple neighbour comparisons: an ion was considered a peak if its intensity was higher than the intensities of its left and right neighbours. Then, all pairs of peaks with a 15 amu difference were found; in each pair, the peak

with higher *m/z* was a candidate for the molecular ion. The candidate with the highest *m/z* was selected as the predicted MW ion. If its intensity exceeded the 75th intensity percentile of the MS, it was included in the metadata for that entry under the "MW_Prediction" metadata field. If no candidates were found or the predicted MW ion had intensity below the enforced percentile, the "MW_Prediction" metadata field was set as "0" or empty. In contrast, the "MW" metadata field was populated under the following cases: 1) Positive MS identification resulting in known chemical structure, 2) analyst determined MW from examination of isotopic ratios and the molecular ion, and 3) analyst determined molecular ion from complementary analysis of samples using vacuum ultraviolet ionization (also reported under "VUV_Exact_Mass" and applicable to only a few datasets in UCB-GLOBES). Custom code was also written to report the base peak (*m/z* of the highest intensity peak) and the five highest intensity ions.

2.4 MS matching criteria

In the forthcoming analyses presented in Section 3, we leverage the UCB-GLOBES database for improving knowledge about previously observed but unidentified MS in ambient datasets. By searching MS from each dataset against UCB-GLOBES, additional MS matches were found besides those within the NIST20 MS Database, Adams Essential Oil MS Library, and a proprietary MS library from the flavor and fragrance company, MANE. Generally, MS matches (MSM) were considered a potential match as long as the NIST MS Search program forward match factor (FMF) ≥ 700, reverse match factor (RMF) ≥ 750, and the difference in the query and database MS retention indices ≤ 10. Retention index (RI) comparison for library search was not used in match factor calculation since RI tolerances were checked separately. For matches external to UCB-GLOBES, retention time differences ≤ 30 were considered a match. The match factor (MF) cutoffs used here are considered to provide "fair" matches (Stein, 1995), which were purposefully lenient in the current analyses to not overattribute MS as unique/unknown by using stricter MF thresholds of 800 or "excellent." Some were still considered matched per analyst discretion after manual inspection and accounting for spurious MS contaminant peaks and/or consideration of influences from signal to noise that might lower the MF scores.

2.5 UCB-GLOBES MS Data Visualization and Comparison Tool

An open-access Python-based tool was developed within Jupyter notebook allowing users to download the latest UCB-GLOBES data, plot MS entries, compare MS, and calculate cosine angle similarity indices between two or more MS as a quantitative indicator for similarity that does not require the NIST MS Search program for match factor calculations. Cosine angle similarity values range from 0 (poor/not a match) to 1 (perfect match). Plots for comparing two user-selected MS are generated following the format developed in Jeon et al. (2023) for comparison of AMS mass spectral data. An example output of the MS data visualization and comparison tool for unidentified MS entries by UCB-GLOBES Universal ID (MS Name) LDY-21509 (UNK_AHG_HZ_SOAS_394) and LDY-22110 (1128_1901_blob_1117) is shown in Figure 2. Overall, the MS are highly similar (cosine similarity score = 0.994) and close in RI (only differs by one RI unit), indicating that these MS are likely from the same compound. Furthermore, the MS here are suspected MS matches with 10 additional unidentified MS

from other datasets in UCB-GLOBES including LDY-7521, LDY-14099, LDY-10911, LDY-6980, LDY-4100, LDY-11940, LDY-5510, LDY-21509, LDY-20100, LDY-9654, LDY-9794. These MSs are derived from monoterpene laboratory oxidation, sesquiterpene laboratory oxidation, and ambient datasets with majority biogenic OA sources. A cosine angle similarity matrix is generated for evaluating the MS similarity for all twelve of these sample types (Table S1) indicating that there is high MS similarity (cosine similarity scores > 0.95) between these compounds which can serve as a potential marker of laboratory and ambient terpene oxidation. Specific information on how to use this Jupyter notebook is discussed in the Supplementary Information.

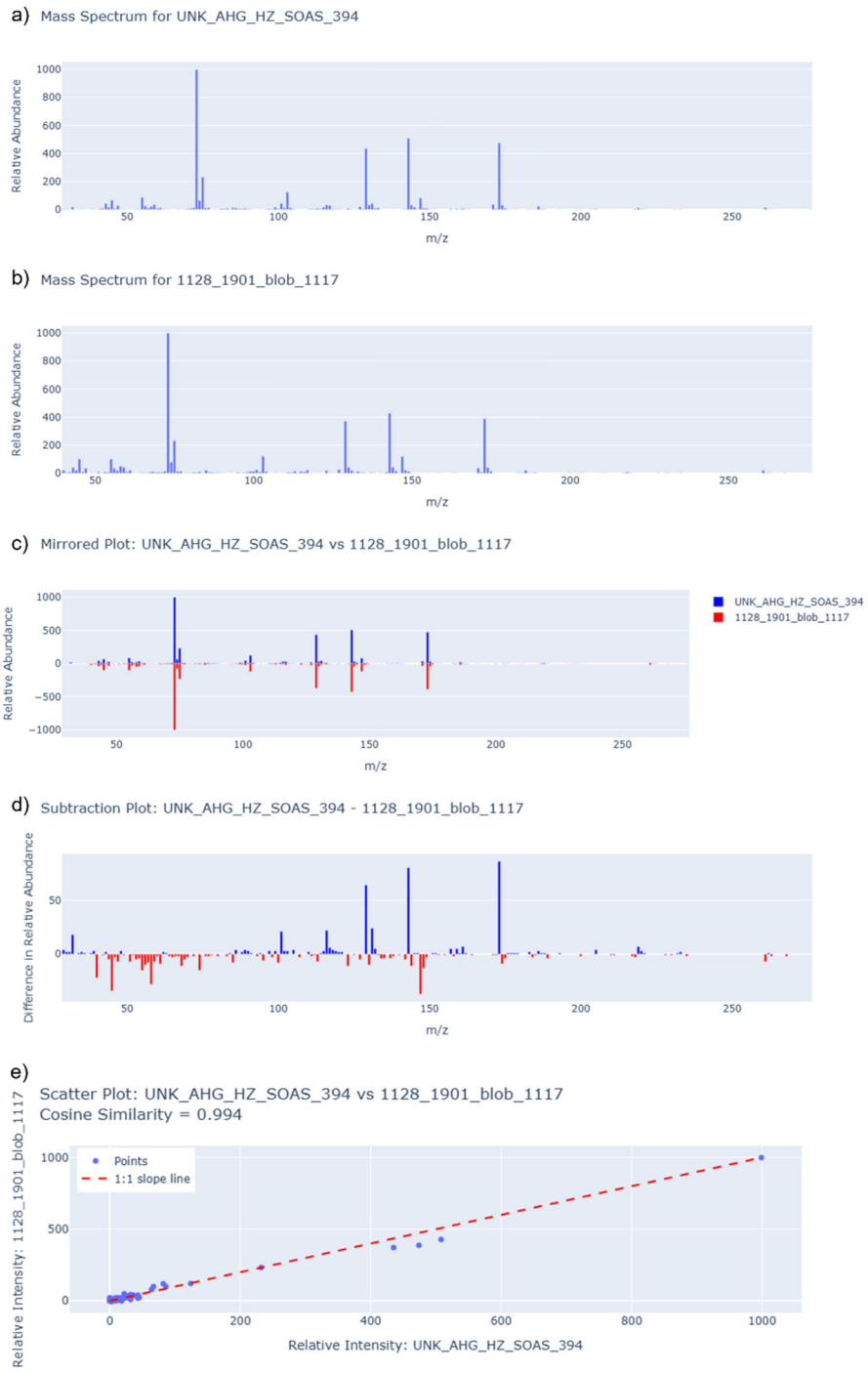


Figure 2: Outputs for MS comparison tool as a) MS for UNK_AHG_SOAS_394 (UID LDY-21509), b) MS for 1128_1901_blob_1117 (UID LDY-22110), c) Mirrored comparison plot for both spectra, d) MS subtraction, and e) Correlation plot of relative intensities with calculated cosine angle similarity between MS.

3 Improving knowledge of unknown atmospheric compounds using UCB-GLOBES

3.1 Progression from targeted vs untargeted tracking of measurable compounds

Previous analysis of GCxGC-HR-TOFMS measurement of compounds sampled during the 2013 Southern Oxidant and Aerosol Study (SOAS) was performed through a semi-targeted approach, in which one filter sample was selected as a representative matrix of compounds to trace throughout the whole 6-week campaign (Zhang et al., 2018). A "template" of ~800 chromatographic peaks stored with their retention times and first dimension RI, and ratios of characteristic ions (manually assigned by the analyst after examining all template MS), was utilized for automated template matching within GCImage for over 300 other filter sample runs. As automated template matching primarily performs matches by measured two-dimensional retention times, this method was highly tedious and required manual quality control to check MS matches according to analyser specified ratios of ion abundances for the assigned characteristic ions. To eliminate the need for manual MS constraints to be assigned by an analyst, we adopted a new workflow in which template peaks' mass spectra along with retention times and first dimension RI are stored in a custom MS library so the peaks within each sample could be searched using their full mass spectra rather than relying on a ratio of two manually chosen characteristic ions as the contingent MS match criteria. This greatly simplified template matching and reduced the time spent in manual quality checks, as MS searches conducted with the NIST MS Search program generate quantitative MS match factors (Stein, 1994) with consideration of retention index matching. We have since generated a custom MS library, Soasox (Table 1), with MS, RI, and metadata entries for the ~800 compounds traced during the SOAS campaign for inclusion in UCB-GLOBES.

By embracing the ability to store full MS data of observed and majority unidentified compounds, we could also generate a consensus template of several compounds viewed over multiple samples during a campaign, rather than select one representative sample that may bias towards time of day/certain chemistry, etc. Such an approach was adopted in Jen et al. (Jen et al., 2019) to generate a dataset with ~4,800 biomass burning marker species spanning over 29 burns featuring varying fuel types burned in a fire laboratory, ~500 species observed in downwind wildfire smoke from Napa, CA (Liang et al., 2021), and ~1,500 species observed during the GoAmazon field campaign (Franklin et al., 2023).

3.2 Improving chemical source category assignment of unknown compounds in ambient data

3.2.1 Source reclassification of compounds observed during SOAS

As an exercise in testing how creation of custom MS libraries can increase the efficiency of classification of unknown compounds to potential sources of chemical origin by MS matching, a MS search was conducted on those of 816 compounds observed during the SOAS 2013 campaign. Previous analysis in Zhang et al. (2018) found 342 MS matches (MSM) among these compounds. Here, we found an additional 91 MS matches (MSM) to UCB-GLOBES entries, and 75 of these were newly classified/reclassified into chemical groups accordingly (Figure 3). This equates to an overall 11% increase in MS match-based confirmations and new information gained about the SOAS dataset. In addition, the number of analytes previously assigned to a chemical source group solely via time series correlation (R^2) with other chemical markers of known chemical

origin dropped by 84; this is a 20% reduction in the number of analytes listed as R^2 for the method of source group determination under the "Group Determined By" column in Table S2 compared to the number listed under the "Group Determined By Reanalysis" column.

Using UCB-GLOBES as an additional reference for added MS confirmations helps fortify inferred chemical source assignments based on time series correlation but importantly leads to potential reassignment of some of the compounds on a basis of chemical structure (through MSM or MS similarity) and origin (MS annotated with campaign/experimental source of MS observation). Moreover, many UCB-GLOBES MS are common between multiple libraries, which means that there is potential for measured compounds to originate from more than one precursor (Figure S3 and Figure S4). This is expected as oxidation of VOCs in the atmosphere typically progresses to organic intermediates of increasing average carbon oxidation state and lower carbon number with CO_2 as an eventual endpoint (Kroll et al., 2011). Thus, many oxidation products along fragmentation pathways (i.e. C-C bond cleavage) will be common among various precursor systems. In the case of this SOAS dataset reanalysis, several occurrences of compound MS having several MS matches necessitates new/hybrid chemical source categories to be employed. For example, Zhang et al. (2018) utilizes single precursor class assignments for biogenic organic aerosols as either BOOA = other biogenic organic aerosols, ISOP = isoprene SOA, MTLab = monoterpene SOA observed in literature and lab experiments, MTR2 = monoterpene SOA based on timeline correlation, or SEQT = sesquiterpene SOA. For figure clarity in this study, we have shortened the MTLab and SEQT designations to MT and SQT hereafter. The reanalysis reveals that for 537 compounds previously ascribed to any of these singular biogenic OA categories, 15 of their MS were in fact crossmatched between multiple libraries/multiple precursor systems. This provides evidence of common oxidation products derived from different precursors. As an example, three MS with UIDs LDY-21129, LDY-21220, and LDY-21183 were found to have corresponding chemical formulae of C8H12O5, C7H8O5, and C7H10O5, respectively. LDY-21220 was originally classified as SQT-derived, while the other two were classified as MT based on timeline correlation. Laboratory oxidation MS in UCB-GLOBES shows these MS can all be generated in MT/SQT oxidation, highlighting that these MS can represent compounds as indistinct chemical makers derived from terpene oxidation. There was even a compound with UID LDY-21336 with proposed molecular formula $C_4H_8O_3$ that crossmatched between MT, SQT, and BBOA (biomass burning OA), which is consistent with previous observations of terpenes and their oxidized derivatives being emitted during biomass burning (Hatch et al., 2019). LDY-21336 also matched to other MS among biomass burning UCB-GLOBES datasets including the Napa, CA fires, FSL_FIREX2016, and laboratory oxidation of biomass burning intermediate 3-methylfuran. Thus additional source group categories reflecting multiple potential sources introduced here include: BBOA/MT/SQT, BBOA/SQT, ISOP/MT/SQT, ISOP/MT, MT, and MT/SQT as shown under the Group Reanalysis column in Table S2 adapted from Dataset S1 (Zhang et al., 2018) and reflected in the Sankey plot (Figure 3). The MS under BBOA/MT/SQT and BBOA/SQT categories could be further interpreted as shared common products derived from terpene oxidation and biomass burning based on Ch3MS-RF predicted carbon numbers <10 and average carbon oxidation states ranging -0.8 to 0 (See Section 3.4 Improving chemical knowledge of unknown UCB-GLOBES compounds through chemical properties prediction). While these hybrid source categories represent an overall small fraction (1.8%) of the SOAS dataset, > 10% changes in the number

of MS assigned to chemical categories in 2018 compared to reassignment through the present reanalysis are non-negligible, with number counts next to each category labelled in Figure 3. Figure 3 shows that the main changes made with 91 new MS match confirmations with UCB-GLOBES data resulted in almost half of those 91 mass spectra getting redistributed from MTR2 to categories like ISOP, MT, MT/SQT, and SQT with MS reanalysis. Changes by percent number of MS within a category include +31% for ISOP, +22% for MT, +34% for SQT, as well as -20% drops each in ASOA (anthropogenic SOA) and in MTR2.

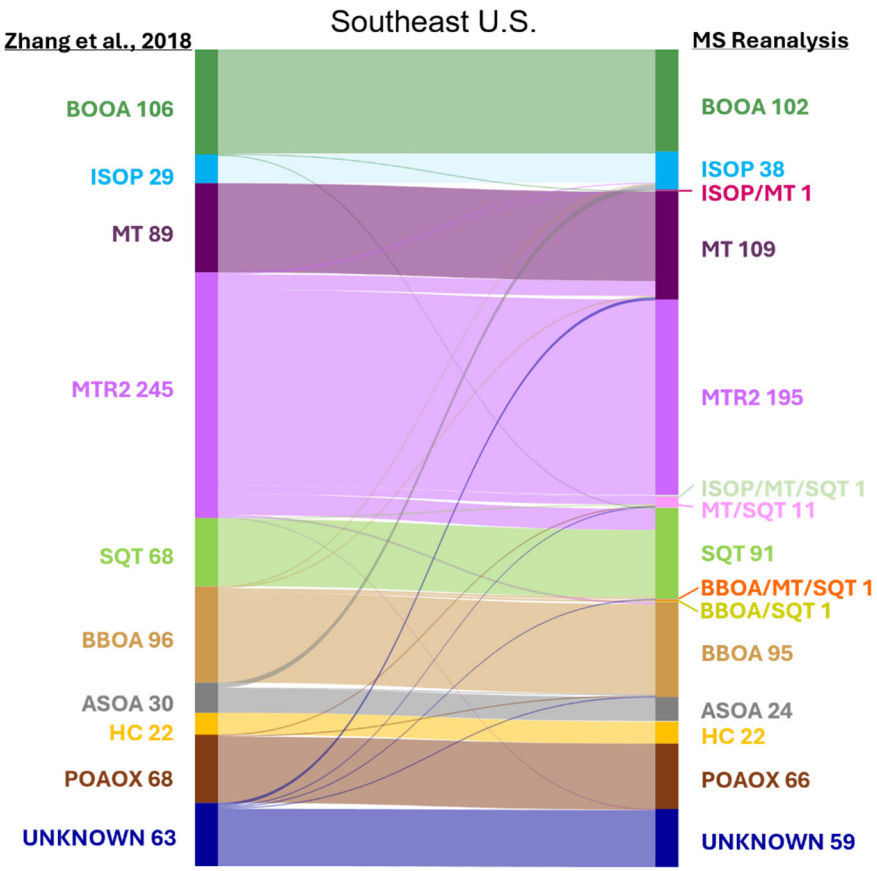


Figure 3: Redistribution of 816 observed compounds during SOAS in the Southeast U.S. by source category based on correlation and mass spectral matching (MSM) performed in Zhang et al., (2018) and reanalysis with MSM utilizing UCB-GLOBES. Reanalysis resulted in 91 new MSM distributed among original and new source categories. Original source categories include other biogenic organic aerosol (BOOA), isoprene-derived organic aerosol (ISOP), monoterpene-derived organic aerosol through MSM (MT), monoterpene-derived organic aerosol inferred by correlation (MTR2), sesquiterpene-derived organic aerosol (SQT), biomass burning organic aerosol (BBOA), anthropogenic SOA (ASOA), hydrocarbon-derived organic aerosol (HC), oxidized primary organic aerosol (POAOX), and unknown. Additional knowledge gained from MSM reanalysis leads to novel hybrid categories from multiple sources including: ISOP/MT, ISOP/MT/SQT, MT/SQT, BBOA/MT/SQT, and BBOA/SQT.

3.2.2 Source reclassification of compounds observed during GoAmazon

A similar MS reanalysis was performed for GoAmazon data, in which original source categorization was performed for 1513 compounds observed during the 2014 wet/dry seasons using hierarchical cluster analyses (HCA) and some MS matching using some terpene laboratory oxidation data in UCB-GLOBES and databases of known compounds (Franklin et al., 2023). Original HCA resulted in three source categories associated with background biogenic (BKGDBIO), biomass burning influence (BB), and urban plume influence (UP). Results of the MS reanalysis are shown in Figure 4a for the wet season and Figure 4b for the dry season. While 136 out of 1513 (9%) of GoAmazon MS were fully identified as originally reported in Franklin et al. (2023), 375 out of 1513 (25%) of the MS in this reanalysis were found to have a MSM when searching among UCB-GLOBES, NIST20, Adams Essential Oil, and MANE databases. Thus, new chemical knowledge for 16% of the dataset is obtained, resulting in further source confirmation determined through HCA or recategorization as shown in Figure 4. Flow links in the Sankey diagram with darker shades within a category indicate MS with "MSM" and lighter shades for MS with "No MSM."

Notable findings for the wet season (Figure 4a) include: 1) 92 MSM out of 544 BKGDBIO determined by HCA were confirmed while 21 BKGDBIO were recategorized into a hybrid category representing BIO/BB influences. For this reanalysis, any MSM with MS entries from laboratory oxidation datasets of isoprene, monoterpene, or sesquiterpenes were categorized as BIO. 2) 16% of the HCA-determined UP category MS (148 out of 390) are further specified as urban plume influence on OA derived from biogenic and/or biomass burning intermediates. These urban plume influence hybrid categories are labelled BIO/UP, BB/UP, and BIO/BB/UP) and constitute ~40% of the urban plume category, underscoring the significant influence the urban plume has on the atmospheric fate of biogenic derived VOCs under these conditions. With additional MS data in the future, it might be possible to further specify the remaining 60% of the urban plume-influence category to determine if the organic carbon is likely from anthropogenic or biogenic sources. Biogenic VOC oxidation was found to be particularly sensitive to anthropogenic emissions in this area (Franklin et al., 2023; de Sá et al., 2017, 2018, 2019; Shrivastava et al., 2019). New experiments focusing on Central Amazon conditions: biogenic VOC precursors (e.g., ~10s of ppt_v monoterpenes) with varying NO_y [0-20 ppb_v] and O_3 [0-40 ppb_v] levels at high relative humidities (80-93 %) and temperatures (26-28 °C) in the absence and presence of anthropogenic VOC precursors and/or primary anthropogenic OA could provide new MS that help elucidate some of the mechanisms behind the compounds in the urban plume-influence category. However, such high relative humidities and low precursor and oxidant concentrations are typically challenging to achieve in a laboratory setting. 3) Lastly, classification was achieved for 64 of 410 MS that were formerly No Assignment (NA) from HCA, 21 of which were reclassified as BKGDBIO, 42 as BB, and 1 as BIO/BB.

In contrast, the dry season (Figure 4b) exhibits more regular biomass burning influence based on the number of MS categorized in BB (196 during wet season, 281 during dry season after MS reanalysis). While 60 MS previously with NA contribute to the increased BKGDBIO (+8%) and BB (+1%), a significant portion of UP MS (38%) is further specified into hybrid categories (i.e. BIO/UP, BB/UP, and BIO/BB/UP). This corroborates the significant impact urban plume influence can

have on SOA formation from biogenic and biomass burning VOCs (Franklin et al., 2023). While modest improvements in MSM (25% of dataset) are achieved, 35% of the dataset remains unclassified from neither HCA nor MSM analyses, underscoring that there is still limited knowledge in reproducing dry season conditions (e.g. exact BBOA sources/BBSOA formation). The MS reanalysis leading to hybrid categories for both wet/dry seasons is not surprising given the multitude of OA sources and conditions sampled during the GoAmazon field campaign. Moreover, there is potential for new chemical markers discovery within these hybrid categories as markers of biogenic OA under urban plume and/or BB influences and processing compared to markers only associated with background conditions.

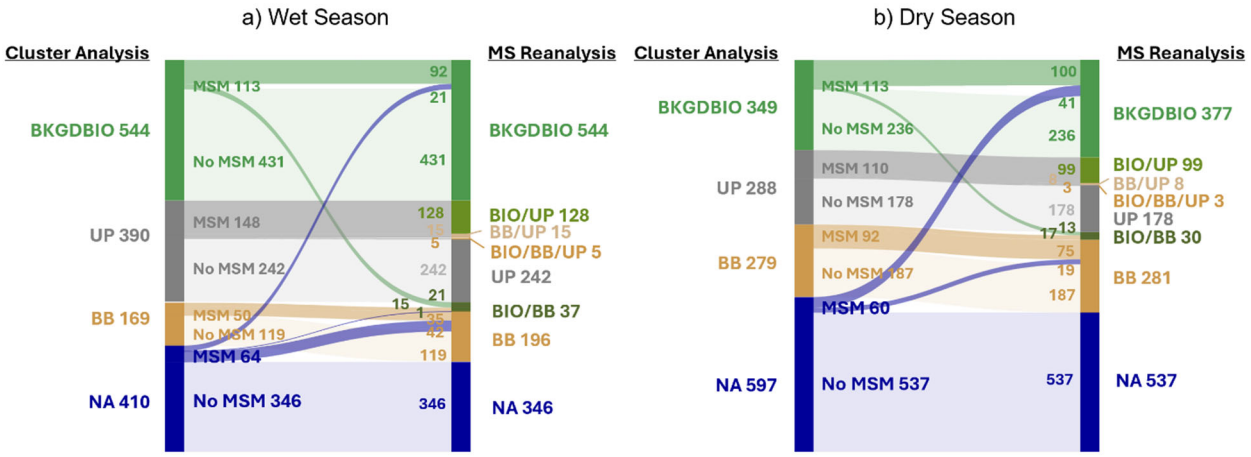
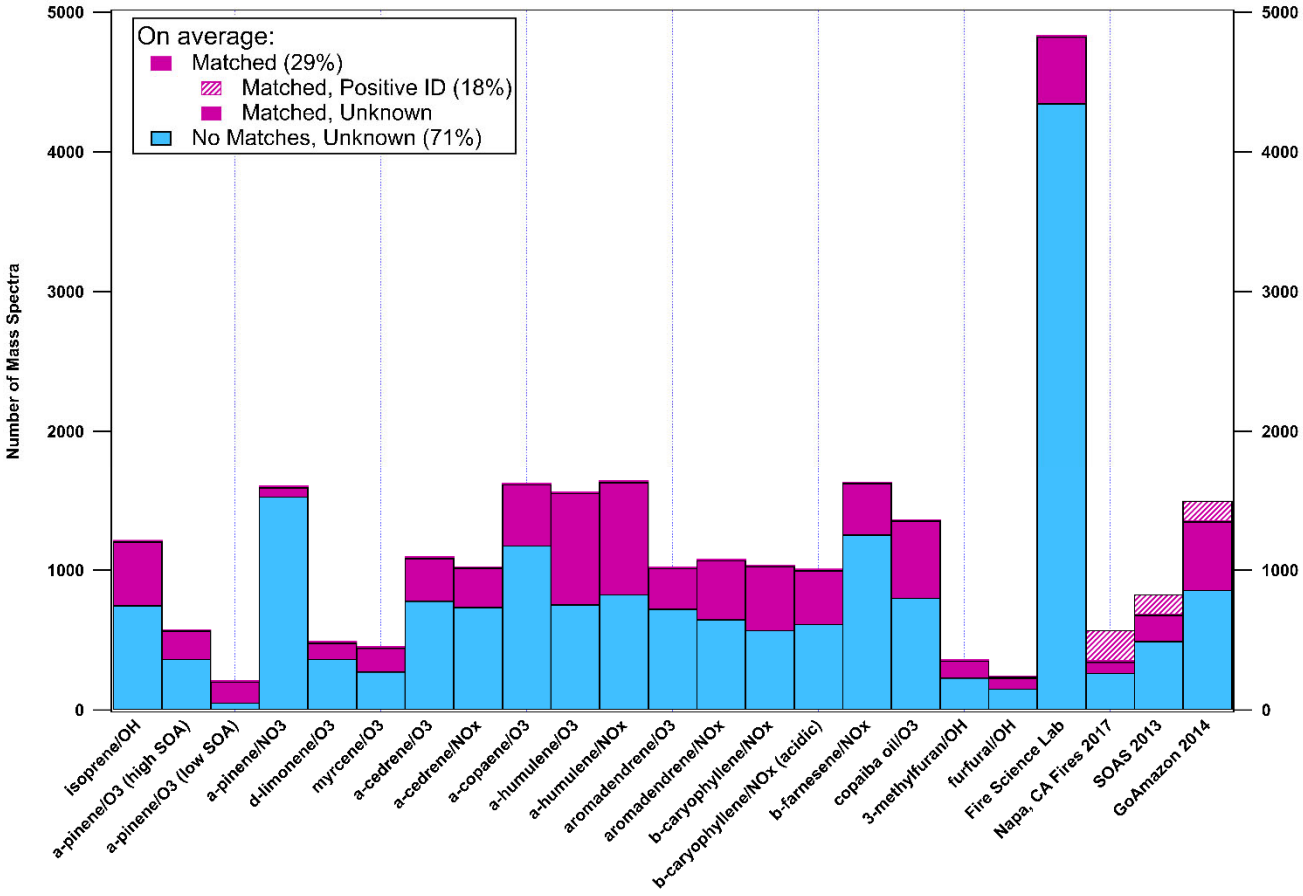


Figure 4: Further specification and source classification of GoAmazon observed compound MS during a) wet season and b) dry season from cluster analyses performed in Franklin et al., (2023) with reanalysis using mass spectral matching (MSM). Source influence categories from cluster analyses include background/biogenic (BKGDBIO), urban plume influence (UP), biomass burning (BB), and no assignment (NA). Additional knowledge gained from MS Reanalysis leads to further specification of urban plume influence on biogenics (BIO/UP), biomass burning (BB/UP), and biogenic/biomass burning (BIO/BB/UP). Additional hybrid category of biogenic/biomass burning (BIO/BB) was also created per MS reanalysis. Darker color links are labelled "MSM" indicating mass spectral match was found; Lighter color links labelled "No MSM" indicate no mass spectral match was found. Outgoing links share the same base color as their cluster analysis source category, while text color of incoming numbers of MS match their destination source category after MS reanalysis.

3.3 UCB-GLOBES MS show large mismatch between laboratory and ambient MS data

Despite generous thresholds for considering potential mass spectral matches (i.e. FMF ≥ 700, RMF ≥ 800), only 29% of all MS in UCB-GLOBES are considered to have a MSM, with the remaining 71% of MS unmatched to other MS within UCB-GLOBES, NIST 20, Adams Essential Oil MS Library, or the MANE Fragrance MS Library (Figure 5). On average, 18% of ambient data MS have MSM amongst an authentic standard or the aforementioned libraries of known compounds leading to positive identification, emphasizing the need to better understand the sources and processing of most atmospheric species catalogued here. This also demonstrates that despite the chemical diversity generated in laboratory oxidation experiments using precursors and oxidants that are deemed to be atmospherically relevant, we have yet to simulate conditions

and processes that generate many atmospheric species that have previously been sampled in the Central Amazon, Southern U.S., and Napa Valley wildfires. Given that biomass burning OA observed during Napa, CA fires, 2017 and biogenic SOA observed during SOAS and GoAmazon typically can contribute significant portions of regional OA budgets, this large mismatch between laboratory and ambient data underscores the need to work towards laboratory simulations that better reveal the sources and origins of ambient OA species. For example, mismatches in simulated relative humidity and OA loadings could be a major source of disconnect (Porter et al., 2021) as well as the relative distribution of RO_2 fate (Kenagy et al., 2024). There are also limited laboratory simulations performed with mixed precursors (Takeuchi et al., 2022) rather than a single precursor, which could allow for cross-system reactions and new SOA products not currently represented in UCB-GLOBES that may better emulate atmospheric chemistry. While UCB-GLOBES MS represent oxidation of several major biogenic precursors and biomass burning intermediates, we recognize that the laboratory simulations are by no means exhaustive of all ambient SOA precursors and OA types.



455

Figure 5: Number of mass spectra (MS) in each library of UCB-GLOBES database coloured by match spectral match status. Blue bars (on average 71%) indicate no mass spectral matches found among UCB-GLOBES, NIST20, Adams Essential Oil Library, MANE Fragrance Library. Magenta bars indicate a mass spectral match (on average 29%) with FMF ≥ 700, RMF ≥ 800, and RI difference < 10 for UCB-GLOBES hits and < 15 for external library hits. Hashed magenta bars indicate mass spectral match with
460 **a known compound in external libraries and/or with an authentic standard leading to positive identification (on average 18%).**

3.4 Improving chemical knowledge of unknown UCB-GLOBES compounds through chemical properties prediction

Since the majority of ambient MS are not matched with available MS databases, we further augment the MS entries by including predicted chemical properties: average carbon oxidation state (Avg. OS_c), carbon number (N_c), oxygen-to-carbon
465 ratio (O:C), and hydrogen-to-carbon ratio (H:C) using Ch3MS-RF (Franklin et al., 2022). Briefly, Ch3MS-RF is a random forest model developed and trained on an authentic standard of ~130 compounds sampled with the same setup in this study. Mass spectral features and first-dimension retention index are utilized for chemical properties prediction of the underivatized analyte, the results of which are included as metadata entries accompanying each MS in UCB-GLOBES. The compounds

used in training were selected to be as representative of the chemical complexity of ambient organic aerosol species as possible with available authentic standards. They span wide-varying volatility, functionality (e.g., alkane, alkene, carboxylic acid, alcohol, aldehyde, ketone, ester, amide), structure (e.g., polyaromatic hydrocarbons, phenol, methoxyphenol, quinone, sugar, sterol), and source types (e.g., known terpene- and isoprene-derived oxidation products). Note that Ch3MS-RF predictions here are inherently limited to compounds amenable to this GC x GC analysis (e.g., organosulfates and organonitrates would be excluded). Franklin et al., (2022) demonstrated Ch3MS-RF to have mean absolute error accuracy of ± 0.25 for Avg. OSc and ± 1.8 for N_c when applied to the test case (~130 authentic standards) and the extrapolation case (71 identified GoAmazon compounds). As random forest modelling does not extrapolate and tends to underpredict property extremes of Avg. OS_c and N_c space (Franklin et al., 2022), we might expect a wider spread of the true chemical properties' values and higher uncertainty in predictions for any encountered MS with very different MS features from the training and extrapolation datasets.

All UCB-GLOBES entries with predicted Avg OS_c and N_C are plotted in Figure 6, demonstrating the variability in these properties by sampled compounds and by source type as previously observed through aerosol mass spectrometry and high resolution mass spectrometry methods (Kroll et al., 2011). For the experimental conditions and field data represented in Figure 6, we note the following observations by panel: a) despite different oxidation conditions and precursors, there is a similar range of predicted Avg. OS_c and N_c for the studied sesquiterpene ($C_{15}H_{24}$) systems, with a good number of predicted $N_c > 15$ suggesting prevalence of accretion type reactions; b) copaiba oil (an essential oil comprised mainly of sesquiterpenes and some monoterpenes) has a similar spread in Avg. OS_c and N_c as the sesquiterpene systems in a) and as the monoterpene ($C_{10}H_{16}$) systems in c) for $N_c < 20$; c) α-pinene reaction with NO_3 radical is quite different compared to ozonolysis, leading to much higher N_c similar to the $N_c > 15$ range achieved by sesquiterpene oxidation; d) there is a wide variety of chemical properties for compounds sampled during laboratory burns of Western U.S. biomass fuel types during FIREX_2016, whereas ambient sampling downwind of Napa, CA 2017 wildfires are more disparate across the chemical space like laboratory oxidation of biomass burning intermediates 3-methylfuran and furfural; e) isoprene (C_5H_8) oxidation occupies a smaller range in N_c, such as monoterpene ozonolysis systems, however there is clearly a signature of extensive oligomerization/accretion type reactions with oxidation products of $N_c > 5$. While several $N_c > 5$ oxidation products have been observed from isoprene oxidation in lab/ambient data, the conditions for this experiment (isoprene, H_2O_2, UV light, wet ammonium sulfate seed, and 57% RH) were intended to simulate SOA formation from isoprene where IEPOX processes are minor, suggesting there are several potential oligomeric species generated from this route under the conditions tested; f) lastly, comparison of ambient datasets from GoAmazon and SOAS show similar occupation of the chemical space for $N_C < 20$ consistent with biogenic SOA formation from isoprene as in e) and monoterpenes as in c), though GoAmazon has greater numbers of compounds with higher predicted $N_c > 25$, which could be consistent with greater prevalence of sesquiterpene oxidation/α-pinene NO_3 oxidation (Figure 6a/c), and higher avg. $OS_c > 0.5$ for $N_c > 12$, which could be consistent with biomass burning influences as shown in Figure 6d. These potential sources are supported by online measurements of the prevalence of sesquiterpenes and their oxidation products during GoAmazon (Yee et al., 2018) relative to the Southeast U.S. during SOAS and prior characterization of OS_c and N_c for biomass burning intermediates and their oxidation (Cai et al., 2020; Gilardoni et al., 2016; Sun et al., 2010).

Regardless of the many overlaps of observed compounds in these datasets with the OS_c and N_c chemical space, we interpret these overlaps as potentially useful indicators of similar chemical processes rather than that of exact mechanisms/sources of observed species, as evidenced by the fact that mass spectral matches are not found for many of the data. Thus, while our predicted OS_c and N_c values here are well within the ranges appropriate by sources previously proposed and observed in Kroll et al. (2011), exact structure matching (as is only achievable through isomer-specific measurements) is still necessary to validate mechanistic hypotheses. UCB-GLOBES thus can serve as an important tool for testing and evaluating chemical mechanisms for observed compounds.

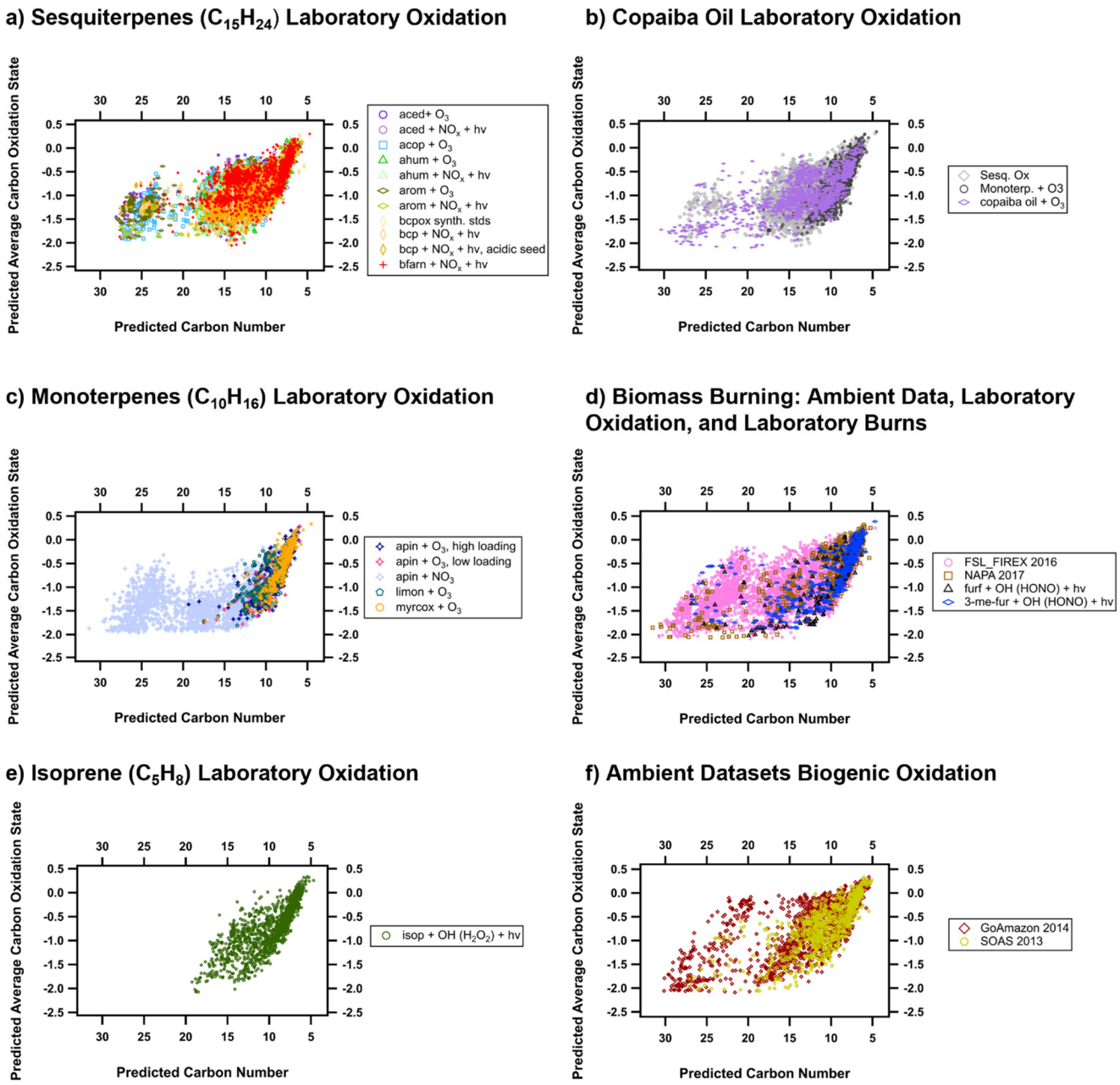


Figure 6: Ch3MS-RF predicted average carbon oxidation state and predicted carbon number for UCB-GLOBES datasets: a) sesquiterpene oxidation systems, b) copaiba oil oxidation, c) monoterpene oxidation systems, d) biomass burning influenced ambient data (NAPA 2017) and laboratory oxidation of biomass burning intermediates (furfural, 3-methylfuran) or Western U.S. fuel types (FSL_FIREX 2016), e) isoprene oxidation system, f) ambient datasets from Central Amazon, GoAmazon and Southeast U.S., SOAS. Data in panels a) and c) repeated in grey in background of panel b) for comparison since copaiba oil is a complex mixture of terpenes.

4 Conclusions

UCB-GLOBES now includes just under 27,000 MS from laboratory and ambient datasets for the purposes of providing: 1) annotated MS data for the wider GC/EI-MS community to use in improving confidence in identification and/or revealing potential origins and sources of organic species in the atmosphere and wider environment, 2) MS data that can be further used in testing and evaluation of MS deconvolution, matching and search algorithms (McGlynn et al., 2025), and *de-novo* chemical structure/properties prediction, 3) development of atmospheric chemical mechanisms and chemical diversity studies, and 4) insights to motivate future research priorities in the identification of novel chemical markers. We provided as examples the utility of UCB-GLOBES in enhancing confidence and accuracy of OA source categorization of real atmospheric species from prior field studies in the Southeast U.S. (SOAS) and the Central Amazon (GoAmazon). One of the most significant advances of using UCB-GLOBES is that for a matched analyte spectrum, there are often several MSMs or closely related MS pointing to the same source(s) (e.g. precursor/s from laboratory oxidation) which strengthens overall confidence in OA source categorizations.

In contrast, statistical techniques such as positive matrix factorization (PMF) have been applied to aerosol mass spectrometry data, in which a timeline of composite mass spectra (also with 70 eV ionization) of all non-refractory flash-vaporized aerosol compounds is generated. In other words, "bulk" OA chemical composition is measured, rather than mass spectra of each individual OA analyte. These timelines of composite mass spectral data are separated into statistical factors, each with an average characteristic mass spectrum generated from the weighted signal contributions of the aerosol compounds' mass spectral fragments assigned to this factor. Thus, a PMF factor mass spectrum will never be a match to any one compound's mass spectrum. Determination of a "specific" source often relies on factor timeline correlation with known chemical tracers measured by complementary contemporaneous techniques and/or evidence of one or a few enriched characteristic ions within the factor mass spectrum that are also enriched in the MS of an authentic standard or known chemical tracers. For example, the isoprene epoxydiol-derived SOA (IEPOX-SOA) factor is characterized by enrichment of m/z = 82 from $C_5H_6O^+$ ion (Robinson et al., 2011), which was later observed in AMS spectra of IEPOX rearrangement products, cis- and trans-3-methyltetrahydrofuran-3,4-diols (Lin et al., 2012). This m/z = 82 ion is also observed in 3-methylfuran oxidation with NO_3 radical, which makes sense given the furan structures in both systems (Joo et al., 2019). Yet, as isoprene is more prevalent in the atmosphere compared to 3-methylfuran, most of the signal associated with m/z = 82 ion tends to be attributed to IEPOX-SOA.

While the majority of UCB-GLOBES data derives from laboratory simulations, we keep in mind that such simulations do not perfectly replicate environmental conditions (e.g., temperature, relative humidity) typical of many regions in the real atmosphere, and additional physicochemical properties (e.g., aerosol phase state, seed chemistry, volume to surface area ratio, etc.) cannot be perfectly controlled/simulated as complexly as in the real atmosphere. Thus, care and general knowledge of atmospheric chemistry must be taken in interpreting the applicability of mass spectral matches to real-world conditions. For example, while there are many cases where MS matches can indicate multiple potential sources of an observed compound, all

these inferred sources may not be atmospherically relevant and/or there could be yet to be measured systems and conditions that produce the same compound. The precise number of sources that leads to a compound, especially a later-generation structure with little similarity to its parent precursor could be numerous beyond what has been included in the current database.

Future directions for this work include additional MS data to be added to UCB-GLOBES including new field data from anthropogenically influenced areas outside Houston, TX and at the Southern Great Plains, OK as well as additional laboratory oxidation simulating aqueous BBSOA compositions. We encourage others employing GC/EI-MS analyses to consider contributing to UCB-GLOBES (see Supporting Information, III. How to Contribute MS Data to UCB-GLOBES) and/or sharing similarly annotated and formatted MS data to address the growing requests of open-access MS libraries and to move untargeted analyses of environmental samples forward. While the majority of atmospheric SVOC species remain unidentified, our future analyses will include clustering and molecular networking of species by chemical similarity (e.g. chemical families), to gain a greater appreciation for real-world chemical diversity of atmospheric SVOCs, their potential sources, and to identify future targeted research priorities on unidentified species of importance.

From the present analyses of UCB-GLOBES, we recommend a few research priorities going forward to advance understanding of the chemistry and fate of organics in the environment. Firstly, the community would benefit from openly shared quality annotated GC/EI-MS mass spectra of a wide range of environmental and laboratory sample types. Second, we would benefit from technical advancements to conduct and catalogue organic species in atmospheric laboratory simulations that are closer to atmospheric conditions with higher relative humidities (> 75 % RH), lower organic aerosol mass loadings, and lower concentrations of precursors and oxidants. We also recommend new datasets simulating oxidation of a mixture of VOC precursors and their oxidation intermediates, as cross-VOC system reactions as would happen in the atmosphere are not sufficiently represented in UCB-GLOBES yet. Inherent to these recommendations are challenges to technical feasibility, ensuring experiments are at sufficient concentrations for detection, and trade-offs of adding too many variables to the experimental matrix. However, achievements toward these recommendations could provide some of the keys to elucidating the true mechanisms and chemical complexity in ambient observations. Finally, we hope that UCB-GLOBES and other available environmental 70eV EI-MS data will be utilized for comparisons across environmental reservoirs (e.g., lithosphere, hydrosphere, biosphere, and atmosphere), so that the emission, transport, and fate of organic carbon across these sectors can be traced and better quantified through processes such as emission, reaction, and deposition.

Data Availability

All datasets in UCB-GLOBES at time of publication are annotated and formatted according to the above described methods and accessible at https://sites.google.com/berkeley.edu/goldstein-lab/resources. Three formatting exceptions include the Worton_VUV_2017 (Worton et al., 2017), GoAmazon2014_terpox_v2 (Yee et al., 2018), and FSL_FIREX2016_v2 (Jen et al., 2019), which were generated before the full standardization of the methods and curation protocols described in this work.

The version of compiled UCB-GLOBES MS data used in this study has the filename ucbglobes2025_v1 with DOI: 10.5281/zenodo.18176760. Versions are denoted by "_v#" notation.

Interactive Computing Environment

The Jupyter Notebook with MS Data Visualization and Comparison Tool is shared online at https://github.com/lyee002/UCB-GLOBES-MS-Data-Visualization-and-Comparison-Tool with DOI: 10.5281/zenodo.18177255 for adaptation and use in analysing MS data from UCB-GLOBES. Additional information on using the Jupyter Notebook is provided in the Supporting Information.

Author Contribution

LY and AG designed the experiments. Field samples were collected by LY, GI, RAW, CJ, YL and analysed by LY, EF, HZ, CJ, YL. Laboratory oxidation samples were analysed with GCxGC by LY, EF, HZ, and RJW. GTEC experiments were designed and carried out by NN, TM, TJ, GE, and WX. PNNL chamber experiments were designed and carried out by YZ and JS. Chemical standards were synthesized by MU, AGB, RT, and FG. LY, RJW, JZ, TZ, FL, SX, and IS developed mass spectral analyses codes and spectral data analyses were performed by JZ, TZ, and LY. LY conducted Ch3MS-RF simulations in consultation with EF. LY prepared the manuscript with contributions from all co-authors.

Competing interests

The authors declare that they have no conflict of interest.

Disclaimer

This work is not a product of the U.S. Government or the U.S. Environmental Protection Agency, and the authors are not doing this work in any governmental capacity. The views expressed are those of the authors only and do not necessarily represent those of the U.S. Government, or the U.S. EPA. It has not been subject to the Agency's review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred. Any mention of trade names, manufacturers or products does not imply an endorsement by the United States Government or the U.S. Environmental Protection Agency. EPA and its employees do not endorse any commercial products, services, or enterprises.

Financial support

The UC Berkeley team gratefully acknowledges support from DOE-ASR program (DE-SC0020051) for further analysis of the GoAmazon 2014/15 field campaign dataset and laboratory oxidation samples for addition of MS datasets to the UCB-GLOBES database. Support from DOE-ASR DE-SC0024032 and NSF AGS-2308645 are also gratefully acknowledged for supporting additional method development and analysis of UCB-GLOBES mass spectral data. TJ, MT, GE, and NLN acknowledge support from NSF AGS-1830727 and NOAA NA18OAR4310112. Pacific Northwest National Laboratory (PNNL) authors were supported by U.S. Department of Energy's (DOE) Atmospheric System Research, an Office of Science (SC) Biological and Environmental Research (BER) program. PNNL is operated for the DOE by Battelle Memorial Institute under Contract DE-AC05-76 RL01830. The Northwestern University team was supported by the National Science Foundation (NSF) under grant no. CHE-1607640. Ariana Gray Bé and Mary Alice Upshur gratefully acknowledge support from NSF Graduate Research Fellowships. Mary Alice Upshur also acknowledges an NSF GROW award, National Aeronautics and Space Administration Earth and Space (NASA ESS) Fellowship, and a P. E. O. Scholar Award.

References

Aiken, A. C., DeCarlo, P. F. and Jimenez, J. L.: Elemental Analysis of Organic Species with Electron Ionization High-Resolution Mass Spectrometry, Anal. Chem., 79(21), 8350–8358, doi:10.1021/ac071150w, 2007.

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, Environ. Sci. Technol., 42(12), 4478–4485, doi:10.1021/es703009q, 2008.

Alam, M. S. and Harrison, R. M.: Recent advances in the application of 2-dimensional gas chromatography with soft and hard ionisation time-of-flight mass spectrometry in environmental analysis, Chem. Sci., 7(7), 3968–3977, doi:10.1039/C6SC00465B, 2016.

Alam, M. S., West, C. E., Scarlett, A. G., Rowland, S. J. and Harrison, R. M.: Application of 2D-GCMS reveals many industrial chemicals in airborne particulate matter, Atmos. Environ., 65, 101 [online] Available from: https://ac.els-cdn.com/S1352231012009685/1-s2.0-S1352231012009685-main.pdf?_tid=091536f0-abd0-11e7-94be-00000aab0f6b&acdnat=1507429746_11828a58fc0beafe30fdef13916f44a1 (Accessed 7 October 2017), 2013.

Aumont, B., Szopa, S. and Madronich, S.: Modelling the evolution of organic carbon during its gas-phase tropospheric oxidation: development of an explicit model based on a self generating approach, Atmos. Chem. Phys., 5(9), 2497–2517, doi:10.5194/acp-5-2497-2005, 2005.

Brophy, P. and Farmer, D. K.: A switchable reagent ion high resolution time-of-flight chemical ionization mass spectrometer for real-time measurement of gas phase oxidized species: characterization from the 2013 southern oxidant and aerosol study, Atmos. Meas. Tech., 8(7), 2945–2959, doi:10.5194/amt-8-2945-2015, 2015.

C. Ditto, J., Joo, T., H. Slade, J., B. Shepson, P., L. Ng, N. and R. Gentner, D.: Nontargeted Tandem Mass Spectrometry Analysis Reveals Diversity and Variability in Aerosol Functional Groups across Multiple Sites, Seasons, and Times of Day, Environ. Sci. & Technol. Lett., 7(2), 60–69, doi:10.1021/acs.estlett.9b00702, 2020.

Cai, J., Zeng, X., Zhi, G., Gligorovski, S., Sheng, G., Yu, Z., Wang, X. and Peng, P.: Molecular composition and photochemical evolution of water-soluble organic carbon (WSOC) extracted from field biomass burning aerosols using high-resolution mass spectrometry, Atmos. Chem. Phys., 20(10), 6115–6128, doi:10.5194/ACP-20-6115-2020, 2020.

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(1), 253–272, doi:10.5194/acp-15-253-2015, 2015.

Chen, Y., Zaveri, R. A., Vandergrift, G. W., Cheng, Z., China, S., Zelenyuk, A. and Shilling, J. E.: Nonequilibrium Behavior in Isoprene Secondary Organic Aerosol, Environ. Sci. Technol., 57(38), 14182–14193, doi:10.1021/acs.est.3c03532, 2023.

Claeys, M., Graham, B., Vas, G., Wang, W., Vermeylen, R., Pashynska, V., Cafmeyer, J., Guyon, P., Andreae, M. O., Artaxo, P. and Maenhaut, W.: Formation of secondary organic aerosols through photooxidation of isoprene, Science (80-.)., 303(5661), 1173–1176, doi:10.1126/science.1092805, 2004.

Crounse, J. D., Mckinney, K. a, Kwan, A. J. and Wennberg, P. O.: Measurement of Gas-Phase Hydroperoxides by Chemical Ionization Mass Spectrometry sponse . Correspondence of airborne observations made illustrated . CF 3 O - appears to be a highly versatile reagent, October, 78(19), 6726–6732, 2006.

van Eijck, A., Opatz, T., Taraborrelli, D., Sander, R. and Hoffmann, T.: New tracer compounds for secondary organic aerosol formation from β-caryophyllene oxidation, Atmos. Environ., 80(3), 122–130, doi:10.1016/j.atmosenv.2013.07.060, 2013.

Franklin, E. B., Yee, L. D., Aumont, B., Weber, R. J., Grigas, P. and Goldstein, A. H.: Ch3MS-RF: a random forest model for chemical characterization and improved quantification of unidentified atmospheric organics detected by chromatography–mass spectrometry techniques, Atmos. Meas. Tech., 15(12), 3779–3803, doi:10.5194/amt-15-3779-2022, 2022.

Franklin, E. B., Yee, L. D., Wernis, R., Isaacman-VanWertz, G., Kreisberg, N., Weber, R., Zhang, H., Palm, B. B., Hu, W., Campuzano-Jost, P., Day, D. A., Manzi, A., Artaxo, P., Souza, R. A. F. De, Jimenez, J. L., Martin, S. T. and Goldstein, A. H.: Chemical Signatures of Seasonally Unique Anthropogenic Influences on Organic Aerosol Composition in the Central Amazon, Environ. Sci. Technol., 57(15), 6263–6272, doi:10.1021/acs.est.2c07260, 2023.

Gagan, S., Sarang, K., Rudzinski, K. J., Liu, R., Szmigielski, R. and Zhang, Y.: Synthetic strategies for oxidation products from biogenic volatile organic compounds in the atmosphere: A review, Atmos. Environ., doi:https://doi.org/10.1016/j.atmosenv.2023.120017, 2023.

Gilardoni, S., Massoli, P., Paglione, M., Giulianelli, L., Carbone, C., Rinaldi, M., Decesari, S., Sandrini, S., Costabile, F., Gobbi, G. P., Pietrogrande, M. C., Visentin, M., Scotto, F., Fuzzi, S. and Facchini, M. C.: Direct observation of aqueous secondary organic aerosol from biomass-burning emissions, Proc. Natl. Acad. Sci., 113(36), 10013–10018, doi:10.1073/pnas.1602212113, 2016.

Goldstein, A. H. and Galbally, I. E.: Known and Unexplored Organic Constituents in the Earth's Atmosphere, Environ. Sci.

Technol., 41(5), 1514–1521, 2007.

Goldstein, A. H., Worton, D. R., Williams, B. J., Hering, S. V., Kreisberg, N. M., Panić, O., Górecki, T., Panic, O. and Gorecki, T.: Thermal desorption comprehensive two-dimensional gas chromatography for in-situ measurements of organic aerosols, J. Chromatogr. A, 1186(1–2), 340–347, doi:10.1016/j.chroma.2007.09.094, 2008.

Hallquist, M., Wenger, J. C., Baltensperger, U., Rudich, Y., Simpson, D., Claeys, M., Dommen, J., Donahue, N. M., George, C., Goldstein, A. H., Hamilton, J. F., Herrmann, H., Hoffmann, T., Iinuma, Y., Jang, M., Jenkin, M. E., Jimenez, J. L., Kiendler-Scharr, A., Maenhaut, W., McFiggans, G., Mentel, T. F., Monod, A., Prévôt, A. S. H., Seinfeld, J. H., Surratt, J. D., Szmigielski, R. and Wildt, J.: The formation, properties and impact of secondary organic aerosol: current and emerging issues, Atmos. Chem. Phys., 9(14), 5155–5236, doi:10.5194/acp-9-5155-2009, 2009.

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

Hamilton, J. F., Lewis, A. C., Reynolds, J. C., Carpenter, L. J. and Lubben, A.: Investigating the composition of organic aerosol resulting from cyclohexene ozonolysis: low molecular weight and heterogeneous reaction products, Atmos. Chem. Phys., 6(12), 4973–4984, doi:10.5194/acp-6-4973-2006, 2006.

Hatch, L. E., Jen, C. N., Kreisberg, N. M., Selimovic, V., Yokelson, R. J., Stamatis, C., York, R. A., Foster, D., Stephens, S. L., Goldstein, A. H. and Barsanti, K. C.: Highly Speciated Measurements of Terpenoids Emitted from Laboratory and Mixed-Conifer Forest Prescribed Fires, Environ. Sci. Technol., 53(16), 9418–9428, doi:10.1021/acs.est.9b02612, 2019.

He, X., Zheng, X., You, Y., Zhang, S., Zhao, B., Wang, X., Huang, G., Chen, T., Cao, Y., He, L., Chang, X., Wang, S. and Wu, Y.: Comprehensive chemical characterization of gaseous I/SVOC emissions from heavy-duty diesel vehicles using two-dimensional gas chromatography time-of-flight mass spectrometry, Environ. Pollut., 305, 119284, doi:10.1016/J.ENVPOL.2022.119284, 2022.

Isaacman-Vanwertz, G., Sueper, D. T., Aikin, K. C., Lerner, B. M., Gilman, J. B., de Gouw, J. A., Worsnop, D. R. and Goldstein, A. H.: Automated single-ion peak fitting as an efficient approach for analyzing complex chromatographic data, J. Chromatogr. A, 1529, 81–92, doi:10.1016/j.chroma.2017.11.005, 2017.

Isaacman, G., Kreisberg, N. M., Yee, L. D., Worton, D. R., Chan, A. W. H., Moss, J. A., Hering, S. V. and Goldstein, A. H.: Online derivatization for hourly measurements of gas- and particle-phase semi-volatile oxygenated organic compounds by thermal desorption aerosol gas chromatography (SV-TAG), Atmos. Meas. Tech., 7(12), 4417–4429, doi:10.5194/amt-7-4417-2014, 2014.

Jaoui, M., Kleindienst, T. E., Docherty, K. S., Lewandowski, M. and Offenberg, J. H.: Secondary organic aerosol formation from the oxidation of a series of sesquiterpenes: alpha-cedrene, beta-caryophyllene, alpha-humulene and alpha-farnesene with O-3, OH and NO3 radicals, Environ. Chem., 10(3), 178–193, doi:10.1071/en13025, 2013.

Jen, C. N., Hatch, L. E., Selimovic, V., Yokelson, R. J., Weber, R., Fernandez, A. E., Kreisberg, N. M., Barsanti, K. C. and Goldstein, A. H.: Speciated and total emission factors of particulate organics from burning western US wildland fuels and their dependence on combustion efficiency, Atmos. Chem. Phys., 19(2), 1013–1026, doi:10.5194/acp-19-1013-2019, 2019.

Jeon, S., Walker, M. J., Sueper, D. T., Day, D. A., Handschy, A. V, Jose, L. and Williams, B. J.: A searchable database and mass spectral comparison tool for aerosol mass spectrometry (AMS) and aerosol chemical speciation monitor (ACSM), , (June), 1–23, 2023.

Johnston, M. V. and Kerecman, D. E.: Molecular Characterization of Atmospheric Organic Aerosol by Mass Spectrometry, Annu. Rev. Anal. Chem., 12, 247–274, doi:10.1146/annurev-anchem-061516-045135, 2019.

Joo, T., Rivera-Rios, J. C., Takeuchi, M., Alvarado, M. J. and Ng, N. L.: Secondary Organic Aerosol Formation from Reaction of 3-Methylfuran with Nitrate Radicals, ACS Earth Sp. Chem., 3(6), 922–934, doi:10.1021/acsearthspacechem.9b00068, 2019.

Joo, T., Machesky, J. E., Zeng, L., Hass-Mitchell, T., Weber, R. J., Gentner, D. R. and Ng, N. L.: Secondary Brown Carbon Formation From Photooxidation of Furans From Biomass Burning, Geophys. Res. Lett., 51(1), doi:10.1029/2023GL104900, 2024.

Kanakidou, M., Seinfeld, J. H., Pandis, S. N., Barnes, I., Dentener, F. J., Facchini, M. C., Van Dingenen, R., Ervens, B., Nenes, A., Nielsen, C. J., Swietlicki, E., Putaud, J. P., Balkanski, Y., Fuzzi, S., Horth, J., Moortgat, G. K., Winterhalter, R., Myhre, C. E. L., Tsigaridis, K., Vignati, E., Stephanou, E. G. and Wilson, J.: Organic aerosol and global climate modelling: a review, Atmos. Chem. Phys., 5(4), 1053–1123, doi:10.5194/acp-5-1053-2005, 2005.

Kenagy, H. S., Heald, C. L., Tahsini, N., Goss, M. B., Kroll, J. H. and Science, C.: Can we achieve atmospheric chemical environments in the laboratory ? An integrated model-measurement approach to chamber SOA studies, , (2), 1–29, doi:10.1126/sciadv.ado1482, 2024.

Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmüller, E., Dörmann, P., Weckwerth, W., Gibon, Y., Stitt, M., Willmitzer, L., Fernie, A. R. and Steinhauser, D.: GMD@CSB.DB: The Golm metabolome database, Bioinformatics, 21(8), 1635–1638, doi:10.1093/bioinformatics/bti236, 2005.

Kourtchev, I., Fuller, S. J., Giorio, C., Healy, R. M., Wilson, E., O'Connor, I., Wenger, J. C., McLeod, M., Aalto, J., Ruuskanen, T. M., Maenhaut, W., Jones, R., Venables, D. S., Sodeau, J. R., Kulmala, M. and Kalberer, M.: Molecular

composition of biogenic secondary organic aerosols using ultrahigh-resolution mass spectrometry: comparing laboratory and field studies, Atmos. Chem. Phys., 14(4), 2155–2167, doi:10.5194/acp-14-2155-2014, 2014.

Kováts, E.: Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone, Helv. Chim. Acta, 41(7), 1915–1932, doi:10.1002/hlca.19580410703, 1958.

Krechmer, J. E., Groessl, M., Zhang, X., Junninen, H., Massoli, P., Lambe, A. T., Kimmel, J. R., Cubison, M. J., Graf, S., Lin, Y.-H., Budisulistiorini, S. H., Zhang, H., Surratt, J. D., Knochenmuss, R., Jayne, J. T., Worsnop, D. R., Jimenez, J. L. and Canagaratna, M. R.: Ion mobility spectrometry–mass spectrometry (IMS–MS) for on- and offline analysis of atmospheric gas and aerosol species, Atmos. Meas. Tech., 9(7), 3245–3262, doi:10.5194/amt-9-3245-2016, 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, Nat. Chem., 3(2), 133–139, doi:10.1038/nchem.948, 2011.

Lee, B. H., Lopez-Hilfiker, F. D., Mohr, C., Kurtén, T., Worsnop, D. R. and Thornton, J. A.: An iodide-adduct high-resolution time-of-flight chemical-ionization mass spectrometer: Application to atmospheric inorganic and organic compounds, Environ. Sci. Technol., 48(11), doi:10.1021/es500362a, 2014.

Liang, Y., Jen, C. N., Weber, R. J., Misztal, P. K. and Goldstein, A. H.: Chemical composition of PM2.5 in October 2017 Northern California wildfire plumes, Atmos. Chem. Phys., 21(7), 5719–5737, doi:10.5194/acp-21-5719-2021, 2021.

Lin, Y.-H., Zhang, Z., Docherty, K. S., Zhang, H., Budisulistiorini, S. H., Rubitschun, C. L., Shaw, S. L., Knipping, E. M., Edgerton, E. S., Kleindienst, T. E., Gold, A. and Surratt, J. D.: Isoprene Epoxydiols as Precursors to Secondary Organic Aerosol Formation: Acid-Catalyzed Reactive Uptake Studies with Authentic Compounds, Environ. Sci. Technol., 46(1), 250–258, doi:10.1021/es202554c, 2012.

Lopez-Hilfiker, F. D., Pospisilova, V., Huang, W., Kalberer, M., Mohr, C., Stefenelli, G., Thornton, J. A., Baltensperger, U., Prevot, A. S. H. and Slowik, J. G.: An extractive electrospray ionization time-of-flight mass spectrometer (EESI-TOF) for online measurement of atmospheric aerosol particles, Atmos. Meas. Tech., 12(9), doi:10.5194/amt-12-4867-2019, 2019.

Mallard, W. G., Andriamaharavo, N. R., Mirokhin, Y. A., Halket, J. M. and Stein, S. E.: Creation of libraries of recurring mass spectra from large data sets assisted by a dual-column workflow, Anal. Chem., 86(20), 10231–10238, doi:10.1021/ac502379x, 2014.

McGlynn, D. F., Yee, L. D., Garraffo, H. M., Geer, L. Y., Mak, T. D., Mirokhin, Y. A., Tchekhovskoi, D. V, Jen, C. N., Goldstein, A. H., Kearsley, A. J. and Stein, S. E.: New Library-Based Methods for Nontargeted Compound Identification by

GC-EI-MS, J. Am. Soc. Mass Spectrom., 36(2), 389–399, doi:10.1021/jasms.4c00451, 2025.

Nizkorodov, S. A., Laskin, J. and Laskin, A.: Molecular chemistry of organic aerosols through the application of high resolution mass spectrometry, Phys. Chem. Chem. Phys., 13(9), 3612–3629, doi:10.1039/c0cp02032j, 2011.

Nozière, B., Kalberer, M., Claeys, M., Allan, J., D'Anna, B., Decesari, S., Finessi, E., Glasius, M., Grgic, 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(10), 3919–3983, doi:10.1021/cr5003485, 2015.

Offenberg, J. H., Lewandowski, M., Kleindienst, T. E., Docherty, K. S., Jaoui, M., Krug, J., Riedel, T. P. and Olson, D. A.: Predicting Thermal Behavior of Secondary Organic Aerosols, Environ. Sci. Technol., 51, 9911–9919, doi:10.1021/acs.est.7b01968, 2017.

Porter, W. C., Jimenez, J. L. and Barsanti, K. C.: Quantifying Atmospheric Parameter Ranges for Ambient Secondary Organic Aerosol Formation, ACS Earth Sp. Chem., 5(9), 2380–2397, doi:10.1021/acsearthspacechem.1c00090, 2021.

Roach, P. J., Laskin, J. and Laskin, A.: Molecular characterization of organic aerosols using nanospray-desorption/ electrospray ionization-mass spectrometry, Anal. Chem., 82(19), 7979–7986, doi:10.1021/ac101449p, 2010a.

Roach, P. J., Laskin, J. and Laskin, A.: Nanospray desorption electrospray ionization: An ambient method for liquid-extraction surface sampling in mass spectrometry, Analyst, 135(9), 2233–2236, doi:10.1039/c0an00312c, 2010b.

Robinson, N. H., Hamilton, J. F., Allan, J. D., Langford, B., Oram, D. E., Chen, Q., Docherty, K. S., Farmer, D. K., Jimenez, J. L., Ward, M. W., Hewitt, C. N., Barley, M. H., Jenkin, M. E., Rickard, A. R., Martin, S. T., Mcfiggans, G. and Coe, H.: Evidence for a significant proportion of Secondary Organic Aerosol from isoprene above a maritime tropical forest, Atmos. Chem. Phys., 11, 1039–1050, doi:10.5194/acp-11-1039-2011, 2011.

Song, K., Guo, S., Gong, Y., Lv, D., Zhang, Y., Wan, Z., Li, T., Zhu, W., Wang, H., Yu, Y., Tan, R., Shen, R., Lu, S., Li, S., Chen, Y. and Hu, M.: Impact of cooking style and oil on semi-volatile and intermediate volatility organic compound emissions from Chinese domestic cooking, Atmos. Chem. Phys., 22(15), 9827–9841, doi:10.5194/acp-22-9827-2022, 2022.

Stein, S. E.: Estimating probabilities of correct identification from results of mass spectral library searches, J. Am. Soc. Mass Spectrom., 5(4), 316–323, doi:10.1016/1044-0305(94)85022-4, 1994.

Stein, S. E.: Chemical Substructure Identification by Mass Spectral Library Searching, J. Am. Soc. Mass Spectrom., 6, 644–

655, 1995.

Stein, S. E.: Mass Spectral Reference Libraries: An Ever-Expanding Resource for Chemical Identification, Anal. Chem., 84, 7274–7282, doi:10.1021/ac301205z, 2012.

Sun, Y. L., Zhang, Q., Anastasio, C. and Sun, J.: Insights into secondary organic aerosol formed via aqueous-phase reactions of phenolic compounds based on high resolution mass spectrometry, Atmos. Chem. Phys., 10(10), 4809–4822, doi:10.5194/acp-10-4809-2010, 2010.

Surratt, J. D., Lewandowski, M., Offenberg, J. H., Jaoui, M., Kleindienst, T. E., Edney, E. O. and Seinfeld, J. H.: Effect of acidity on secondary organic aerosol formation from isoprene, Environ. Sci. Technol., 41(15), 5363–5369, doi:10.1021/es0704176, 2007.

Szmigielski, R., Surratt, J. D., Gómez-González, Y., Van der Veken, P., Kourtchev, I., Vermeylen, R., Blockhuys, F., Jaoui, M., Kleindienst, T. E., Lewandowski, M., Offenberg, J. H., Edney, E. O., Seinfeld, J. H., Maenhaut, W. and Claeys, M.: 3-methyl-1,2,3-butanetricarboxylic acid: An atmospheric tracer for terpene secondary organic aerosol, Geophys. Res. Lett., 34(24), L24811, doi:10.1029/2007GL031338, 2007.

Takeuchi, M., Berkemeier, T., Eris, G. and Ng, N. L.: Non-linear effects of secondary organic aerosol formation and properties in multi-precursor systems, Nat. Commun., 13(1), 1–13, doi:10.1038/s41467-022-35546-1, 2022.

Ulbrich, I. M., Canagaratna, M. R., Zhang, Q., Worsnop, D. R. and Jimenez, J. L.: Interpretation of organic components from Positive Matrix Factorization of aerosol mass spectrometric data, Atmos. Chem. Phys., 9(9), 2891–2918, doi:10.5194/acp-9-2891-2009, 2009.

Walters, W. W., Takeuchi, M., Blum, D. E., Eris, G., Tanner, D., Xu, W., Rivera-Rios, J., Liu, F., Xu, T., Huey, G., Min, J. B., Weber, R., Ng, N. L. and Hastings, M. G.: Evaluating NOx fate and organic nitrate chemistry from α-pinene oxidation using stable oxygen and nitrogen isotopes, Atmos. Chem. Phys., 25(18), 10707–10730, doi:10.5194/acp-25-10707-2025, 2025.

Wang, X., Hayeck, N., Brüggemann, M., Yao, L., Chen, H., Zhang, C., Emmelin, C., Chen, J., George, C. and Wang, L.: Chemical Characteristics of Organic Aerosols in Shanghai: A Study by Ultrahigh-Performance Liquid Chromatography Coupled With Orbitrap Mass Spectrometry, J. Geophys. Res. Atmos., 122(21), 11,703-11,722, doi:10.1002/2017JD026930, 2017.

Wen, Z., Xu, L. and Zhang, H.: Recent advances in mass spectrometry techniques for atmospheric chemistry research on molecular-level, Mass Spectrom. Rev., 43(5), 1091–1134, doi:10.1002/mas.21857, 2023.

Worton, D. R., Kreisberg, N. M., Isaacman, G., Teng, A. P., McNeish, C., Górecki, T., Hering, S. V. and Goldstein, A. H.: Thermal Desorption Comprehensive Two-Dimensional Gas Chromatography: An Improved Instrument for In-Situ Speciated Measurements of Organic Aerosols, Aerosol Sci. Technol., 46(4), 380–393, doi:10.1080/02786826.2011.634452, 2012.

Worton, D. R., Decker, M., Isaacman-Vanwertz, G., Chan, A. W. H., Wilson, K. R. and Goldstein, A. H.: Improved molecular level identification of organic compounds using comprehensive two- dimensional chromatography, dual ionization energies and high resolution mass spectrometry, Analyst, 142, 2395–2403, doi:10.1039/c7an00625j, 2017.

Yee, L. D., Isaacman-Vanwertz, G., Wernis, R. A., Meng, M., Rivera, V., Kreisberg, N. M., Hering, S. V., Bering, M. S., Glasius, M., Upshur, M. A., Gray Bé, A., Thomson, R. J., Geiger, F. M., Offenberg, J. H., Lewandowski, M., Kourtchev, I., Kalberer, M., de Sá, S. S., Martin, S. T., Alexander, M. L., Palm, B. B., Hu, W., Campuzano-Jost, P., Day, D. A., Jimenez, J. L., Liu, Y. J., Mckinney, K. A., Artaxo, P., Viegas, J., Manzi, A., Oliveira, M. B., de Souza, R. A. F., Machado, L. A. T., Longo, K. and Goldstein, A. H.: Observations of sesquiterpenes and their oxidation products in central Amazonia during the wet and dry seasons, Atmos. Chem. Phys, 18, 10433–10457, doi:10.5194/acp-18-10433-2018, 2018.

Zhang, H., Yee, L. D., Lee, B. H., Curtis, M. P. M. P., Worton, D. R., Isaacman-VanWertz, G., Offenberg, J. H., Lewandowski, M., Kleindienst, T. E. T. E., Beaver, M. R., Holder, A. L., Lonneman, W. A., Docherty, K. S., Jaoui, M., Pye, H. O. T., Hu, W., Day, D. A. D. A., Campuzano-Jost, P., Jimenez, J. L., Guo, H., Weber, R. J., de Gouw, J. A., Koss, A. R., Edgerton, E. S., Brune, W., Mohr, C., Lopez-Hilfiker, F. D., Lutz, A., Kreisberg, N. M., Spielman, S. R., Hering, S. V., Wilson, K. R., Thornton, J. A. and Goldstein, A. H.: Monoterpenes are the largest source of summertime organic aerosol in the southeastern United States, Proc. Natl. Acad. Sci., 115(9), 2038–2043, doi:10.1073/pnas.1717513115, 2018.

Zhang, M., Cai, D., Lin, J., Liu, Z., Li, M., Wang, Y. and Chen, J.: Molecular characterization of atmospheric organic aerosols in typical megacities in China, npj Clim. Atmos. Sci., 7(1), doi:10.1038/s41612-024-00784-1, 2024.