Elucidating atmospheric brown carbon – Supplanting chemical intuition with exhaustive enumeration and machine learning

Enrico Tapavicza,*,† Guido Falk von Rudorff,‡ David O. De Haan,¶ Mario Contin,§ Christian George,‖ Matthieu Riva,‖ and O. Anatole von Lilienfeld‡

†*Department of Chemistry and Biochemistry, California State University, Long Beach, 1250 Bellflower Boulevard, Long Beach, CA, 90840, USA*

‡*University of Vienna, Faculty of Physics, Kolingasse 14-16, AT-1090 Wien, Austria, and Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials (MARVEL), Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland*

¶*Department of Chemistry and Biochemistry, University of San Diego, 5998 Alcala Park, San Diego, CA, 92110, USA*

§*Universidad de Buenos Aires, Facultad de Farmacia y Bioquímica, Departamento de Química Analitica y Fisicoquímica, Junín 956, Buenos Aires, C1113AAD, Argentina*

‖*Université Lyon, Université Claude Bernard Lyon 1, CNRS, IRCELYON, 69626 Villeurbanne, France*

E-mail: enrico.tapavicza@csulb.edu

Abstract

Brown carbon (BrC) is involved in atmospheric light absorption, climate forcing, and can cause adverse health effects. Understanding the formation mechanisms and

molecular structure of BrC is of key importance in developing strategies to control its impact on environment and health. Structure determination of BrC, however, is challenging, mainly due to the lack in experiments providing characteristic fingerprints of the molecules and the sheer size of possible molecular structures with identical molecular mass. Chemical intuition and *ad hoc* assumptions often provide the basis for suggesting particular structures, but are prone to errors due to their biased nature. This study develops an unbiased algorithm, based on a combined graph-based molecule generator and machine learning workflow, to identify the molecular structure of compounds involved in biomass burning (BB) and the composition of BrC. We apply this algorithm to unravel the structures of $C_{12}H_{12}O_7$ isomers, identified in chamber experiments as light-absorbing photooxidation products of syringol, a prevalent marker in wood smoke. Of the 260 million initial molecular graphs with sum formula $C_{12}H_{12}O_7$, the algorithm reduces the number of candidates to 0.01%. Further reduction strategies are discussed and analyzed according to their power to condense the number structures consistent with experimental observations. The method will potentially make isomers extracted from lab and field aerosol particles more readily and rapidly identified without introducing human bias.

Introduction

Visible light-absorbing secondary organic aerosols (SOAs), also known as brown carbon (BrC), interfere in atmospheric processes, impact climate forcing, and cause adverse health effects due to their oxidative character.[1–5] Emerging from biomass burning (BB) and from natural and industrial emissions, SOAs constantly undergo several chemical modifications due to reactions in the atmosphere, eventually forming light-absorbing oligomers with large absorption coefficients. Understanding the molecular details of the formation mechanisms, precursor identification, and knowledge of the exact molecular structure of BrC is of key importance in designing strategies and policies to control its impact on environment and

public health. Structural knowledge is not only important to evaluate their toxicology,[6,7] carcinogenic activity, and receptors binding[8] in order to assess public health impact, but also does it allow to make predictions on molecular stability and chemical fate; the latter are important properties to assess BrC's further implications on atmospheric processes and climate forcing.[9] Thus, structure and precursor identification takes up a key role in atmospheric chemistry, connecting field studies with lab experiments and computer modeling. Unravelling the structure of BrC compounds and the characterization of the molecular composition, particularly identifying the major constitutional isomers, is a pressing challenge for atmospheric chemistry.[10–12] The difficulty of this process is mainly caused by the lack of experiments providing characteristic fingerprints of these molecules and due to the sheer size of possible molecular structures associated with a given molecular mass. Chemical intuition and *ad hoc* assumptions for structural elements often provide the basis for suggesting particular structures, but are prone to errors due to their biased nature. This study develops an unbiased algorithm to identify the molecular structure of compounds involved in BB and the composition of BrC. One source of secondary brown carbon believed to be atmospherically significant is the formation of oligomers during the aqueous-phase photooxidation of phenolic compounds, such as syringol, that are prevalent in wood smoke. This oligomer formation is thought to change the optical properties of BB aerosol particles during cloud processing, partially counteracting photobleaching and other aging processes. We focus here on the structural identification of a light-absorbing dimer identified from this reaction with the formula $C_{12}H_{12}O_7$.

Typically, the formation and further reactions of SOA under specific atmospheric conditions can be simulated in atmospheric chamber experiments.[13] In these experiments, aerosol-phase reaction products are often extracted from filters and characterized by high-resolution liquid chromatography (LC)/mass spectrometry (MS) analysis with inline UV/Vis absorbance spectroscopy. While the detected mass of the compounds gives complete information about the chemical sum formula, the exact chemical structure remains unknown. In

chamber experiments, often molecular structures are proposed on the basis of mass spectrometric fragment data, absorption estimates, and chemical intuition. Chemical intuition, however, introduces a human bias that might prevent the discovery of the exact molecular constitution by not considering specific isomers if they do not seem probable based on the chemist's experience. This bias might constitute a hurdle in finding novel, undiscovered constitutional isomers. In view of the large number of constitutional isomers of medium sized molecules, the probability of picking the right structure from the first estimate is small. Moreover, chemical intuition requires manual intervention, limiting the number of potential target compounds that can be identified.

Thus, it is advisable to support proposed structures by comparison of as many physically measurable properties as available to gain confidence in the correctness of the chosen structure or rule out structures not consistent with experimental observations. Due to the low concentrations of compounds in gas phase experiments, spectroscopic methods that provide conclusive information about the chemical constitution, such as NMR, are unfortunately not applicable. The high absorption coefficients of BrC, however, allow the use of UV/Vis spectroscopy to probe if the proposed structure is consistent with the experimental observations. Furthermore, it is questionable if only one constitutional isomer is present. Given the chemical complexity of SOA, it is often more realistic to assume a composition of different isomers with similar physical chemical properties.

In an attempt to rule out any human bias in the proposition of candidate structures, we attack the problem of finding consistent isomers by initially considering *all* possible isomers exhaustively; this is in stark contrast to the common approaches based on chemical intuition.[11,14] However, due to the quickly growing size of the chemical space with the number of atoms, this approach is already challenging for small and medium-sized molecules and becomes impossible for larger molecules. As a test case, we consider syringol (2,6-dimethoxyphenol), an aromatic phenolic C_8 compound (Fig. 1) that has been used as a marker for wood smoke emissions in the atmosphere.[15] When syringol is photooxidized with

OH radicals or triplet carbon ($^3C^*$) species in the aqueous phase, one of the seven major products detected by negative-mode nano-desorption electrospray mass spectrometry has the sum formula $C_{12}H_{12}O_7$.[14] In experiments at the CESAM chamber[16] where syringol was oxidized with OH radicals, product peaks identified by UHPLC-(+)ESI-MS in aerosol extracts were categorized by whether their concentrations were higher in experiments where brown carbon formed. Within this group of peaks that correlated with brown carbon, 2% of the peak area was due to C_6 (monomer) products, 94% was due to C_{10} - C_{15} (dimer) products, and 4% to larger products, up to C_{29}. Among molecules with less than 20 heavy atoms, $C_{12}H_{12}O_7$ was the largest peak, responsible for 7% of the peak area correlating with brown carbon, and co-eluting with an absorbance peak. Thus, $C_{12}H_{12}O_7$ is an appropriate brown carbon candidate (Fig. 1). Considering all possible molecular graphs (i.e. the set of all atoms and their bonds including bond orders) of $C_{12}H_{12}O_7$, we assign one graph node for each heavy atom. In total there are more than 10^{35} simple connected graphs of 19 nodes.[17] The number of molecular graphs is even higher, since this count neither includes elemental composition nor bond orders. These large numbers show that the major bottleneck in exploring this chemical subspace lies in the efficiency of the computer generation of molecular structures, which ultimately limits this approach for molecules larger than a given size. A second challenge arises from the prediction of physical chemical data for all these structures needed to determine the candidate molecules consistent with experimental measurements. In BrC, the observable usually is the UV/Vis spectrum, which can be predicted reasonably well by correlated quantum chemistry methods.[18–22] However, the computational resources necessary for the spectra prediction of such a large number of isomers quickly becomes out of reach.

Here, we developed a computational workflow to find possible constitutional $C_{12}H_{12}O_7$ isomers consistent with the recorded absorption spectra. To tackle the exhaustive generation of constitutional isomers we present a graph-based, bias-free molecule generator, that leverages massively parallel computation. The problem of quantum chemical spectra predic-

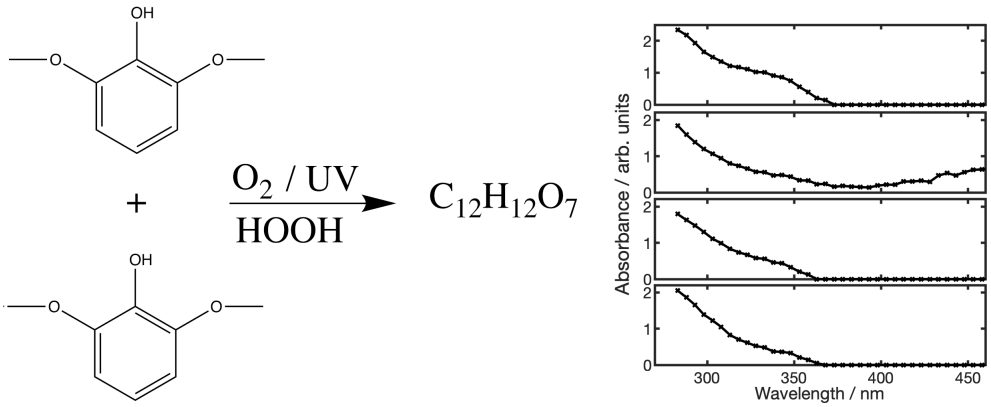

Figure 1: In the proposed reaction scheme (left), aqueous-phase syringol photooxidation forms $C_{12}H_{12}O_7$, a product with unknown structure that correlates with brown carbon formation. Right: four different UV/Vis absorbance spectra, measured in a filter extract from the CESAM chamber at four different retention times; each corresponds with elution of a different $C_{12}H_{12}O_7$ isomer.

tion of a large number of molecules is solved by making use of machine learning to predict spectral properties of the molecules.[23] In a Monte-Carlo procedure, we then determine the likelihood that specific feature groups give rise to the experimentally observed spectrum.

The work flow starts from an unbiased and exhaustive generation of all possible molecular graphs. The number of graphs is further reduced by molecular stability and steric criteria based on tight-binding density functional theory. After prediction of electronic excitation energies and oscillator strengths, we filter the compounds by the probability of agreement with experimental UV/Vis absorption spectrum. Finally, we explore how additional information about structure or functional groups could further reduce the number of possible $C_{12}H_{12}O_7$ isomers consistent with experimental data.

Materials and Methods

Experimental

The filter extraction protocol has been described previously.[24] Briefly, each collected Teflon filter (1 μm pores, 47 mm diam.) was spiked with caffeine (final concentration 100 ppb), as

internal standard and then extracted twice with 6 mL of acetonitrile and agitated for 20 minutes with an orbital shaker at 1000 rpm. The extracts were then filtered with a syringe filter (0.2 μm, Pall Acrodisc® PSF, with GHP membrane, hydrophilic polypropylene) to remove any insoluble particles and blown dry under a gentle N_2 (g) stream at ambient temperature. The residues were reconstituted in 0.2 mL of water:methanol (v/v 1:1, Optima®LC/MS, Fischer Scientific). Finally, the filter extracts were analyzed by ultra-high performance liquid chromatography (Dionex 3000, Thermo Scientific) using a Water Acquity HSS C18 column (1.8 μm, 100 x 2.1mm) coupled with a diode array UV/Vis absorbance detector and a Q-Exactive Hybrid Quadrupole-Orbitrap mass spectrometer (Thermo Scientific) equipped with an electrospray ionization (ESI) source operated in positive or negative mode. The mobile phase used was constituted of (A) 0.1% formic acid in water (Optima® LC/MS, Fischer Scientific) and (B) 0.1% formic acid in acetonitrile (Optima® LC/MS, Fischer Scientific). Gradient elution was carried out by the A/B mixture at a total flow rate of 300 μL/min: 0 to 13 min B from 1% to 100% , 13.1 min B 1% for 9 min.

Raw data was processed with MZmine 2.51. Features with a higher intensity than 1×10^6 and at least 10 times higher than the blank intensity were selected. The chromatographic peaks of all ID selected were visually analyzed and a proposed molecular formula was obtained using Xcalibur 2.2 (Thermo Scientific) software package. A subset of peaks was then identified that had areas averaging at least 5 times larger in experiments where brown carbon formed than when it did not; $C_{12}H_{12}O_7$ was prominent within this subset.

Molecule generation

For the sum formula $C_{12}H_{12}O_7$, we systematically[25] enumerate all molecules that potentially could be a product of the reaction in the atmospheric chamber. We rationalize that the product forms via radical-initiated coupling[26] of two syringol units to $C_{16}H_{18}O_6$, the most abundant SOA product identified in previous studies,[14,27] followed by further oxidation and fragmentation to $C_{12}H_{12}O_7$.[14] We limited ourselves to those candidates where the two C_6-

rings found in the two reactant molecules persist in the product, which requires the loss of methoxy carbons. We note that half of the syringol SOA product structures proposed by Yu et al.[14] have lost at least one methoxy carbon, and 16% of their proposed structures have lost all methoxy carbons. Demethylation of methoxy groups during photooxidation has been observed for vanillin,[28] syringaldeyde, and acetosyringol.[29] Technically, this enumeration is performed by a) enumerating all potential molecular graphs ignoring hydrogens, b) constructing all possible hydrogen saturations of these graphs, c) filtering all molecules which are not stable in GFN2-xTB calculations.[30] The protocol for these steps is detailed in the SI and is based on Refs.[25,31–34]

Electronic structure methods and machine learning

To assess the absorption spectrum, we computed the lowest three excitation energies and their corresponding oscillator strengths using the Algebraic Diagramatic Construction to Second Order (ADC(2)) method.[35,36] To include effects of water solvation in the calculation, we employed the Conductor-like Screening Model (COSMO)[37] using a dielectric constant of 80.1 and a refractive index of 1.3325.[38] The def2-TZVP basis set was used.[39] This approach has been shown to yield accurate excitation energies.[40] Calculations were carried out with TURBOMOLE V7.2.[41,42]

Since it is prohibitively expensive to apply this reliable method to the exhaustive list of all molecules, we calculated 10,000 randomly selected molecules as training set for the Kernel-Ridge-Regression (KRR) method[43] with the FCHL molecular representation[44] as implemented in the QML toolkit.[45] Machine learning in general and KRR in particular have been successfully used to predict excited state properties,[46–48] typically highlighting the need for high-quality reference data. 82 molecules were excluded since they exhibited negative excitation energies, which indicates a non-stable ground state. We determined optimal hyperparameters for the kernel widths and regularizer with 5-fold cross validation (see SI). Once both the excitation energies and oscillator strengths for the lowest three

excitations have been predicted for all compounds from machine learning, we can model the spectrum[3,49] and compare it to the experimental ones. We employ a Monto-Carlo method (see SI) to assess whether these predicted spectra are compatible with the experimental spectra. In this work, a predicted spectrum is considered compatible if the experimental spectrum and predicted spectrum are separated by at most one standard deviation of both modeling and experimental uncertainties.

Results and discussion

Analysis of the generated molecules

At first we will analyze the distribution of features in the molecular graphs, before the structures have been optimized. According to our initial assumption that two C_6-rings exist in the structure, there are two different possibilities how the rings are connected: either directly by a carbon-carbon bond, or by one or more oxygen atoms, serving as a bridging unit. These two possibilities are reflected by having either 13 or 12 C-C bonds, respectively (Fig. 2). Analyzing C-O bonds, we find a peaked distribution ranging from 1 to 13, with a maximum probability at 7. Oxygen-oxygen bonds range from 0 to a maximum of 6, with the maximum of 6 corresponding to a structure where an O_7 chain exists (blue, middle Fig. 2). We also note that the longer the oxygen chain, the fewer graphs are found, as expected. We note that most structures have 0-2 carbonyl groups (Fig. 2, right), which are important for absorption properties.

Of the 263 million graphs about 123 million lead to stable three-dimensional structures according to GFN2-xTB. All their coordinates are available online[50] together with the reference data for the machine learning model.[51] Since we are mainly interested in the structures that are consistent with the experimental spectra, we skip a more detailed analysis of the features of this large structure set. However, it is important to say that we observe a substantial amount of structures that are not commonly seen. For instance, we find a considerable

number of stable molecules with chains up to seven oxygen atoms and dioxiranes (i.e. three rings with two oxygen atoms). There has been an ongoing discussion about the possible length of oxygen chains.[52] While it might seem unlikely to find oxygen chains with more than three members, theory has predicted the stability of oxygen chains up to at least 6 members. Experimentally, four-membered chains have been confirmed.[52] Dioxiranes have been known experimentally since 1978, although their existence were already predicted in 1899 by Bayer and Villiger.[53] An indication of the relative stability of the molecules can also be based on total electronic energies (see Fig. 1, SI).

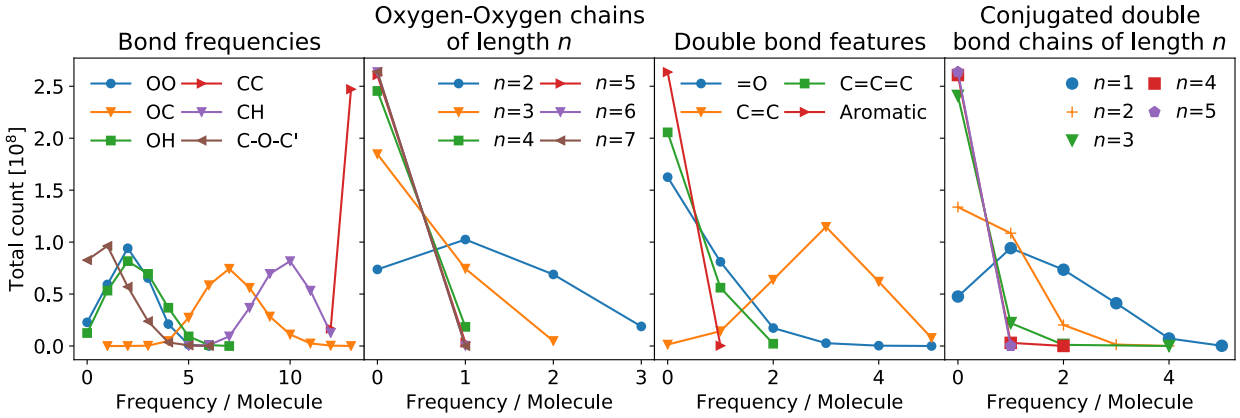

Figure 2: Total number of molecular graphs with given feature occurrences. Panel 1: bond count frequencies, and ether bridges connecting the two carbon rings; panel 2: count of oxygen-oxygen chains of length n; panel 3: count of carbonyl sites, allene sites, double bonds, aromatic rings; and panel 4: conjugated double bond chains of length n. Each curve adds up to the total number of molecular graphs of 263'917'411.

Electronic spectra prediction

For a random set of 9,918 molecules ADC(2)/COSMO calculations resulted in positive excitation energies and oscillator strengths for the lowest three states (Fig. 3). The data is available online.[51] From Fig. 3, we see that the first excited state contributes with the highest oscillator strengths in the region between 0.12 and 0.16 au, where the experimental absorption band is located, but S_2 and S_3 also show substantial absorption in this region. Using KKR, we predicted the lowest three excitation energies and oscillator strengths based on

different training set sizes (Fig. 4). For a training set of 9000 molecules, predictions exhibit mean absolute errors (MAEs) of 9, 8, and 7 mHa, for S_1, S_2, and S_3, respectively. Thus, ML errors are similar to the expected error of ADC(2) with respect to experimental values, which was previously determined to be 8 mHa (0.21 eV).[54] The curves confirm the learning abilities of the model as it makes use of additional training data to improve prediction accuracy. The accuracy of the machine learning predictions is set into perspective by comparison to the null model (dashed lines in Fig. 4), which is obtained when the mean excitation energy over all training molecules is used as prediction. MAEs for the oscillator strengths amount to 0.035, 0.038, and 0.036 au, for S_1, S_2, and S_3, respectively (Fig. 4, right). Interestingly, oscillator strengths of S_1 benefit the most from KKR, whereas, for S_2 and S_3, learning curves are comparably flat. Using the ML model based on the 9,918 training molecules, we predicted the lowest three excitation energies and oscillator strengths for the remaining 120 million stable structures.

Establishing matching characteristics

We present the analysis for the first spectrum on the top right of Fig. 1; results for the remaining three spectra are very similar (see Fig. 2, SI). Out of the 123 million stable molecules, 55 million match this spectrum according to the criteria defined in the SI. For every structure, we determined a feature vector that describes the structural features in the molecules (Fig. 5). Features considered were: a) bond types, b) oxygen chains of different lengths, c) carbonylic groups, d) double bonds, e) conjugated double bonds, f) aromatic rings, g) ether bridges, and h) allene groups. The total dimension of the feature vector amounts to 21, whereas the length of the entries vary between 2 and 6. For instance for the carbon-carbon bonds, only two values are possible (12 and 13), but for two-membered oxygen chains the number of entries amounts to four, because possible values are 0, 1, 2, 3. For every feature and for every number value thereof, we calculate the fraction of the molecules that are compatible with the experimental spectrum as defined above. This allows to correlate

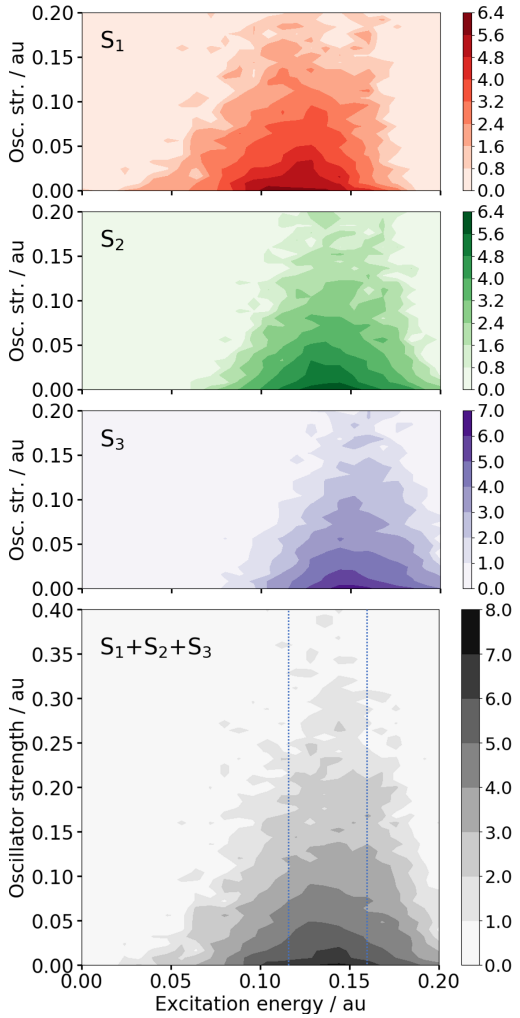

Figure 3: Distribution of ADC(2)/COSMO excited states as a function of excitation energy and oscillator strengths of the 9,918 training molecules. The distribution is shown separately for the lowest three excited states (top three panels) and combined for all three states (bottom panel). The color code refers to the decadic logarithm of the density found in a square of an area of 0.008×0.008 au^2. The blue dotted lines in the bottom panel indicate the region in which the experimental band is located.

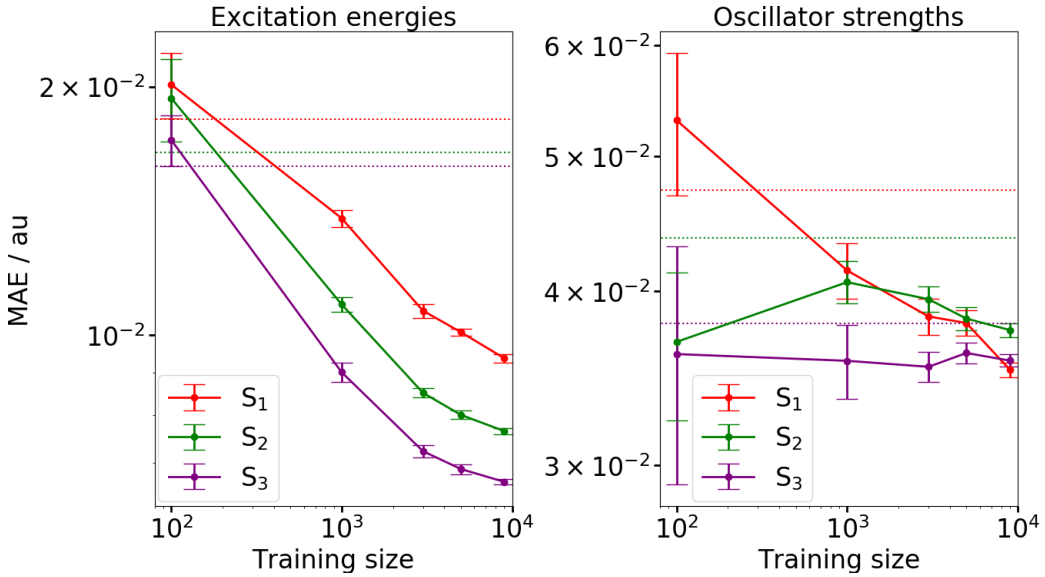

Figure 4: Left: Mean absolute error (MAE) as a function of training set size for the learning of the lowest three excitation energies. Right: Mean absolute error (MAE) as a function of training size for the learning of the lowest three oscillator strengths. In both plots, the dashed lines indicate the error of the null model, using the same color code for the different states.

molecular features with the probability that it causes the experimentally observed spectrum (see Fig. 1). Analyzing the features in Fig. 5, we find that for example the probability that a molecule is consistent with experimental spectrum increases with the number of (O-O) bonds (blue line, left panel). As another example (orange line, right panel), we see that the probability of a matching molecule decreases if there are more than two cases of conjugated double bond chains of length two.

To illustrate the chemical diversity of stable molecular structures that are compatible with the experimentally observed spectrum, we group all molecules by their feature vector. Representatives of large feature groups of matching and non-matching molecules are given on top and bottom of Fig. 5, respectively. The corresponding groups of molecules are huge: just for the first molecule on the top left in Fig. 5, there are 695,039 stable molecules that match the experimental spectrum and have an identical feature vector.

Each of the other molecules shown in Fig. 5 is just one representative of similarly large groups of feature-identical stable molecules. While the presence of individual molecular

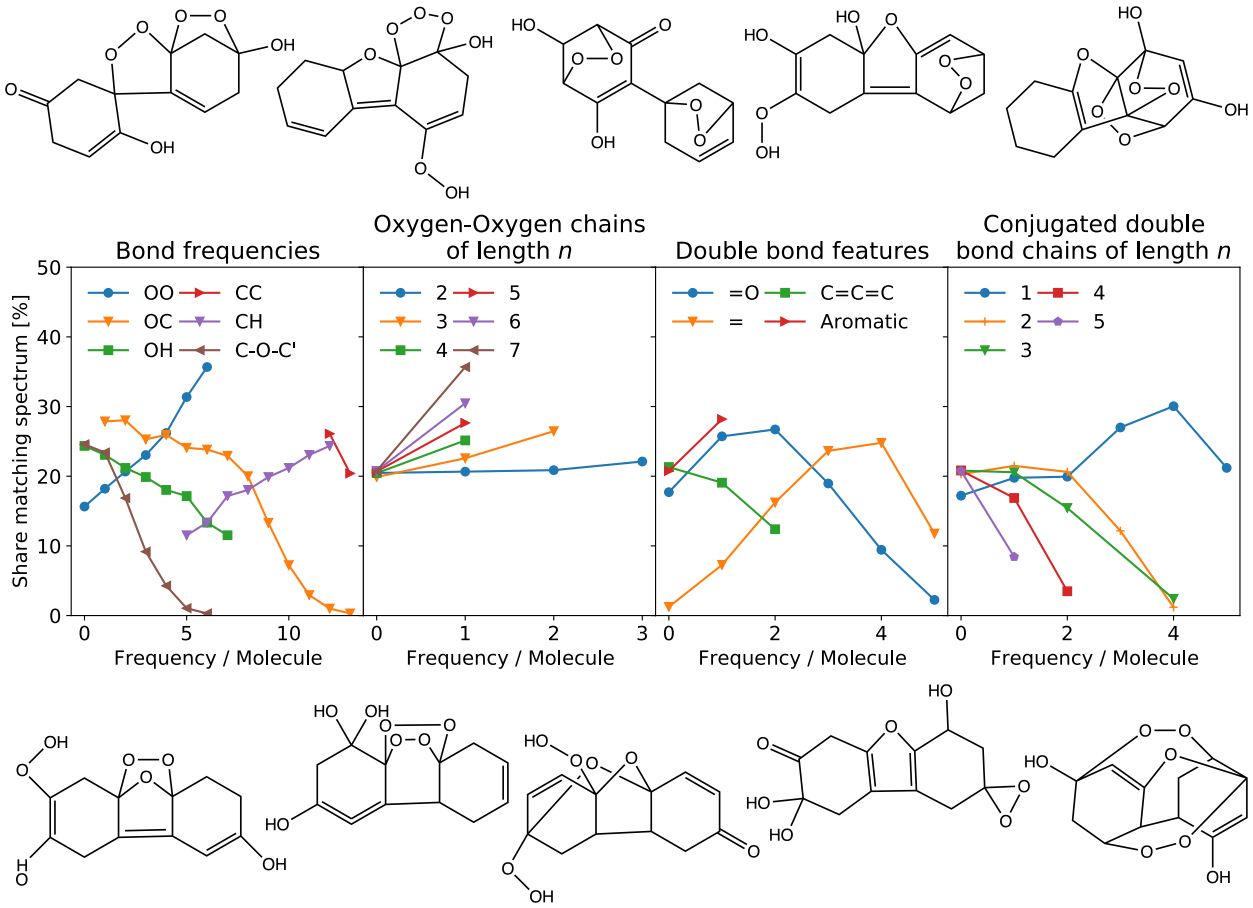

Figure 5: Per-feature probability (Share matching spectrum) of molecular structures being compatible with the first spectrum in Figure 1 shown in the panels. Conditional probabilities are exemplified by grouping all molecules by their feature vector and calculating the share of matching molecules for each group. Low-energy representatives of the largest groups are shown for matching (top) and non-matching (bottom) molecules.

features can significantly reduce the number of molecules, the sheer size of these groups highlights that the share of molecules matching any spectrum is still by far too large to claim unique identification. Thus, the extent of the structural ambiguity of brown carbon absorption spectra is made clear by the exhaustive enumeration of all possible molecular structures.

Filtering strategies

In view of these large numbers of candidate structures, it is evident that any identification of individual molecules based on their spectra needs more criteria derived from experiments to reduce the number of possible candidates. In practice, misidentifications are likely if too few additional constraints are included in the search. Furthermore, a comparison between the representative matching and non-matching feature groups (see Fig. 5) shows that it is not trivial to establish obvious structural characteristics that would increase the likelihood of being consistent with the experimental spectrum. Hence, common textbook relationships between structural elements and absorption properties (e.g. batochromic shift) are of limited utility in the selection of candidate brown carbon molecules.

Table 1: Summary of how given structural features reduce the number of possible $C_{12}H_{12}O_7$ structures.

Total molecules with two C_6 rings	263,917,411
and which have OH groups	263,917,411
and which have no oxygen chain longer than 2	161,160,394
and which have an oxygen connecting the carbon rings	115,715,458
and which have one aromatic ring	134,944
and which are stable	64,121
and which match spectrum 1	36,518

Strategies to obtain more decisive criteria in establishing the possible candidates can be based on structural motifs found in MS fragmentation data, MS ionization data, and/or stability criteria. Applied to our first spectrum, Table 1 lists how these criteria reduce the number of possible structures. In the present case, although fragmentation spectra of the

individual $C_{12}H_{12}O_7$ isomers are not available, the detection of both hydrogen and sodium ion adducts of the $C_{12}H_{12}O_7$ isomer in question suggests that it contains OH and ether groups rather than carbonyls.[55] Furthermore, if we exclude oxygen chains longer than two (which most likely are not stable enough to endure the analytical procedure), only 36'516 stable molecules are left that match the experimental spectrum. This constitutes 0.01 and 0.03 % of the initially generated molecular graphs and stable structures, respectively. Given sufficiently accurate computational chemistry methods, the total energies of the structures (Fig. 1, SI) could be used to select or exclude certain structures; due to the approximative character of the GFN2-xTB calculations, we do not pursue this route further.

Starting from a complete list of all molecules is key to allow a bias-free filtering based on experimental input. Most importantly, filtering molecular graphs by MS fragments (or electrospray ionization information) is free of approximations from the theory side as no filtering based on computational chemistry or machine-learning methods is done at these early stages. The presence or absence of a structural feature in a given molecular graph can be determined readily.

Having a substantially shortened list (e.g. the 0.01 % for our case) allows for better calculations on the theory side once the filtering possibilities based on MS data are exhausted. There are two reasons for this: not only does a smaller chemical space require fewer training points to be accurately modeled with a machine learning approach, but also the reference data for the individual training points can be calculated using a better level of theory with fewer approximations.

Figure 6 illustrates how this filtering could be employed in a systematic fashion by repeatedly searching for structural features that divide the current set into two new sets of as equal size as possible. Similar to the method of binary search, this filters the total list of molecules in the fastest possible way if only tests for the existence of particular MS fragments are allowed. For the chemical space under consideration, an average of 15 fragment tests would be required to reduce the number of candidate molecules to below 10,000, if

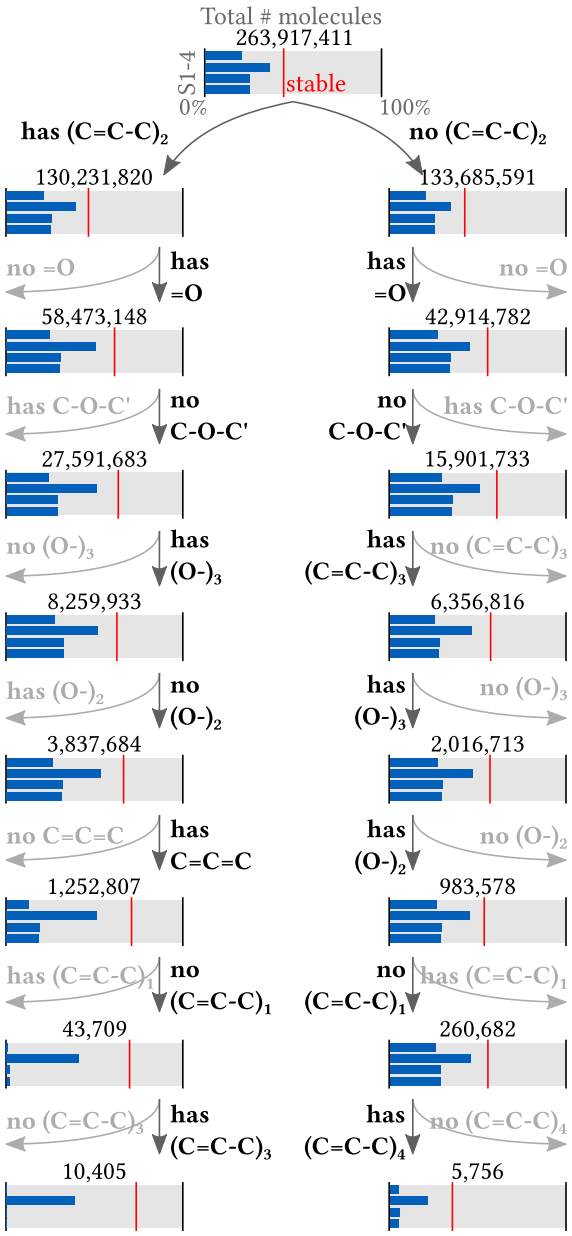

Figure 6: Idealized reduction in number of molecules as more and more conditions on the molecular graph are applied (from top to bottom). Note that these conditions are not founded in experimental fragmentation data, but rather illustrate the filtering process. Including all features would give a wide tree, so only two branches of the tree are shown. After each filter step, the total number of remaining molecules is shown where the red bar denotes the share thereof that is stable. The four bars illustrate how many of the residual molecules are compatible with the spectra 1-4 in this work. $(\mathrm{X})_n$ denotes that the structural feature X appears n-times in a row.

fragments were randomly distributed amongst all stable molecules and independent of each other. Typically, this is not the case, as exemplified by Figure 6 where we require tests for eight fragments until the number of molecules has been reduced to about 10,000 which would then be accessible for quantum chemistry calculations. In practice, this means that typically on the order of ten fragment tests would be needed to narrow the molecules down. For larger molecules, and particularly for molecules with a potentially branched structure, the number of fragments tests required will be larger. Starting from the exhaustive list of molecules however, it is clear exactly how many molecules remain to be analysed and thus going down such a decision tree could guide experimental work or MS data analysis. Furthermore, such an approach can not only determine whether additional criteria are still needed to identify a molecule, but can also identify which criteria will most efficiently narrow down the candidate molecule list.

The information whether molecules with these features are stable is typically not available while filtering the molecules, as long as the list of molecules is too long to render the required calculations feasible for multiple spectra. In this work however, we have performed the stability calculations for the complete list to illustrate in Figure 6 that the structural features alone are not always sufficient to determine stability or similarity to a UV/Vis spectrum. As the number of fragmentation results included increases in Figure 6, the share of stable molecules and those matching the four spectra in this work are initially roughly constant along the two paths shown. Only at the final stages does the feature list become more sensitive to the spectra in question. This emphasizes that real-world structure determinations will typically require a substantial number of confirmed/missing MS fragment determinations in addition to the UV/vis spectrum.

We have systematically enumerated all molecules with the sum formula $C_{12}H_{12}O_7$ containing two C_6-rings. We investigated whether the specific $C_{12}H_{12}O_7$ isomer behind an experimental brown carbon UV/Vis spectrum can be identified uniquely if a bias-free systematic comparison is done. To this end, we used a machine learning model to predict spectra for

all possible 123 million stable molecules in the set. We find that the experimental spectrum alone only halves the set of possible candidate molecules, so much additional information is required to determine the structure of a brown carbon molecule. Even with multiple MS fragments identified, there are tens to hundreds of thousands of potential structures that are compatible with the spectrum. The true scale of this problem only becomes clear once the exhaustive enumeration is done.

In light of our findings, we still consider identifying functional groups from MS the most promising strategy to reduce the number of candidates, especially if this information can be used early during the generation of molecular graphs. The advantage of using this information early is that it can be used to accelerate the graph generation. In addition, it reduces the chemical diversity, which may then reduce the error of the machine learning model.

Without the systematic enumeration of molecular targets, it becomes unclear whether sufficiently numerous molecular fragments have been identified to narrow down the list of potential molecules. This might lead to mis-identifications of molecules: Laskin et al.[14] suggested a possible structure for a $C_{12}H_{12}O_7$ product found in a syringol photooxidation chamber study, but our calculation shows that because of its dominant absorption band between 350 and 400 nm, the spectrum of this structure (Fig. 4, SI) is not consistent with any of the four experimentally measured spectra shown in Figure 1.

Based on the numerical evidence in this work, we expect that a systematic enumeration approach, where high-quality MS fragmentation data is included early on and where calculated spectra come from machine learning predictions based on quantum chemistry calculations, will make possible the rapid identification of individual brown carbon molecules based on their exact mass, MS fragmentation spectrum, and UV/Vis spectrum. In addition, such an approach will also yield guarantees that there are no other molecules that also would fit the experimental data.

Unraveling the chemistry behind SOA formation, and BrC formation in particular, is necessary until the work makes it possible to quantify their varied atmospheric sources.

The identification of molecular tracers and major products is important for connecting field measurements with lab studies and computer modeling of particular precursor chemistry. High resolution LCMS methods are currently the state-of-the-art for molecular identification of SOA and BrC species, but often return long lists of molecular formulae and associated UV-Vis absorption spectra. Even with further structural information from mass spectrometry fragmentation data for the most abundant ions, it is extremely time-consuming to work out chemical structures one by one with their associated reaction mechanisms, especially for larger oligomeric species. Furthermore, given the vast number of possible structures matching a chemical formula, there is no guarantee that the published structures generated in this way are even correct.

One source of secondary BrC believed to be atmospherically significant is the formation of oligomers during the aqueous-phase photooxidation of phenolic compounds, such as syringol, that are prevalent in wood smoke. This oligomer formation is thought to change the optical properties of BB aerosol particles during cloud processing, partially counteracting photobleaching and other aging processes. One light-absorbing dimer identified from this reaction has the formula $C_{12}H_{12}O_7$, which it shares with more than 260 million other dual-ring-retention products. Our study shows, that with further experimental constraints, our algorithm is able to shorten the list of possible structures to a few thousands to ten thousand candidates. As automated methods like those described here develop further and incorporate matching to experimental optical spectra and mass spectrometry fragmentation datasets, particular isomers extracted from lab and field aerosol particles may be more readily and rapidly identified. This will allow more detailed understanding of reaction mechanisms and precursor identification, and make it possible to design control strategies to reduce the climate effects of BrC and the adverse health effects of SOAs.

Acknowledgement

We would like to thank Stefan Heinen and Anders S. Christensen for support with the QML code. Research reported in this paper was supported by National Institute of General Medical Sciences of the National Institutes of Health (NIH) under award numbers UL1GM118979-02, TL4GM118980, and RL5GM118978 and NSF award number AGS-1826593. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We acknowledge technical support from the Division of Information Technology of CSULB. O.A.v.L. acknowledges support from the Swiss National Science foundation (407540_167186 NFP 75 Big Data) and from the European Research Council (ERC-CoG grant QML and H2020 projects BIG-MAP and TREX). This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreements #952165 and #957189. This result only reflects the author's view and the EU is not responsible for any use that may be made of the information it contains. This work was partly supported by the NCCR MARVEL, funded by the Swiss National Science Foundation.

Supporting Information Available

- Histogram of total ground state energies of 2 % of randomly selected structures.
- Correlation of features with compatibility of spectra 2 –4.
- Figure with tolerance regions of the spectra and illustration of matching probability.
- Description of procedure to generate molecules.
- Computational details of machine learning procedure.
- Description of Monte-Carlo procedure to determine matching probability with experimental spectra.
- Computational results of the proposed structure of Laskin et al.

References

(1) Laskin, A.; Laskin, J.; Nizkorodov, S. A. Chemistry of atmospheric brown carbon. *Chemical reviews* **2015**, *115*, 4335–4382.

(2) Feng, Y.; Ramanathan, V.; Kotamarthi, V. Brown carbon: a significant atmospheric absorber of solar radiation? *Atmospheric Chemistry & Physics Discussions* **2013**, *13*.

(3) Epstein, S. A.; Tapavicza, E.; Furche, F.; Nizkorodov, S. A. Direct photolysis of carbonyl compounds dissolved in cloud and fog droplets. *Atmos. Chem. Phys.* **2013**, *13*, 9461–9477.

(4) Kasthuriarachchi, N. Y.; Rivellini, L.-H.; Adam, M. G.; Lee, A. K. Light Absorbing Properties of Primary and Secondary Brown Carbon in a Tropical Urban Environment. *Environmental Science & Technology* **2020**, *54*, 10808–10819.

(5) Hettiyadura, A. P. S.; Garcia, V.; Li, C.; West, C. P.; Tomlin, J.; He, Q.; Rudich, Y.; Laskin, A. Chemical Composition and Molecular-Specific Optical Properties of Atmospheric Brown Carbon Associated with Biomass Burning. *Environmental Science & Technology* **2021**,

(6) Verma, V.; Rico-Martinez, R.; Kotra, N.; King, L.; Liu, J.; Snell, T. W.; Weber, R. J. Contribution of water-soluble and insoluble components and their hydrophobic/hydrophilic subfractions to the reactive oxygen species-generating potential of fine ambient aerosols. *Environmental science & technology* **2012**, *46*, 11384–11392.

(7) Chowdhury, P. H.; He, Q.; Carmieli, R.; Li, C.; Rudich, Y.; Pardo, M. Connecting the oxidative potential of secondary organic aerosols with reactive oxygen species in exposed lung cells. *Environmental science & technology* **2019**, *53*, 13949–13958.

(8) Shiraiwa, M.; Ueda, K.; Pozzer, A.; Lammel, G.; Kampf, C. J.; Fushimi, A.; Enami, S.;

Arangio, A. M.; Fröhlich-Nowoisky, J.; Fujitani, Y., et al. Aerosol health effects from molecular to global scales. *Environmental science & technology* **2017**, *51*, 13545–13567.

(9) Wang, X.; Heald, C.; Ridley, D.; Schwarz, J.; Spackman, J.; Perring, A.; Coe, H.; Liu, D.; Clarke, A. Exploiting simultaneous observational constraints on mass and absorption to estimate the global direct radiative forcing of black carbon and brown carbon. *Atmospheric Chemistry and Physics* **2014**, *14*, 10989–11010.

(10) Schilling Fahnestock, K. A.; Yee, L. D.; Loza, C. L.; Coggon, M. M.; Schwantes, R.; Zhang, X.; Dalleska, N. F.; Seinfeld, J. H. Secondary organic aerosol composition from C12 alkanes. *The Journal of Physical Chemistry A* **2015**, *119*, 4281–4297.

(11) De Haan, D. O.; Tapavicza, E.; Riva, M.; Cui, T.; Surratt, J. D.; Smith, A. C.; Jordan, M.-C.; Nilakantan, S.; Almodovar, M.; Stewart, T. N., et al. Nitrogen-containing, light-absorbing oligomers produced in aerosol particles exposed to methylglyoxal, photolysis, and cloud cycling. *Environ. Sci. Technol.* **2018**, *52*, 4061–4071.

(12) Fleming, L. T.; Lin, P.; Roberts, J. M.; Selimovic, V.; Yokelson, R.; Laskin, J.; Laskin, A.; Nizkorodov, S. A. Molecular composition and photochemical lifetimes of brown carbon chromophores in biomass burning organic aerosol. *Atmospheric Chemistry & Physics* **2020**, *20*.

(13) Denjean, C.; Formenti, P.; Picquet-Varrault, B.; Katrib, Y.; Pangui, E.; Zapf, P.; Doussin, J. A new experimental approach to study the hygroscopic and optical properties of aerosols: application to ammonium sulfate particles. *Atmospheric Measurement Techniques* **2014**, *7*, 183.

(14) Yu, L.; Smith, J.; Laskin, A.; Anastasio, C.; Laskin, J.; Zhang, Q. Chemical characterization of SOA formed from aqueous-phase reactions of phenols with the triplet excited state of carbonyl and hydroxyl radical. *Atmos. Chem. Phys* **2014**, *14*, 13801–13816.

(15) Lauraguais, A.; Coeur-Tourneur, C.; Cassez, A.; Seydi, A. Rate constant and secondary organic aerosol yields for the gas-phase reaction of hydroxyl radicals with syringol (2,6-dimethoxyphenol). *Atmospheric Environment* **2012**, *55*, 43 – 48.

(16) Wang, J.; Doussin, J.; Perrier, S.; Perraudin, E.; Katrib, Y.; Pangui, E.; Picquet-Varrault, B. Design of a new multi-phase experimental simulation chamber for atmospheric photosmog, aerosol and cloud chemistry research. *Atmospheric Measurement Techniques* **2011**, *4*, 2465.

(17) The On-Line Encyclopedia of Integer Sequences: A001349. 2020; `http://oeis.org/A001349`.

(18) Send, R.; Kuhn, M.; Furche, F. Assessing excited state methods by adiabatic excitation energies. *Journal of chemical theory and computation* **2011**, *7*, 2376–2386.

(19) Cisneros, C.; Thompson, T.; Baluyot, N.; Smith, A. C.; Tapavicza, E. The role of tachysterol in vitamin D photosynthesis - a non-adiabatic molecular dynamics study. *Phys. Chem. Chem. Phys.* **2017**, *19*, 5763–5777.

(20) Tapavicza, E.; Thompson, T.; Redd, K.; Kim, D. Tuning the photoreactivity of Z-hexatriene photoswitches by substituents – a non-adiabatic molecular dynamics study. *Phys. Chem. Chem. Phys.* **2018**, *20*, 24807–24820.

(21) Tapavicza, E. Generating Function Approach to Single Vibronic Level Fluorescence Spectra. *J. Phys. Chem. Lett.* **2019**, *10*, 6003–6009.

(22) Loos, P.-F.; Lipparini, F.; Boggio-Pasqua, M.; Scemama, A.; Jacquemin, D. A Mountaineering Strategy to Excited States: Highly Accurate Energies and Benchmarks for Medium Sized Molecules. *Journal of chemical theory and computation* **2020**, *16*, 1711–1741.

(23) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; Von Lilienfeld, O. A. Electronic spectra from TDDFT and machine learning in chemical space. *The Journal of chemical physics* **2015**, *143*, 084111.

(24) Wang, X.; Hayeck, N.; Brüggemann, M.; Yao, L.; Chen, H.; Zhang, C.; Emmelin, C.; Chen, J.; George, C.; Wang, L. Chemical Characteristics of Organic Aerosols in Shanghai: A Study by Ultrahigh-Performance Liquid Chromatography Coupled With Orbitrap Mass Spectrometry. *Journal of Geophysical Research: Atmospheres* **2017**, *122*, 11,703–11,722.

(25) McKay, B. D.; Piperno, A. Practical graph isomorphism, {II}. *Journal of Symbolic Computation* **2014**, *60*, 94 – 112.

(26) Chang, J. L.; Thompson, J. E. Characterization of colored products formed during irradiation of aqueous solutions containing H2O2 and phenolic compounds. *Atmospheric environment* **2010**, *44*, 541–551.

(27) Sun, Y.; Zhang, Q.; Anastasio, C.; Sun, J. Insights into secondary organic aerosol formed via aqueous-phase reactions of phenolic compounds based on high resolution mass spectrometry. *Atmospheric Chemistry & Physics* **2010**, *10*.

(28) Vione, D.; Albinet, A.; Barsotti, F.; Mekic, M.; Jiang, B.; Minero, C.; Brigante, M.; Gligorovski, S. Formation of substances with humic-like fluorescence properties, upon photoinduced oligomerization of typical phenolic compounds emitted by biomass burning. *Atmospheric Environment* **2019**, *206*, 197–207.

(29) Huang, D. D.; Zhang, Q.; Cheung, H. H.; Yu, L.; Zhou, S.; Anastasio, C.; Smith, J. D.; Chan, C. K. Formation and evolution of aqSOA from aqueous-phase reactions of phenolic carbonyls: comparison between ammonium sulfate and ammonium nitrate solutions. *Environmental science & technology* **2018**, *52*, 9215–9224.

(30) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **2019**, 1652–1671.

(31) Cordella, L. P.; Foggia, P.; Sansone, C.; Vento, M. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2004**, *26*, 1367–1372.

(32) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of cheminformatics* **2011**, *3*, 33.

(33) Open Babel version 2.4.0. `openbabel.org`, accessed 2020-03-01.

(34) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of computational chemistry* **1996**, *17*, 490–519.

(35) Schirmer, J. Beyond the random-phase approximation: A new approximation scheme for the polarization propagator. *Physical Review A* **1982**, *26*, 2395.

(36) Hättig, C.; Weigend, F. CC2 excitation energy calculations on large molecules using the resolution of the identity approximation. *J. Chem. Phys.* **2000**, *113*, 5154–5161.

(37) Klamt, A.; Schürmann, G. COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc. Perkin Trans.2* **1993**, *5*, 799.

(38) Lunkenheimer, B.; Köhn, A. Solvent effects on electronically excited states using the conductor-like screening model and the second-order correlated method ADC (2). *J. Chem. Theory Comput.* **2013**, *9*, 977–994.

(39) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

(40) Thompson, T.; Tapavicza, E. First-Principles Prediction of Wavelength-Dependent Product Quantum Yields. *J. Phys. Chem. Lett.* **2018**, *9*, 4758–4764.

(41) TURBOMOLE V7.2, TURBOMOLE GmbH, Karlsruhe, 2017; available from `http://www.turbomole.com`.

(42) Balasubramani, S. G.; Chen, G. P.; Coriani, S.; Diedenhofen, M.; Frank, M. S.; Franzke, Y. J.; Furche, F.; Grotjahn, R.; Harding, M. E.; Hättig, C., et al. TURBOMOLE: Modular program suite for ab initio quantum-chemical and condensed-matter simulations. *The Journal of chemical physics* **2020**, *152*, 184107.

(43) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(44) Faber, F. A.; Christensen, A. S.; Huang, B.; Von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *The Journal of chemical physics* **2018**, *148*, 241717.

(45) Christensen, A.; Faber, F.; Huang, B.; Bratholm, L.; Tkatchenko, A.; Muller, K.; von Lilienfeld, O. QML: A Python Toolkit for Quantum Machine Learning. 2017; `https://github.com/qmlcode/qml`.

(46) Häse, F.; Galván, I. F.; Aspuru-Guzik, A.; Lindh, R.; Vacher, M. How machine learning can assist the interpretation of ab initio molecular dynamics simulations and conceptual understanding of chemistry. *Chemical science* **2019**, *10*, 2298–2307.

(47) Westermayr, J.; Marquetand, P. Machine learning for electronically excited states of molecules. *Chemical Reviews* **2020**,

(48) Xue, B.-X.; Barbatti, M.; Dral, P. O. Machine Learning for Absorption Cross Sections. *The Journal of Physical Chemistry A* **2020**, *124*, 7199–7210.

(49) Schalk, O.; Geng, T.; Thompson, T.; Baluyot, N.; Thomas, R. D.; Tapavicza, E.; Hansson, T. Cyclohexadiene Revisited: A Time-Resolved Photoelectron Spectroscopy and ab Initio Study. *J. Phys. Chem. A* **2016**, *120*, 2320.

(50) Tapavicza, E.; von Rudorff, G. F.; De Haan, D. O.; Contin, M.; George, C.; Riva, M.; von Lilienfeld, O. A. Elucidating atmospheric brown carbon - Supplanting chemical intuition with exhaustive enumeration and machine learning. 2021; `https://doi.org/10.5281/zenodo.4432153`.

(51) Tapavicza, E.; von Rudorff, G. F.; De Haan, D. O.; Contin, M.; George, C.; Riva, M.; von Lilienfeld, O. A. Elucidating atmospheric brown carbon - Supplanting chemical intuition with exhaustive enumeration and machine learning. 2021; `https://doi.org/10.5281/zenodo.4432606`.

(52) Mckay, D. J.; Wright, J. S. How long can you make an oxygen chain? *Journal of the American Chemical Society* **1998**, *120*, 1003–1013.

(53) Rappoport, Z. *The Chemistry of Peroxides, Parts 1 and 2*; John Wiley & Sons, 2007; Vol. 168.

(54) Sarkar, R.; Boggio-Pasqua, M.; Loos, P.-F.; Jacquemin, D. Benchmarking TD-DFT and Wave Function Methods for Oscillator Strengths and Excited-State Dipole Moments. 2020.

(55) Swanson, K. D.; Spencer, S. E.; Glish, G. L. Metal cationization extractive electrospray

ionization mass spectrometry of compounds containing multiple oxygens. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 1030–1035.

Graphical TOC Entry