# Elucidating an Atmospheric Brown Carbon Species-Toward Supplanting Chemical Intuition with Exhaustive Enumeration and Machine Learning

Enrico Tapavicza, Guido Falk von Rudorff, David De Haan, Mario Contin, Christian George, Matthieu Riva, O. Anatole Von Lilienfeld

# Elucidating an Atmospheric Brown Carbon Species—Toward 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

ABSTRACT: Brown carbon (BrC) is involved in atmospheric light absorption and 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 environment and health impact. Structure determination of BrC is challenging, due to the lack of experiments providing molecular fingerprints and the sheer number of molecular candidates with identical mass. Suggestions based on chemical intuition are prone to errors due to the inherent bias. We present an unbiased algorithm, using graph-based molecule generation and machine learning, which can identify all molecular structures of compounds involved in biomass burning and the composition of BrC. We apply this algorithm to C12H12O7, a light-absorbing "test case" molecule identified in chamber experiments on the aqueous photo-oxidation of syringol, a prevalent marker in wood smoke. Of the 260 million molecular graphs, the algorithm leaves only 36,518 (0.01%) as viable candidates matching the spectrum. Although no unique molecular structure is obtained from only a chemical formula and a UV/vis absorption spectrum, we discuss further reduction strategies and their efficacy. With additional data, the method can potentially more rapidly identify isomers extracted from lab and field aerosol particles without introducing human bias.

**KEYWORDS:** *biomass burning, chemical diversity, chemical space, structure determination, oligomers, light absorption*

## 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 chemical modification due to reactions in the atmosphere, sometimes forming light-absorbing oligomers with large absorption coefficients.[6] 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 not only is important to evaluate their toxicology,[7,8] carcinogenic activity, and receptor binding[9] in order to assess public health impact but also allows us to make predictions on molecular stability and chemical fate; the latter properties are important properties to assess BrC's further implications on atmospheric processes and climate forcing.[10] Thus, structure and precursor identification takes up a key role in atmospheric chemistry, connecting field studies with lab experiments and computer modeling. Unraveling 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.[11−13]

Typically, the formation and further reactions of SOA under specific atmospheric conditions can be simulated in atmospheric chamber experiments.[14,15] 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.[5] While the detected exact mass of the compounds gives complete information about the chemical sum formula, the exact chemical structures often remain unknown. Due to the generally low concentrations of precursor compounds that must be used in order to realistically simulate atmospheric conditions in chamber experiments, spectroscopic methods that provide conclusive information about the

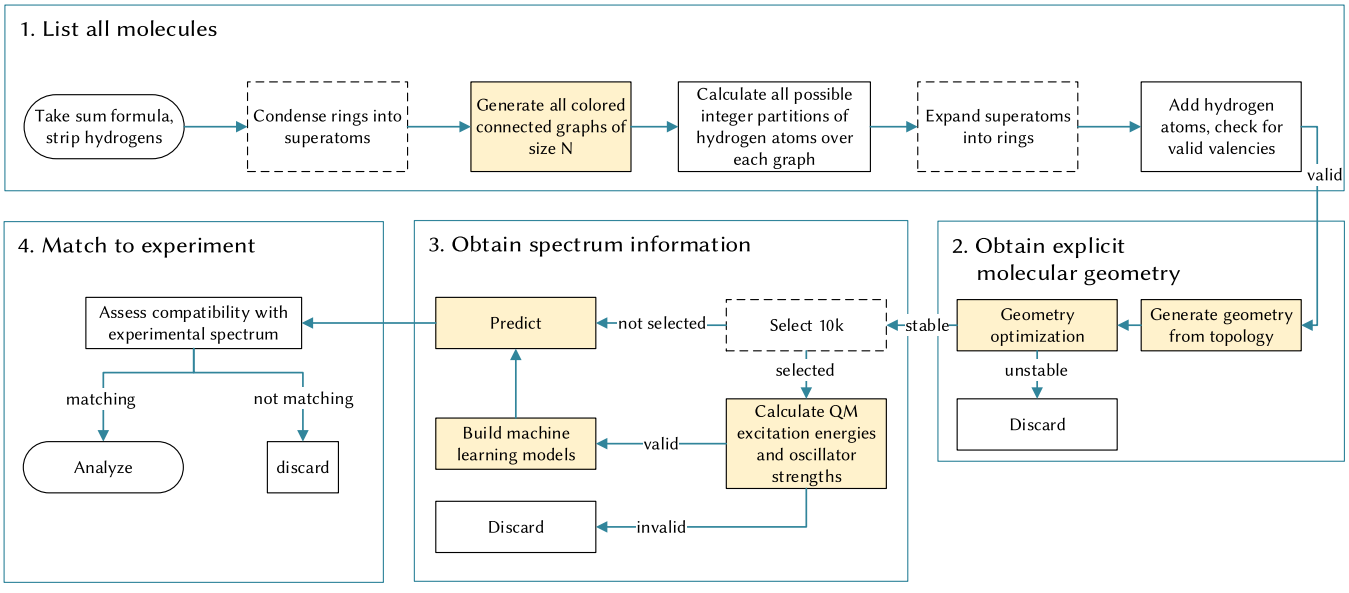


**Figure 1.** Workflow to identify all molecules of a given sum formula that fit an experimental UV/vis absorbance spectrum. Of all possible molecules of this sum formula (box 1), first, the stable ones are selected (box 2) for which machine learning-assisted spectra are obtained (box 3) which then can be compared to experiment (box 4). Steps for which we used external software are shaded orange, and steps which only improve performance but are not strictly required have a dashed outline.

chemical constitution of reaction products, such as NMR, are unfortunately rarely applicable. Furthermore, it is questionable if only one constitutional isomer of a detected chemical sum formula is present, and imperfect LC separations are common. Given the chemical complexity of SOA, it is often more realistic to assume that a detected chemical sum formula may be composed of different isomers with similar physical chemical properties. Likewise, UV/vis spectra may be complicated by contributions from coeluting products. Thus, to gain confidence in the correctness of the chosen structure or to rule out structures not consistent with experimental observations, one should use as many experimental observables as available. Molecular structures of products generated in atmospheric chamber experiments are typically proposed on the basis of mass spectrometric data and chemical intuition. The high absorption coefficients of BrC allow the use of UV/vis spectra to not only identify which mass spectrometric peaks correspond with light-absorbing molecules but also probe if proposed BrC product structures are consistent with experimental observations.[12,16,17]

The difficulty of determining the structure of a BrC compound, even when its exact mass is known, is caused by a lack of experiments providing characteristic fingerprints of these molecules and amplified by the sheer number of possible molecular structures associated with a given molecular formula. Chemical intuition and ad hoc assumptions for structural elements often provide the basis for suggesting particular structures but are time-consuming to develop and are prone to errors due to their biased nature. The correct structure of a detected product molecule may not be considered due to the vast number of possibilities, or it may get rejected because it does not seem probable based on the chemist's prior experience. Furthermore, the choice of isomers considered might also be guided by the availability of standards for comparison of their physicochemical properties. These biases work against the discovery of novel constitutional isomers and the atmospheric chemistry which they represent.

In an attempt to minimize human bias in the proposition of candidate structures, here, we approach 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.[12,18] We develop an unbiased algorithm with the goal of determining the feasibility of automated molecular structure identification of compounds involved in BB and the composition of BrC. 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.

One source of secondary brown carbon believed to be atmospherically significant is the formation of oligomers during the aqueous-phase photo-oxidation 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. When syringol is photo-oxidized with OH radicals or triplet carbon ($^3C^*$) species in the aqueous phase, one of the seven major products detected by negative-mode nanodesorption electrospray MS has the sum formula $C_{12}H_{12}O_7$.[18] Isomers with this formula were associated with BrC absorbance peaks in the aqueous aerosol phase in as-yet-unpublished syringol photo-oxidation experiments at the CESAM chamber[19] where syringol was oxidized with OH radicals. In addition, $C_{12}H_{12}O_7$ was the largest peak identified by UHPLC-(+)ESI-MS with a formula containing less than 20 heavy atoms. Thus, $C_{12}H_{12}O_7$ was selected as an appropriate brown carbon candidate for this study.

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[20] even without assigning elements to the individual nodes and without assigning bond orders. Annotating the graphs with the atom elements drastically increases this number, while following

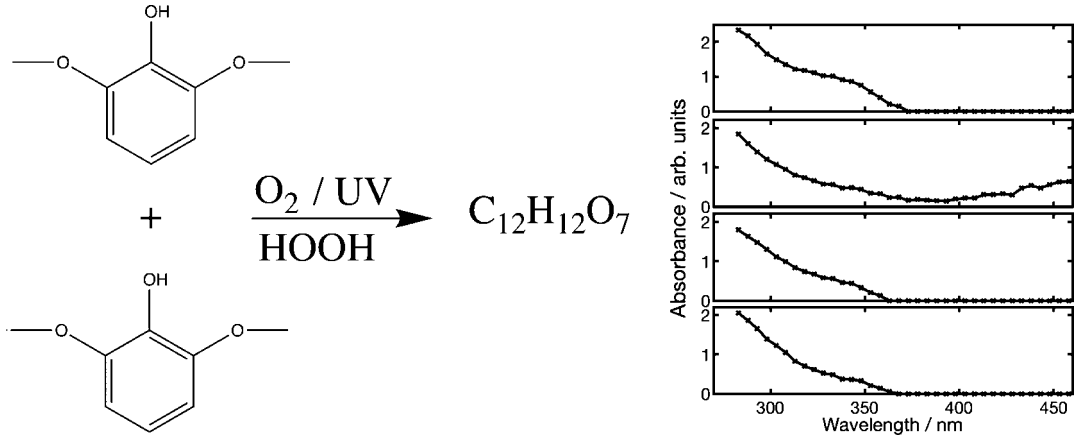


**Figure 2.** In the proposed reaction scheme (left), aqueous-phase syringol photo-oxidation forms $C_{12}H_{12}O_7$, a dimer product with an unknown structure that correlates with brown carbon formation. Right: four different experimental UV/vis absorbance spectra, measured at four different retention times corresponding with elution of a different $C_{12}H_{12}O_7$ isomer.

chemical bonding rules again reduces the total count of molecular graphs with this sum formula. The unrestricted total count of molecular graphs with valid Lewis structures of that sum formula is unknown. 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 physicochemical 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.[21−25] However, the computational resources necessary for the spectra prediction quickly become unfeasible with increasing number of atoms, since the size of the chemical space for one given sum formula rapidly grows.

Here, we developed a computational workflow (cf. Figure 1) 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, which leverages massively parallel computation. The problem of quantum chemical spectra prediction of a large number of molecules is solved by making use of machine learning to predict spectral properties of the molecules.[26] In a Monte-Carlo procedure, we then determine the likelihood that specific feature groups give rise to the experimentally observed spectrum.

The workflow starts from an unbiased and exhaustive generation of all possible molecular graphs (box 1 in Figure 1). 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 the 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 Section.** The filter collection and extraction protocols have been described previously.[27] Briefly, to determine UV/vis spectra corresponding to $C_{12}H_{12}O_7$ products, SOA formed in syringol photo-oxidation experiments at the CESAM chamber were collected overnight onto a Teflon filter (1 $\mu$m pores, 47 mm diam.), which was kept at −20 C until analysis. Each Teflon filter was spiked with caffeine (final concentration 100 ppb) as the internal standard and then extracted twice with 6 mL of acetonitrile and agitated for 20 min with an orbital shaker at 1000 rpm. The extracts were then filtered with a syringe filter (0.2 $\mu$m, 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, Fisher Scientific). Finally, the filter extracts were analyzed by ultrahigh performance LC (Dionex 3000, Thermo Scientific) using a Water Acquity HSS C18 column (1.8 $\mu$m, 100 × 2.1 mm) 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 the positive or negative mode. The mobile phase used was constituted of (A) 0.1% formic acid in water (Optima LC/MS, Fisher 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 $\mu$L/min: 0 to 13 min B from 1 to 100% and 13.1 min B 1% for 9 min. Raw data were processed with MZmine 2.51. A single-ion chromatogram for (+) mode ion signal matching formula $C_{12}H_{12}O_7$ was generated. UV/vis spectra corresponding to $C_{12}H_{12}O_7$ peak elution times (minus the time delay between the two detectors) were then generated.

Wavelength-dependent molar absorption coefficients of $C_{12}O_7H_{12}$ were estimated from absorbance spectra recorded at four different retention times of sodium adducts of $C_{12}O_7H_{12}$ (Figure 2) using the Beer−Lambert law. To obtain an estimate of the necessary concentrations, surrogate standards were used to quantify ESI efficiencies. Since ionization efficiencies vary significantly between compounds, the use of surrogate standards introduces an order-of-magnitude uncertainty of the estimated molar absorption coefficients. This uncertainty is taken into account in the evaluation of the compatibility between the computed spectra and the experimental spectra (see below and the Supporting Information).

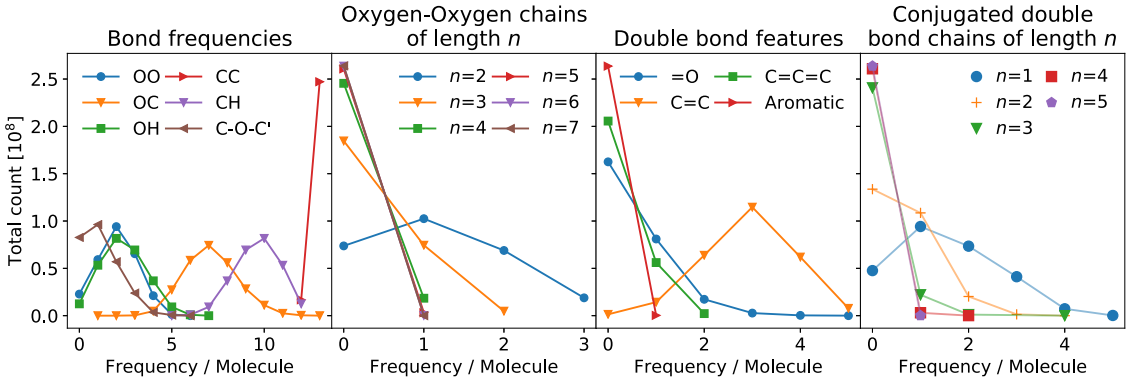


**Figure 3.** Total number of molecular graphs with the given feature occurrences. Panel (1): bond count frequencies and ether bridges connecting the two carbon rings; panel (2): count of oxygen−oxygen chains of length $n$; panel (3): count of carbonyl sites, allene sites, double bonds, and aromatic 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.

**Molecule Generation.** For the sum formula $C_{12}H_{12}O_7$, we systematically[28] 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[29] of two syringol units to $C_{16}H_{18}O_6$, the most abundant SOA product identified in previous studies,[18,30] followed by further oxidation and fragmentation to $C_{12}H_{12}O_7$.[18] To develop the method in this work as a proof of concept, 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. While other products are possible, including a proposed $C_{12}H_{12}O_7$ structure with only one $C_6$ ring remaining,[18] we note that demethylation of methoxy groups is commonly observed during photo-oxidation of vanillin,[31] syringaldehyde, and acetosyringol.[32] Furthermore, half of the syringol SOA product structures proposed by Yu et al.[18] have lost at least one methoxy carbon, and 16% of their proposed structures have lost all methoxy carbons, making it probable that at least some $C_{12}H_{12}O_7$ isomers have two $C_6$ rings. Of course, with increasing computational power, this analysis would ideally be performed without any initial structural assumption. Technically, this enumeration is performed by (a) enumerating all potential molecular graphs ignoring hydrogens, (b) constructing all possible hydrogen saturations of these graphs, and (c) filtering all molecules which are not stable in GFN2-xTB calculations.[33] The protocol for these steps is summarized in Figure 1 and detailed in the Supporting Information and is based on refs 28 and 34353637.

**Electronic Structure Methods and Machine Learning.** Electronic excitation energies and their oscillator strengths are readily computed by many electronic structure methods and can be converted into absorption spectra. To assess the absorption spectrum, we computed the lowest three excitation energies and their corresponding oscillator strengths using the algebraic diagrammatic construction to second order (ADC(2)) method.[38,39] To include effects of water solvation in the calculation, we employed the conductor-like screening model (COSMO)[40] using a dielectric constant of 80.1 and a refractive index of 1.3325.[41] The def2-TZVP basis set was used.[42] This approach has been shown to yield accurate excitation energies.[43−45] Calculations were carried out with TURBOMOLE V7.2.[46,47]

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 the training set for the Kernel-ridge-regression (KRR) method[48] with the FCHL molecular representation[49] as implemented in the QML toolkit.[50]

Note that these 10,000 molecules have been selected as representatives of the whole chemical space under consideration. Since homogeneous data sets can be expected to exhibit better learning, future refinements of our approach might repeat this training set selection step after the first few features of molecular graphs have been eliminated by experimental results, thus reducing the chemical space under consideration. This might provide more accurate models, further increasing the filtering efficiency of our work. In this work, we only select one training set for high-quality reference calculations to showcase how this alone is sufficient to narrow down the list of candidate molecules for a given spectrum.

Machine learning, in general, and KRR, in particular, have been successfully used to predict excited-state properties,[51−53] typically highlighting the need for high-quality reference data. A total of 82 molecules were excluded since they exhibited negative excitation energies, which indicates that the ground state has multireference character and is close to or at a singlet instability.[54] We determined optimal hyperparameters for the kernel widths and regularizer with fivefold cross validation (see the Supporting Information). 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[3,55] and compare it to the experimental ones. The electronic absorption spectrum for one excitation and oscillator strength $f$ is simulated using a Gaussian line shape $\rho(\lambda)$ according to

$$\epsilon(\lambda) = 2.303\frac{4\pi^2 q_e}{3\hbar c}\rho(\lambda)N_A f \qquad (1)$$

where $\hbar$ is Planck's reduced constant, $c$ denotes the speed of light, $q_e$ is the electronic charge, and $N_A$ is Avogadro's number. We employ a Monto-Carlo method (see the Supporting Information) 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 (see Figure 2) and predicted spectrum are separated by at most one standard deviation of both modeling and experimental uncertainties.

## RESULTS AND DISCUSSION

In our analysis of filters collected after syringol + OH photooxidation experiments in a chamber containing aqueous aerosol, $C_{12}H_{12}O_7$ isomers were observed to elute at four different retention times. The rather similar UV/vis spectra that were observed at the corresponding elution times of the four isomers are shown in Figure 2.

**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, the two rings can either be directly connected 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 (Figure 3). Analyzing the number of C−O bonds, we find a peaked distribution ranging from 1 to 13, with a maximum probability at 7. The number of 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 Figure 3). 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 (Figure 3, 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[56] together with the reference data for the machine learning model[57] and the code used in this work.[58] 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.[59] 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 six members. Experimentally, four-membered chains have been confirmed.[59] Dioxiranes have been known experimentally since 1978, although their existence was already predicted in 1899 by Bayer and Villiger.[60] An indication of the relative stability of the molecules can also be based on total electronic energies (see Figure S1).

**Electronic Spectra Prediction.** For a random set of 9918 molecules, ADC(2)/COSMO calculations resulted in positive excitation energies and oscillator strengths for the lowest three states (Figure 4) and therefore could be used for training the models. The data are available online.[57] From Figure 4, we see that the first excited state ($S_1$) contributes with the highest oscillator strengths in the region between 0.12 and 0.16 au, where the experimental absorption band is located, but the second ($S_2$) and third ($S_3$) excited states 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 (Figure 5). We assess the accuracy with the mean absolute error (MAE) which gives the mean absolute deviation of the predicted value from the in silico ADC(2)/COSMO value. It is obtained separately for each of the excitation energies and oscillator strengths and only measures the individual spectrum component, not the overall accuracy of the spectrum. For a training set of 9000 molecules, predictions exhibit 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

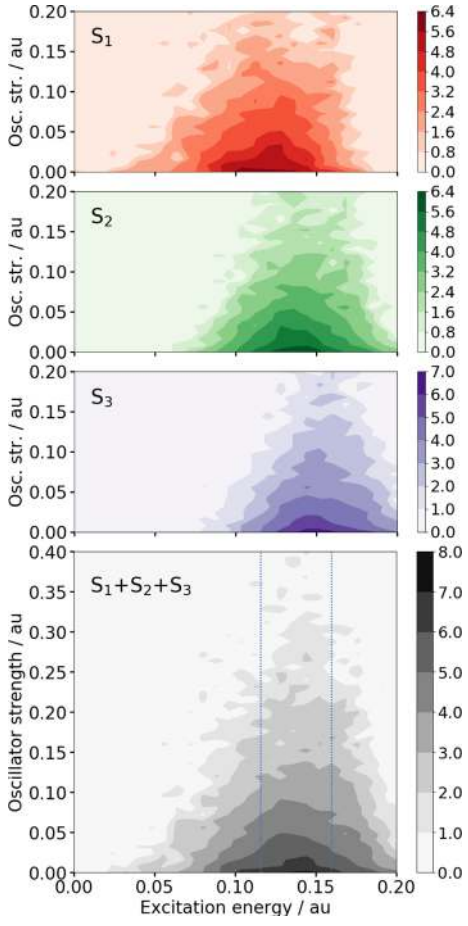


**Figure 4.** Distribution of ADC(2)/COSMO excited states as a function of excitation energy and oscillator strengths of the 9918 training molecules. The distribution is shown separately for the lowest three excited states (top three panels) and combined for all three states (bottom panel). The color 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.

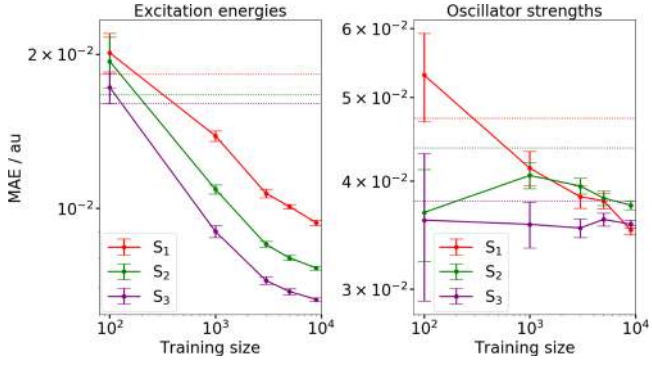


**Figure 5.** Left: MAE as a function of training set size for the learning of the lowest three excitation energies. Right: MAE as a function of training size for the learning of the lowest three oscillator strengths. In both plots, the dashed lines indicate the error of the null model, using the same color code for the different states.

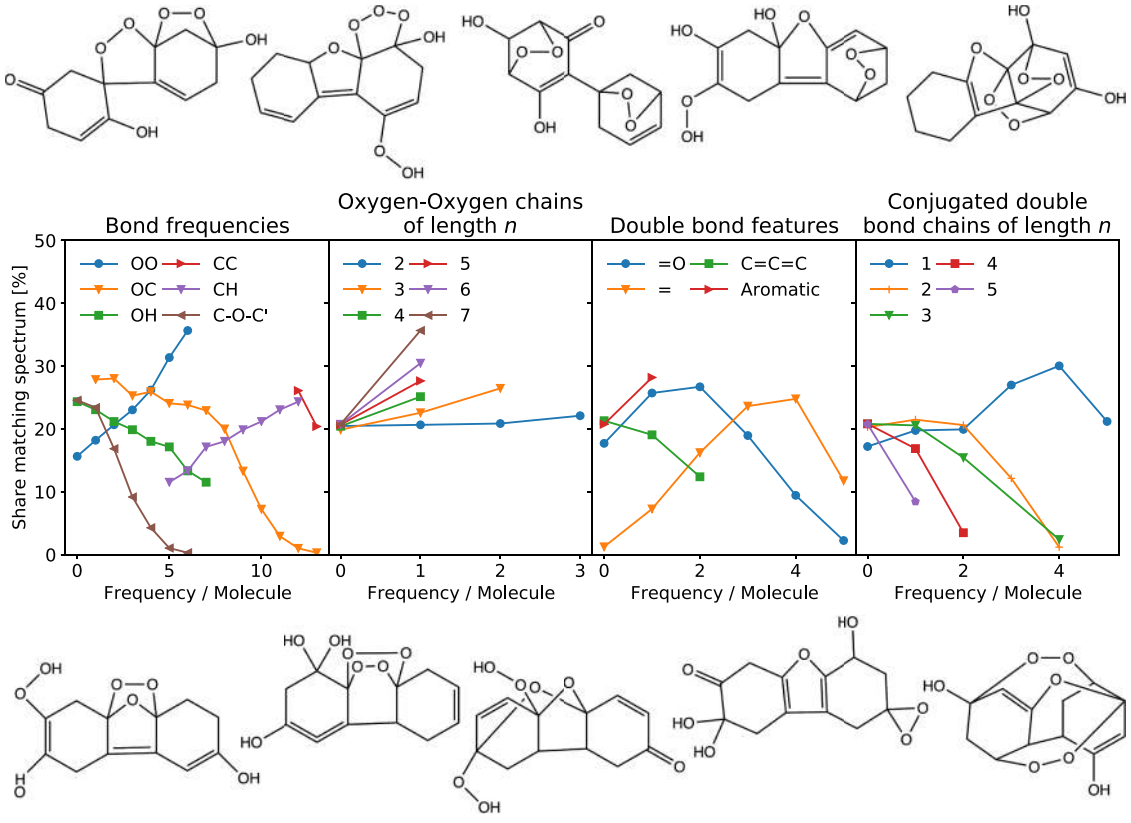


**Figure 6.** Per-feature probability (share matching spectrum) of molecular structures being compatible with the first spectrum in Figure 2 shown in the panels. Conditional probabilities are exemplified by grouping all molecules by their feature vector and calculating the share of matching molecules for each group. Low-energy representatives of the largest groups are shown for matching (top) and nonmatching (bottom) molecules.

of ADC(2) with respect to experimental values, which was previously determined to be 8 mHa (0.21 eV).[61]

As the learning curves in Figure 5 show, the model becomes systematically more accurate as the number of training points is increased, which confirms that the model indeed is learning. From learning theory,[62] one would expect a negative linear relationship in the learning curves in the limit of large number of training points where the slope denotes how fast a model is learning. 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 Figure 5), 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 (Figure 5, right). Interestingly, oscillator strengths of $S_1$ benefit the most from KKR, whereas, for $S_2$ and $S_3$, learning curves are comparably flat. It has been noted in previous work[26] that representations based on the coordinates and atomic number of the molecules might not be optimal for oscillator strengths, which depend on the transition density between ground and excited states and therefore might benefit from representations that include quantities related to the transition density. The reason why the oscillator strengths for $S_3$ are better predicted than for $S_2$ and $S_1$ is probably due to the fact that they are more uniform since they can be described more "Rydberg like" than oscillator strengths of $S_2$ and $S_1$ and therefore are less sensitive to variations in the molecular structure. The same holds for the comparison between oscillator strengths of $S_2$ and $S_1$; in this case, $S_2$ is predicted with better accuracy than $S_1$.

As illustrated by Figure 5, adding more training data typically improves the prediction accuracy. Increasing the data set beyond 10,000 training points, however, not only is costly in terms of the number of quantum chemistry calculations but also increases prediction times. We expect better performance by incorporating active learning schemes, where a preselection of the training data improves prediction accuracy for smaller training sets. Evaluating the different active learning schemes is beyond the scope of this work.

Using the ML model based on the 9918 training molecules, we predicted the lowest three excitation energies and oscillator strengths for the remaining 120 million stable structures. For our models, training including hyperparameter optimization took about 1 day on a workstation, while a single prediction requires about four core seconds. For larger models, the training time would increase quadratically with the number of training points (unless approximation schemes are employed[63]), but query times which are more relevant would increase only linearly.

**Establishing Matching Characteristics.** We present the analysis for the first spectrum on the top right of Figure 2; results for the remaining three spectra are very similar (see Figure S2). Out of the 123 million stable molecules, 55 million match this spectrum according to the criteria defined in the Supporting Information. For every structure, we determined a feature vector that describes the structural features in the molecules (Figure 6). 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 varies between 2 and 6. For instance, for the number of carbon−carbon bonds, only two values are possible (12 and 13), but for a number of two-membered oxygen chains, the number of entries amounts to four, because possible values are 0, 1, 2, and 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 us to correlate molecular features with the probability that it causes the experimentally observed spectrum (see Figure 2). Note that this constitutes a conditional probability, so, for example, molecules that have no carbonyl group and those that have no double bond are different groups and are of different sizes. Therefore, no direct comparisons between different lines in Figure 6 are allowed, only within one line for one feature. Analyzing the features in Figure 6, we find that, for example, the probability that a molecule is consistent with the 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 nonmatching molecules are given on top and bottom of Figure 6, respectively. The corresponding groups of molecules are huge: just for the first molecule on the top left in Figure 6, there are 695,039 stable molecules that match the experimental spectrum and have an identical feature vector.

Each of the other molecules shown in Figure 6 is just one representative of similarly large groups of feature-identical stable molecules. While the presence of individual molecular 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 nonmatching feature groups (see Figure 6) 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., bathochromic shift) are of limited utility in the selection of candidate brown carbon molecules.

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.[64] 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 (263 million) and stable structures (123 million), respectively. Given sufficiently accurate computational chemistry methods, the total energies of the structures (Figure S1) 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.

**Table 1. Summary of How the 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 |

Starting from a complete list of all molecules is key to allow a bias-free filtering based on the experimental input. Most importantly, filtering molecular graphs by MS fragments (or ESI information) is free of approximations from the theory side as no filtering based on computational chemistry or machine-learning methods is carried out 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 7 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 fragments were randomly distributed among all stable molecules and independent of each other. Typically, this is not the case, as exemplified by Figure 7, 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 analyzed 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

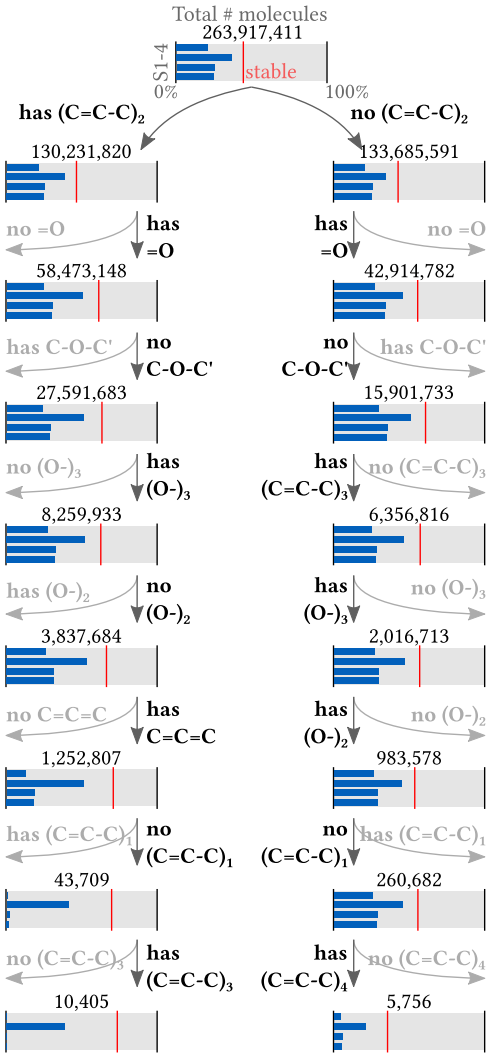


**Figure 7.** Idealized reduction in the number of molecules as more and more conditions on the molecular graph are applied (from top to bottom). Note that these conditions are not founded in experimental fragmentation data but rather illustrate the filtering process. Including all features would give 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. $(X)_n$ denotes that the structural feature X appears *n*-times in a row.

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 7 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 7, 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 (regardless of whether they are aromatic or not). 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 carried out. 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 carried out.

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 misidentifications of molecules: Laskin et al.[18] suggested a possible structure for a $C_{12}H_{12}O_7$ product found in a syringol photo-oxidation chamber study, but our calculation shows that because of its dominant absorption band between 350 and 400 nm, the spectrum of this structure (Figure S4) 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 are 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.

Nevertheless, even in the case of a large ensemble of valid structures, one could potentially rule out certain reaction pathways in both, the formation and also the further modifications of these compounds, if common structural patterns can be identified and be associated with a particular reaction; this, however, is beyond the scope of this study.

## ENVIRONMENTAL IMPLICATIONS

Unraveling the chemistry behind SOA formation, 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 MS 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 photo-oxidation of phenolic compounds, such as syringol, that are prevalent in wood smoke.[29,65−68] This oligomer formation is thought to change the optical properties of BB aerosol particles during cloud processing,[69] partially counteracting photobleaching and other aging processes.[70] 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 unbiased, automated methods such as those described here develop further and incorporate matching to experimental optical spectra and MS fragmentation data sets, 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.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.1c00885.

Method description for determination of UV/vis spectra corresponding to $C_{12}H_{12}O_7$ product isomers; histogram of total ground-state energies for a stratified subset of structures; correlation of features with compatibility of spectra 2−4; figure with tolerance regions of the spectra and illustration of matching probability; description of the procedure to generate molecules; computational details of the machine learning procedure; description of the Monte-Carlo procedure to determine matching probability with experimental spectra; and computational results of the proposed structure of Laskin et al. (PDF)

## AUTHOR INFORMATION


### Corresponding Author

**Enrico Tapavicza** − *Department of Chemistry and Biochemistry, California State University, Long Beach, Long Beach, California 90840, United States;* orcid.org/0000-0002-0640-0297; Email: enrico.tapavicza@csulb.edu

### Authors

**Guido Falk von Rudorff** − *Faculty of Physics, University of Vienna, AT-1090 Wien, Austria; Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials (MARVEL), Department of Chemistry, University of Basel, CH-4056 Basel, Switzerland*

**David O. De Haan** − *Department of Chemistry and Biochemistry, University of San Diego, San Diego, California 92110, United States;* orcid.org/0000-0003-4559-2284

**Mario Contin** − *Facultad de Farmacia y Bioquímica, Departamento de Química Analitica y Fisicoquímica, Universidad de Buenos Aires, Buenos Aires C1113AAD, Argentina*

**Christian George** − *Université Lyon, Université Claude Bernard Lyon 1, CNRS, IRCELYON, 69626 Villeurbanne, France;* orcid.org/0000-0003-1578-7056

**Matthieu Riva** − *Université Lyon, Université Claude Bernard Lyon 1, CNRS, IRCELYON, 69626 Villeurbanne, France;* orcid.org/0000-0003-0054-4131

**O. Anatole von Lilienfeld** − *Faculty of Physics, University of Vienna, AT-1090 Wien, Austria; Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials (MARVEL), Department of Chemistry, University of Basel, CH-4056 Basel, Switzerland;* orcid.org/0000-0001-7419-0466


### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS


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 the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under award numbers R15GM126524, 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.


## REFERENCES


(1) Laskin, A.; Laskin, J.; Nizkorodov, S. A. Chemistry of atmospheric brown carbon. *Chem. Rev.* **2015**, *115*, 4335−4382.

(2) Feng, Y.; Ramanathan, V.; Kotamarthi, V. Brown carbon: a significant atmospheric absorber of solar radiation? *Atmos. Chem. Phys. Discuss.* **2013**, *13*, 8607.

(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. Y. Light Absorbing Properties of Primary and Secondary Brown Carbon in a Tropical Urban Environment. *Environ. Sci. Technol.* **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. *Environ. Sci. Technol.* **2021**, *55*, 2511.
(6) Kasthuriarachchi, N. Y.; Rivellini, L.-H.; Chen, X.; Li, Y. J.; Lee, A. K. Y. Effect of Relative Humidity on Secondary Brown Carbon Formation in Aqueous Droplets. *Environ. Sci. Technol.* **2020**, *54*, 13207−13216.
(7) 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. *Environ. Sci. Technol.* **2012**, *46*, 11384−11392.
(8) 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. *Environ. Sci. Technol.* **2019**, *53*, 13949−13958.
(9) Shiraiwa, M.; Ueda, K.; Pozzer, A.; Lammel, G.; Kampf, C. J.; Fushimi, A.; Enami, S.; Arangio, A. M.; Fröhlich-Nowoisky, J.; Fujitani, Y.; Furuyama, A.; Lakey, P. S. J.; Lelieveld, J.; Lucas, K.; Morino, Y.; Pöschl, U.; Takahama, S.; Takami, A.; Tong, H.; Weber, B.; Yoshino, A.; Sato, K. Aerosol health effects from molecular to global scales. *Environ. Sci. Technol.* **2017**, *51*, 13545−13567.
(10) Wang, X.; Heald, C. L.; Ridley, D. A.; Schwarz, J. P.; Spackman, J. R.; Perring, A. E.; Coe, H.; Liu, D.; Clarke, A. D. Exploiting simultaneous observational constraints on mass and absorption to estimate the global direct radiative forcing of black carbon and brown carbon. *Atmos. Chem. Phys.* **2014**, *14*, 10989−11010.
(11) 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. *J. Phys. Chem. A* **2015**, *119*, 4281−4297.
(12) 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.; de Loera, A.; De Haan, A. C.; Cazaunau, M.; Gratien, A.; Pangui, E.; Doussin, J.-F. Nitrogen-containing, light-absorbing oligomers produced in aerosol particles exposed to methylglyoxal, photolysis, and cloud cycling. *Environ. Sci. Technol.* **2018**, *52*, 4061−4071.
(13) 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. *Atmos. Chem. Phys.* **2020**, *20*, 1105.
(14) Denjean, C.; Formenti, P.; Picquet-Varrault, B.; Katrib, Y.; Pangui, E.; Zapf, P.; Doussin, J. F. A new experimental approach to study the hygroscopic and optical properties of aerosols: application to ammonium sulfate particles. *Atmos. Meas. Tech.* **2014**, *7*, 183.
(15) Cocker, D. R.; Flagan, R. C.; Seinfeld, J. H. State-of-the-art chamber facility for studying atmospheric aerosol chemistry. *Environ. Sci. Technol.* **2001**, *35*, 2594−2601.
(16) Shapiro, E. L.; Szprengiel, J.; Sareen, N.; Jen, C. N.; Giordano, M. R.; McNeill, V. F. Light-absorbing secondary organic material formed by glyoxal in aqueous aerosol mimics. *Atmos. Chem. Phys.* **2009**, *9*, 2289−2300.
(17) Grace, D. N.; Sharp, J. R.; Holappa, R. E.; Lugos, E. N.; Sebold, M. B.; Griffith, D. R.; Hendrickson, H. P.; Galloway, M. M. Heterocyclic product formation in aqueous brown carbon systems. *ACS Earth Space Chem.* **2019**, *3*, 2472−2481.
(18) 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.
(19) Wang, J.; Doussin, J. F.; 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. *Atmos. Meas. Tech.* **2011**, *4*, 2465.
(20) The On-Line Encyclopedia of Integer Sequences: A001349, 2020. http://oeis.org/A001349 (accessed 2021-05-31).
(21) Send, R.; Kühn, M.; Furche, F. Assessing excited state methods by adiabatic excitation energies. *J. Chem. Theory Comput.* **2011**, *7*, 2376−2386.
(22) 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.
(23) 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.
(24) Tapavicza, E. Generating Function Approach to Single Vibronic Level Fluorescence Spectra. *J. Phys. Chem. Lett.* **2019**, *10*, 6003−6009.
(25) 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. *J. Chem. Theory Comput.* **2020**, *16*, 1711−1741.
(26) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; Von Lilienfeld, O. A. Electronic spectra from TDDFT and machine learning in chemical space. *J. Chem. Phys.* **2015**, *143*, 084111.
(27) 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. *J. Geophys. Res.: Atmos.* **2017**, *122*, 11703−11722.
(28) McKay, B. D.; Piperno, A. Practical graph isomorphism, {II}. *J. Symbolic Comput.* **2014**, *60*, 94−112.
(29) Chang, J. L.; Thompson, J. E. Characterization of colored products formed during irradiation of aqueous solutions containing H2O2 and phenolic compounds. *Atmos. Environ.* **2010**, *44*, 541−551.
(30) Sun, Y. L.; 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. *Atmos. Chem. Phys.* **2010**, *10*, 4809.
(31) 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. *Atmos. Environ.* **2019**, *206*, 197−207.
(32) Huang, D. D.; Zhang, Q.; Cheung, H. H. Y.; 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. *Environ. Sci. Technol.* **2018**, *52*, 9215−9224.
(33) 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**, *15*, 1652−1671.
(34) Cordella, L. P.; Foggia, P.; Sansone, C.; Vento, M. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 1367−1372.
(35) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
(36) Open Babel version 2.4.0. http://openbabel.org (accessed March 1, 2020).
(37) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(38) Schirmer, J. Beyond the random-phase approximation: A new approximation scheme for the polarization propagator. *Phys. Rev. A: At., Mol., Opt. Phys.* **1982**, *26*, 2395.
(39) 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.
(40) Klamt, A.; Schüürmann, 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.

(41) 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.
(42) 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.
(43) Thompson, T.; Tapavicza, E. First-Principles Prediction of Wavelength-Dependent Product Quantum Yields. *J. Phys. Chem. Lett.* **2018**, *9*, 4758−4764.
(44) Benkyi, I.; Tapavicza, E.; Fliegl, H.; Sundholm, D. Calculation of vibrationally resolved absorption spectra of acenes and pyrene. *Phys. Chem. Chem. Phys.* **2019**, *21*, 21094−21103.
(45) Grathwol, C. W.; Wössner, N.; Swyter, S.; Smith, A. C.; Tapavicza, E.; Hofstetter, R. K.; Bodtke, A.; Jung, M.; Link, A. Azologization and repurposing of a hetero-stilbene-based kinase inhibitor: towards the design of photoswitchable sirtuin inhibitors. *Beilstein J. Org. Chem.* **2019**, *15*, 2170−2183.
(46) *TURBOMOLE V7.2*; TURBOMOLE GmbH: Karlsruhe, 2017. available from http://www.turbomole.com.
(47) 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.; Hellweg, A.; Helmich-Paris, B.; Holzer, C.; Huniar, U.; Kaupp, M.; Marefat Khah, A.; Karbalaei Khani, S.; Müller, T.; Mack, F.; Nguyen, B. D.; Parker, S. M.; Perlt, E.; Rappoport, D.; Reiter, K.; Roy, S.; Rückert, M.; Schmitz, G.; Sierka, M.; Tapavicza, E.; Tew, D. P.; van Wüllen, C.; Voora, V. K.; Weigend, F.; Wodyński, A.; Yu, J. M. TURBOMOLE: Modular program suite for ab initio quantum-chemical and condensed-matter simulations. *J. Chem. Phys.* **2020**, *152*, 184107.
(48) 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.
(49) Faber, F. A.; Christensen, A. S.; Huang, B.; Von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **2018**, *148*, 241717.
(50) 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.
(51) Häse, F.; Fdez Galván, I.; 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. *Chem. Sci.* **2019**, *10*, 2298−2307.
(52) Westermayr, J.; Marquetand, P. Machine learning for electronically excited states of molecules. *Chem. Rev.* **2020**, DOI: 10.1021/acs.chemrev.0c00749.
(53) Xue, B.-X.; Barbatti, M.; Dral, P. O. Machine Learning for Absorption Cross Sections. *J. Phys. Chem. A* **2020**, *124*, 7199−7210.
(54) Tuna, D.; Lefrancois, D.; Wolański, Ł.; Gozem, S.; Schapiro, I.; Andruniów, T.; Dreuw, A.; Olivucci, M. Assessment of Approximate Coupled-Cluster and Algebraic-Diagrammatic-Construction Methods for Ground-and Excited-State Reaction Paths and the Conical-Intersection Seam of a Retinal-Chromophore Model. *J. Chem. Theory Comput.* **2015**, *11*, 5758−5781.
(55) 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.
(56) 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**. DOI: 10.5281/zenodo.4432153.
(57) 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**. DOI: 10.5281/zenodo.4432606.
(58) von Rudorff, G. F.; Tapavicza, E. *Ferchault/Spectrumscan, Code for Publication*, 2021.10.5281/zenodo.4742333.
(59) Mckay, D. J.; Wright, J. S. How long can you make an oxygen chain? *J. Am. Chem. Soc.* **1998**, *120*, 1003−1013.
(60) Rappoport, Z. *The Chemistry of Peroxides, Parts 1 and 2*; John Wiley & Sons, 2007; Vol. *168*.
(61) 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. *J. Chem. Theory Comput.* **2021**, *17*, 1117−1132.
(62) Cortes, C.; Jackel, L. D.; Solla, S. A.; Vapnik, V.; Denker, J. S. Learning curves: Asymptotic values and rate of convergence. *Adv. Neural Inf. Process. Syst.* **1994**, 327−334.
(63) Meanti, G.; Carratino, L.; Rosasco, L.; Rudi, A. Kernel methods through the roof: handling billions of points efficiently, **2020**. arXiv:2006.10350.
(64) 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.
(65) Hawthorne, S. B.; Miller, D. J.; Barkley, R. M.; Krieger, M. S. Identification of methoxylated phenols as candidate tracers for atmospheric wood smoke pollution. *Environ. Sci. Technol.* **1988**, *22*, 1191−1196.
(66) Hawthorne, S. B.; Miller, D. J.; Langenfeld, J. J.; Krieger, M. S. PM-10 high-volume collection and quantitation of semi-and non-volatile phenols, methoxylated phenols, alkanes, and polycyclic aromatic hydrocarbons from winter urban air and their relationship to wood smoke emissions. *Environ. Sci. Technol.* **1992**, *26*, 2251−2262.
(67) Xu, J.; Cui, T.; Fowler, B.; Fankhauser, A.; Yang, K.; Surratt, J. D.; McNeill, V. F. Aerosol brown carbon from dark reactions of syringol in aqueous aerosol mimics. *ACS Earth Space Chem.* **2018**, *2*, 608−617.
(68) Fleming, L. T.; Lin, P.; Laskin, A.; Laskin, J.; Weltman, R.; Edwards, R. D.; Arora, N. K.; Yadav, A.; Meinardi, S.; Blake, D. R.; Pillarisetti, A.; Smith, K. R.; Nizkorodov, S. A. Molecular composition of particulate matter emissions from dung and brushwood burning household cookstoves in Haryana, India. *Atmos. Chem. Phys.* **2018**, *18*, 2461−2480.
(69) Jiang, W.; Misovich, M. V.; Hettiyadura, A. P.; Laskin, A.; McFall, A. S.; Anastasio, C.; Zhang, Q. Photosensitized Reactions of a Phenolic Carbonyl from Wood Combustion in the Aqueous Phase—Chemical Evolution and Light Absorption Properties of AqSOA. *Environ. Sci. Technol.* **2021**, *55*, 5199.
(70) Forrister, H.; Liu, J.; Scheuer, E.; Dibb, J.; Ziemba, L.; Thornhill, K. L.; Anderson, B.; Diskin, G.; Perring, A. E.; Schwarz, J. P.; Campuzano-Jost, P.; Day, D. A.; Palm, B. B.; Jimenez, J. L.; Nenes, A.; Weber, R. J. Evolution of brown carbon in wildfire plumes. *Geophys. Res. Lett.* **2015**, *42*, 4623−4630.