JGR Atmospheres

RESEARCH ARTICLE
10.1029/2023JD039459

Key Points:

- Appropriate synthetic minority over-sampling based few-shot learning could decipher the light-absorbing molecules from a small data set
- $C_4H_6O_4NS$, $C_8H_6O_4NS$, $C_{11}H_{15}O_3N_2$, $C_{12}H_{15}O_3N_2$, and $C_{19}H_{21}O_6$ dominate the Abs_{365} of $HULIS_{WS}$ in Nanjing
- Nitrogen-containing functional groups dominate the light-absorbing ability of $HULIS_{WS}$ at a micro-level

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Y.-L. Zhang,
zhangyanlin@nuist.edu.cn;
dryanlinzhang@outlook.com

Citation:

Hong, Y., Zhang, Y.-L., Bao, M., Fan, M.-Y., Lin, Y.-C., Xu, R., et al. (2023). Nitrogen-containing functional groups dominate the molecular absorption of water-soluble humic-like substances in air from Nanjing, China revealed by the machine learning combined FT-ICR-MS technique. *Journal of Geophysical Research: Atmospheres*, *128*, e2023JD039459. https://doi.org/10.1029/2023JD039459

Received 13 JUN 2023
Accepted 19 NOV 2023

Author Contributions:

Conceptualization: Yan-Lin Zhang
Funding acquisition: Yan-Lin Zhang, Fang Cao
Methodology: Yihang Hong, Mengying Bao, Zhiyang Shu, Ji-Yan Wu, Hongxing Jiang, Zhineng Cheng, Jun Li, Gan Zhang
Software: Yihang Hong
Supervision: Yan-Lin Zhang, Jun Li, Gan Zhang
Validation: Zhiyang Shu
Visualization: Yihang Hong
Writing – original draft: Yihang Hong
Writing – review & editing: Yan-Lin Zhang, Mei-Yi Fan, Yu-Chi Lin, Rongshuang Xu, Fang Cao

Nitrogen-Containing Functional Groups Dominate the Molecular Absorption of Water-Soluble Humic-Like Substances in Air From Nanjing, China Revealed by the Machine Learning Combined FT-ICR-MS Technique

Yihang Hong[1,2], Yan-Lin Zhang[1,3], Mengying Bao[4], Mei-Yi Fan[1,3,5], Yu-Chi Lin[1,3], Rongshuang Xu[1,3], Zhiyang Shu[6], Ji-Yan Wu[1,3], Fang Cao[1,3], Hongxing Jiang[7,8], Zhineng Cheng[7,8], Jun Li[7,8], and Gan Zhang[7,8]

[1]Atmospheric Environment Center, Joint Laboratory for International Cooperation on Climate and Environmental Change, Ministry of Education (ILCEC), Nanjing University of Information Science and Technology, Nanjing, China, [2]School of Politics, Economics and International Relations, University of Reading, Reading, UK, [3]School of Ecology and Applied Meteorology, Nanjing University of Information Science and Technology, Nanjing, China, [4]Huzhou Meteorological Administration, Huzhou, China, [5]Air Quality Studies, Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong, China, [6]Morrissey College of Arts and Sciences, Boston College, Boston, MA, USA, [7]State Key Laboratory of Organic Geochemistry, Guangdong Province Key Laboratory of Environmental Protection and Resources Utilization, Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou, China, [8]CAS Center for Excellence in Deep Earth Science, Guangzhou, China

Abstract The light absorption capacity of water-soluble humic-like substances ($HULIS_{WS}$) at the molecular level is crucial for reducing the uncertainties in modeling the radiative forcing. This study proposed a machine learning approach to allocate the light absorption coefficient at 365 nm (Abs_{365}) of $HULIS_{WS}$ into 8084 Fourier transform-ion cyclotron resonance mass spectrometry (FT-ICR-MS) detached molecular markers and their potential functional groups. The ML model showed an acceptable uncertainty (<5%) to the whole Abs_{365} value based on the prediction errors. The results showed that five critical light-absorbing molecules ($C_4H_6O_4NS$, $C_8H_6O_4NS$, $C_{11}H_{15}O_3N_2$, $C_{12}H_{15}O_3N_2$, and $C_{19}H_{21}O_6$) could explain 74% (±3%) of the variation of Abs_{365} in the winter, whereas no crucial light-absorbing molecules were found in the summer. Besides, the nitrogen-containing functional groups were found to dominate (61% ± 8%) the molecular absorption near the 365 nm of the spectrum. This work illustrated how functional groups affect the absorption of $HULIS_{WS}$, providing critical information for future research of $HULIS_{WS}$ on the molecular level.

Plain Language Summary Water-soluble humic-like substances are a crucial light-absorbing component of fine particulate matter in China. Understanding the crucial light-absorbing chromophores of the light-absorbing mixtures was helpful in controlling the global warming event. In this study, CHON and CHONS species were found to dominate the light absorption coefficient of water-soluble humic-like substances in Nanjing by combining the machine learning algorithm with the FT-ICR-MS technique. Then, the dominant light-absorbing molecules were further combined with the UV-Vis spectra from the chemical database to further estimate the light-absorbing functional groups. Our results indicated that nitrogen-containing functional groups dominate the light-absorbing ability of water-soluble humic-like substances.

1. Introduction

Water-soluble humic-like substances ($HULIS_{WS}$) have been widely detected in clouds, fog, rainwater, and snow (Bao et al., 2022; Fan et al., 2018; Lin et al., 2010; Santos et al., 2012), which can be emitted by primary sources and formed from secondary processes (Arya et al., 2020; Bao et al., 2022; Fan, Song, & Peng, 2016; Li et al., 2019; Zhang et al., 2021). As an essential fraction of atmospheric brown carbon (BrC), $HULIS_{WS}$ possess significant light-absorbing abilities at the ultraviolet to visible (UV-Vis) wave band, contributing 20%–40% of the direct radiative forcing to the atmosphere (Chung et al., 2012; Wang et al., 2018; Zhang et al., 2017, 2020). The light-absorbing abilities of $HULIS_{WS}$ are predominantly governed by their molecular sub-structures (Fan et al., 2018), which depend on their atmospheric processes and emission sources (Baduel et al., 2010; Fan, Wei, et al., 2016). Many previous studies have investigated the light absorption coefficient (Abs) of individual species

(Bao et al., 2022; Guo et al., 2022; Jiang, Li, Sun, Liu, et al., 2021). However, the dominant chromophores of $HULIS_{WS}$ at the micro level are still unknown since they are composed of a variety of chemical components (Laskin et al., 2018; Noziere et al., 2015).

The Fourier transform-ion cyclotron resonance mass spectrometry (FT-ICR-MS) coupled with electrospray ionization (ESI) ion source [ESI-FT-ICR-MS] is a feasible method for the detection of elements in multi-component mixtures, which can provide high-accuracy molecular information (structural markers) and associated relative abundance (molecular intensity) of mixed aerosols (Chen et al., 2016; Jiang, Li, Sun, Tian, et al., 2021; Jiang et al., 2020; Lin et al., 2012; Wang, Hu, Lin, et al., 2019; Zeng et al., 2020). For instance, Jiang, Li, Sun, Tian, et al. (2021) quantified the molecular compositions of dissolved organic aerosols using the ESI-FT-ICR-MS technique. Zeng et al. (2020) used ESI-FT-ICR-MS to quantify the molecular compositions of BrC, and further allocated the Abs of the mixtures into the individual molecular formula with the aid of the statistic method (Zeng et al., 2020). However, the mix of complex atmospheric aerosols and the interaction of their chromophores could cause some synergistic or antagonistic effects on the light-absorbing abilities (Li et al., 2022), leading to a more complex non-linear relationship between molecular composition and Abs value. Such a complex relationship might limit the application of traditional statistical methods in deciphering high-dimensional relationships.

Machine learning (ML) is an emerging tool that could build complex non-linear relationships between input and output variables, having been successfully used to reveal the complex relationships between mixtures and their molecular intensities (Jiang, Li, Sun, Tian, et al., 2021; Tapavicza et al., 2021). Although ML can easily fit the non-linear relationships between targets and variables, overfitting may occur when the training data set is insufficient. Based on this, some over-sampling techniques like the synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002) and other ML-based data augment methods (Mumuni & Mumuni, 2022) were provided to ensure good performance.

This work uses the ML approach combined with the SMOTE technique (described in Text S10 of the Supporting Information S1) to allocate the Abs values of $HULIS_{WS}$ mixtures to each molecule level by the molecular marker using the ESI-FT-ICR-MS technique (Hong, Cao, Fan, Lin, Bao, et al., 2022; Jiang, Li, Tang, Cui, et al., 2022). Further, the molecular absorption was allocated to the functional groups by combining the molecular absorption data collected from the chemical database—Reaxys (Goodman, 2009). This ML approach enables to select the light-absorbing molecules and functional groups, which could be used to allocate the Abs value at the macro-level to the micro-level when the chemical structures of $HULIS_{WS}$ and the interaction between each chromophore (Boyle et al., 2009; Yakimov et al., 2022), water-soluble metallic ions (Li et al., 2022), and solutions (Phillips et al., 2017) are not perfectly identified.

2. Methods and Data Analysis

2.1. Sample Collection and Chemical Analysis

$PM_{2.5}$ samples were collected at an open urban site in Nanjing, China, using a high-volume air sampler (flow rate: 1 m^3 min^{-1}, see details in Text S1 of the Supporting Information S1). As shown in the flow chart (Figure 1), $HULIS_{WS}$ fractions were extracted from atmospheric samples and determined, following the protocol reported by Fan et al. (2018) using hydrophile-lipophile balance (HLB) resin (Text S2 in Supporting Information S1). According to the concentration levels of $HULIS_{WS}$, 12 samples (with 2 high, 2 moderate, and 2 low concentrations of $HULIS_{WS}$ from winter and summer, Figure S1 and Table S1 in Supporting Information S1) were selected (Bao et al., 2023) from the whole samples for light absorption (Bao et al., 2022) (Text S3 in Supporting Information S1) and ESI-FT-ICR-MS analyses (Bao et al., 2023; Jiang, Li, Sun, Tian, et al., 2021) (Texts S4 in Supporting Information S1). In addition, concentrations of water-soluble ions in $PM_{2.5}$ samples were also determined following the procedure described elsewhere (Bao et al., 2022).

During the sampling period, the gas pollutant data were obtained from the China National Environmental Monitoring Centre (www.cnemc.cn) while the meteorological parameters were acquired from the China Meteorological Bureau (www.cma.gov.cn). In addition, fire-spot maps were obtained from the Fire Information for Resource Management System (firms.modaps.eosdis.nasa.gov) (Figure S2 in Supporting Information S1). Aerosol liquid water content (ALWC) and pH value were calculated by the ISORROPIA-II model (Text S5 in Supporting Information S1) (Lin et al., 2020).

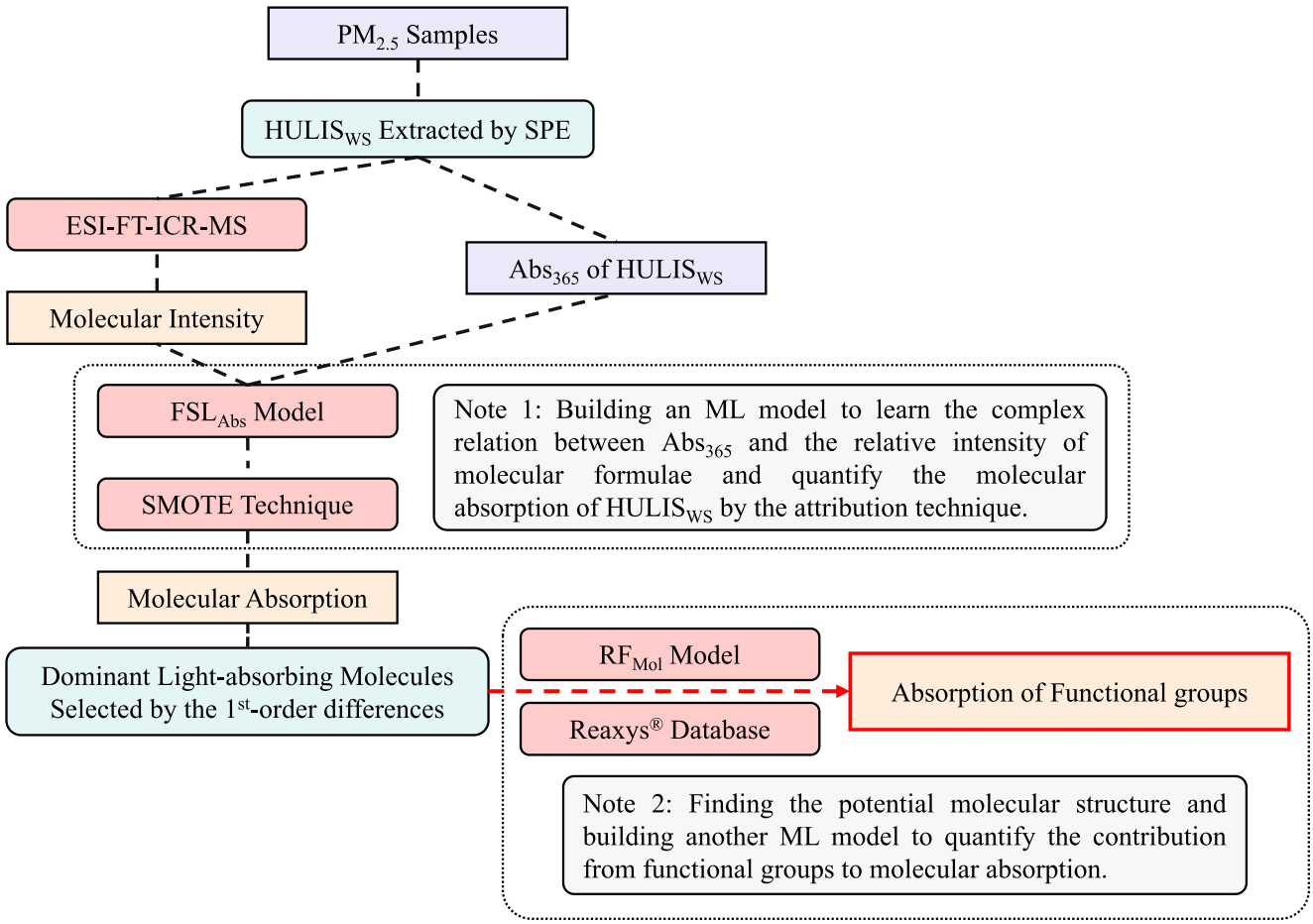

Figure 1. The flow chart of this work. The red squares represent the important technique used in this work; the cyan squares represent the important intermediate output of the research flow and their separated methods; the orange squares represent the important results on the molecular level; the purple squares represent the mixed samples.

2.2. Light Absorption From Individual Molecules to Bulk

For the molecular formula j detached from sample i using ESI-FT-ICR-MS, the molecular intensity (I_{ij}) equals its concentration (c_{ij}) multiplied by the ionization efficiency (γ_j) ($I_{ij} = c_{ij} \cdot \gamma_j$). Combining with the Beer-Lambert law, the light absorption coefficient (Abs_i) of sample i could be calculated by I_{ij} as follows (Zeng et al., 2020):

$$\mathrm{Abs}_i = \sum_{i=0}^{n} \left(I_{ij} \times \frac{\varepsilon_j}{\gamma_j} \right) + \mathrm{Interactions}, \quad (1)$$

where ε_j represents the Abs value of each molecular formula and the Interactions represent the intermolecular effects on the Abs value. When the ML model fully learns the relationship of Equation 1, the relative contribution calculated by the ML-based attribution technique should equal $k \cdot \varepsilon_j / \gamma_j$ (k is the proportional coefficient). Then, the molecular light absorption could be calculated (see details in Text S6 of the Supporting Information S1).

2.3. Machine Learning Model

As the workflow shown in Figure 1, then, an approach using the few-shot learning (FSL) method ($\mathrm{FSL}_{\mathrm{Abs}}$) was built to decipher the relationship from Abs value to the relative intensity of molecular markers in 12 samples determined by the ESI-FT-ICR-MS (see Text S6 in Supporting Information S1) (Wright & Ziegler, 2017). FSL is an algorithm of the ML that is aimed to learn the underlying pattern from a few samples (Parnami & Lee, 2022), and has been widely used in previous object detection (Kisantal et al., 2019), cheminformatics (Chen et al., 2023), and environmental studies (Huang et al., 2023). Here, we used the synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002) combined with the random forest (RF) algorithm provided by Ranger

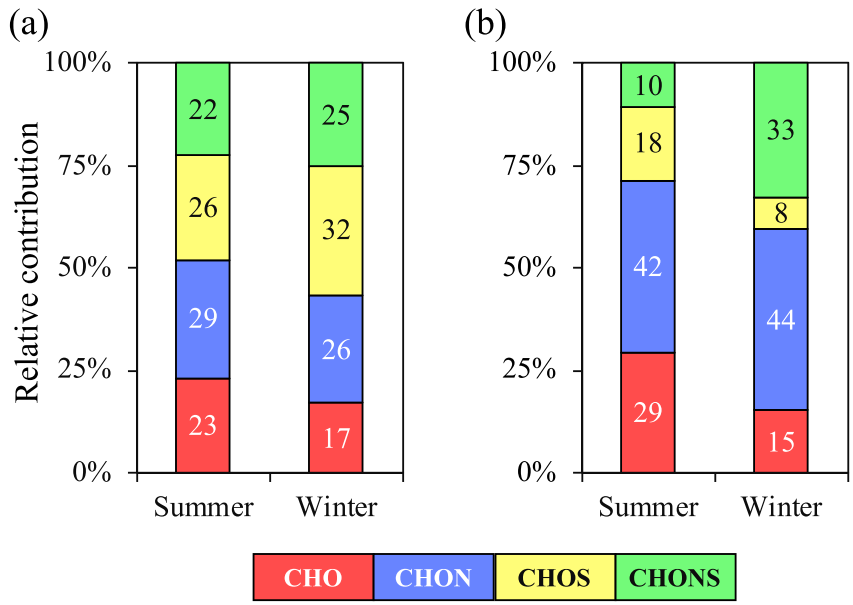

Figure 2. The contribution of CHO, CHON, CHOS, and CHONS subgroups to the molecular intensity (a) and Abs_{365} (b) of the $HULIS_{WS}$.

package (Fan et al., 2023; Hong, Cao, Fan, Lin, Bao, et al., 2022; Wright & Ziegler, 2017) to successfully build an FSL model without the risk of overfitting (has been proved in a 55 × 35799 ESI-FT-ICR-MS data set, see details in Text S8 of the Supporting Information S1) (Belgiu & Drăguţ, 2016; Cortes-Ciriano & Bender, 2015; Jablonka et al., 2020), and then proved by the validation data set and model outputs (Text S9 in Supporting Information S1) (Arulkumaran et al., 2017).

After that, the Abs_{365} of $HULIS_{WS}$ (target variable) were allocated to each molecular marker (input variables) using the attribution technique combined with the FSL_{Abs} (Davies et al., 2021; Hong, Cao, Fan, Lin, Bao, et al., 2022; Wei et al., 2015), and the main light-absorbing molecules were selected by the first-order difference of molecular absorption among each molecule (Text S7 in Supporting Information S1).

Meanwhile, the molecular information (i.e., UV-Vis spectra, molecular structures, and functional groups) of 5 selected light-absorbing molecular markers were collected from Reaxys (www.reaxys.com) (Goodman, 2009). After that, as the flow showed in Figure 1, an RF model (RF_{Mol}) was built to learn the relationship between functional groups and the molecular absorption near 365 nm using the 296 selected chemical spectra and structures (Davies et al., 2021; Hong, Cao, Fan, Lin, Bao, et al., 2022; Wei et al., 2015), and validated by other 827 light-absorbing molecules. Here, other influencing factors like the number of each element (i.e., C, H, O, N, and S), the solution (i.e., water, ethanol, etc.), and some UV-Vis spectra parameters were also considered (see Texts S11 and S12 in Supporting Information S1) and the relative contribution of each functional group was estimated by the attribution technique (Davies et al., 2021; Hong, Cao, Fan, Lin, Bao, et al., 2022; Wei et al., 2015).

3. Results and Discussion

3.1. Molecular Composition of $HULIS_{WS}$

The ESI-FT-ICR-MS analysis detected 8084 unique molecular formulas from $HULIS_{WS}$ fractions in total (Figures S3 and S4 in Supporting Information S1), and the number of chemical formulas ranged from 1,334 to 4,939 (Table S1 in Supporting Information S1) for each extracted solution. In general, a total of 6,121 molecules, which have molecular absorption at 365 nm, were determined by the FSL_{Abs} model with a consideration of chromophores' interaction in solution. After that, the contributions of four subgroups to the molecular absorptions were apportioned using elemental composition (Zeng et al., 2020) (see Figure 2): CHO (molecules containing only C, H, and O), CHON (only C, H, O, and N), CHOS (only C, H, O, and S), and CHONS (only C, H, O, N, and S). The result showed that the relative intensity of CHON components dominated the Abs_{365} value in both summer and winter (42%–44%) (All contributions to mass concentration and Abs_{365} came from the relative intensity of subgroups). However, the contributions

of other subgroups to Abs_{365} varied in the summer and winter. For example, the contributions of CHO and CHOS to Abs_{365} in the summer (CHO∼29% and CHOS∼18%) were much higher than those in the winter. On the contrary, the contribution of CHONS to ABS_{365} in the winter was 33%, exceeding that (10%) in the summer by a factor of 3.

In the summer, the higher temperature (29.4 ± 1.9°C) and relative humidity (81.5% ± 6.0%) might promote the chemical reaction, leading to a photo-bleaching effect (Jiang, Li, et al., 2020; Lin et al., 2017; Song et al., 2018). In the atmosphere, NO_x ($NO_x = NO + NO_2$) can be converted into nitrate (NO_3^-) and organic nitrate (light-absorbing fractions) upon atmospheric oxidation (Zhang, Zhang, et al., 2022), and organic nitrate may further allocate into some smaller N-containing compounds (e.g., isocyanic acid) (Hems & Abbatt, 2018). Here, the higher contribution (42%) of CHON species (carbon number > 3) and lower nitrate (NO_3^-) concentration (3.2 μg m^{-3}) in the summer indicated that more organic nitrates were formed through atmospheric oxidation. Such phenomena could be attributed to the favorable occurrence of liquid-liquid phase separation (LLPS) of summer aerosols, owing to their comparatively higher pH value of 4.2 and lower ALWC of 13.8 ± 8.8 μg m^{-3} in summer relative to those values in winter (pH of 3.6 and ALWC of 166.2 μg m^{-3}) (Dallemagne et al., 2016; Schmedding et al., 2020). Since phase-separated aerosols typically exhibit an inorganic-rich core and an organic-rich shell with higher viscosity, the entry of gas species (i.e., NO_x) could be hindered, thereby limiting the formation of secondary inorganic ions (NO_3^-: 3.2 ± 2.9 μg m^{-3}, SO_4^{2-}: 7.6 ± 2.6 μg m^{-3}) (Schmedding et al., 2020). In addition, the low oxidation extent of aerosols found in this work, where 84% of molecular formulas have an O/C ratio lower than 0.8 (Figure S5 in Supporting Information S1), provides further evidence for the occurrence of LLPS (Kucinski et al., 2021).

Compared to the O/C value, the (O-3S-2N)/C ratio is a better index to express the number of oxygen-containing functional groups of each carbon atom since each organo-sulfate/nitrate usually contains three/two more oxygen atoms than the common CHO species.[48] Previously, the distribution of the (O-3S-2N)/C combined with the H/C value could be used to roughly distinguish the chemical composition (Jiang, Li, Tang, Cui, et al., 2022). As shown in Figure 2, the molecular compositions of $HULIS_{WS}$ were separated into six different groups using the Van-Krevelen diagrams (see details in Table S10 of the Supporting Information S1) (Jiang, Li, Sun, Tian, et al., 2021). Generally, highly unsaturated molecules dominated both molecular intensity (44.5%) and Abs_{365} (39.0%) of $HULIS_{WS}$ during the whole period, which could be attributed to the biomass burning (see Figure S2 in Supporting Information S1) and cooking (mainly the used of fuel) emissions (Jiang, Li, Tang, Cui, et al., 2022).

To further discuss the potential formation of organic nitrate, the molecular formulas with an O/N ratio equal to 6 were selected as organic nitrate in the further discussion (Jiang, Li, Tang, Zhao, et al., 2022). Compared with summer, organic nitrate has lower O/C (0.77) and H/C (1.45) ratios based on the relative intensity weighted (Table S9 in Supporting Information S1). Winter organic nitrate showed a higher relative intensity weighted double-bond equivalent (DBE) and aromaticity index (AI) compared with spring (Table S9 in Supporting Information S1).

Here, the N-containing compounds (CHON and CHONS) contributed large proportions to light-absorbing ability (55% in the summer and 77% in the winter), highlighting the existence of nitro-aromatic compounds (NACs) (Wang, Hu, Wang, et al., 2019). Besides, the high observed RH (77.9% ± 12.8%) and NO_x concentration (109.8 ± 18.0 μg m^{-3}), indicate that the secondary formations of NO_x might be a formation pathway of $HULIS_{WS}$ (Wang, Hu, Wang, et al., 2019).

Previous studies have documented that NO_x participated in atmospheric chemical reactions and produced light-absorbing N-containing compounds (Chow et al., 2016; Li et al., 2020). Based on the Van-Krevelen diagrams divided results in Figure 2 (Jiang, Li, Tang, Cui, et al., 2022), the organic nitrate in winter was mainly composed of polyphenolic aromatics, which formation was significantly affected by the NO_2 concentration level (Wang, Hu, Wang, et al., 2019). Besides, the rising SO_2 and SO_4^{2-} abundance could also contribute to the formation of organic nitrate at nighttime (Chen et al., 2022). Here, $HULIS_{WS}$ showed a well-positive correlation with NO_3^- ($R^2 = 0.81$, $p < 0.01$), NO_2 ($R^2 = 0.76$, $p < 0.01$), SO_2 ($R^2 = 0.57$, $p < 0.01$), and SO_4^{2-} ($R^2 = 0.96$, $p < 0.01$) in the winter (Figures S6 and S7 in Supporting Information S1), highlighting the NO_2 and SO_2 joint formation of organic nitrate in winter (Chen et al., 2022).

3.2. Possible Molecular Formulas of the Dominant Light-Absorbing Molecules of $HULIS_{WS}$

Here, 5 crucial molecular formulas ($C_4H_6O_4NS$, $C_8H_6O_4NS$, $C_{11}H_{15}O_3N_2$, $C_{12}H_{15}O_3N_2$, and $C_{19}H_{21}O_6$) selected from 6,121 molecules by the first-order differences of molecular absorption (Text S8 in Supporting Information S1) in the winter contributed ∼74% (±3%) to the Abs_{365}, whereas no significant light-absorbing molecules were detected in the summer (all molecules contributed <1% to Abs_{365}). Such results could be attributed to the higher emissions from

biomass burning and anthropogenic (e.g., coal combustion) activities that happened in winter (Bao et al., 2023). Such events could emit more highly unsaturated pollutants (i.e., $C_4H_6O_4NS$, $C_8H_6O_4NS$, $C_{11}H_{15}O_3N_2$, $C_{12}H_{15}O_3N_2$, and $C_{19}H_{21}O_6$), contributing more light absorbing abilities in 365 nm (Text S9 in Supporting Information S1). In summer, however, the higher biogenic events emitted and the secondary reactions produced more saturated pollutants under the higher temperature and RH conditions, which have limited light-absorbing abilities (Bao et al., 2023). Besides, compared to the CHO (0.45), CHON (0.49), and CHONS (0.26) species, the lower (0.08) relative intensity weighted aromaticity index (AI_W) of CHOS subgroups in winter indicated that the CHOS species mainly comprised of aliphatic species, which contributed lower or no light-absorbing abilities near 365 nm (Bai et al., 2023; Jiang, Li, Sun, Tian, et al., 2021). Such results could be used to explain why no CHOS species were selected.

To further prove the dominant species could contribute larger light-absorbing abilities at the micro level, a total of 296 existing isomers of $C_8H_6O_4NS$, $C_{11}H_{15}O_3N_2$, $C_{12}H_{15}O_3N_2$, and $C_{19}H_{21}O_6$ (see details in Table S3 of the Supporting Information S1) were collected from previous literature using the Reaxys database. Among these four light-absorbing molecules, $C_8H_6O_4NS$, $C_{11}H_{15}O_3N_2$, and $C_{12}H_{15}O_3N_2$ possessed lower (O-3S-2N)/C ratios, indicating that these molecules mainly originated from the less oxidation state sulfur-containing molecules (Figure 3). On the contrary, $C_{19}H_{21}O_6$ exhibited a higher (O-3S-2N)/C ratio, reflecting that this molecule was characterized by highly unsaturated structures (Jiang, Li, Tang, Cui, et al., 2022). Using the Reaxys database, the possible isomer of these 4 critical light-absorbing molecules were identified as methyl trans-β-(5-nitro-2-thienyl)acrylate ($C_8H_6O_4NS$), 2-(2-Phenoxy-ethylazo)-prop-2-yl-hydroperoxide ($C_{11}H_{15}O_3N_2$), syn-(1-hydroxyethyl,methyl)(ethyl,methyl)bimane ($C_{12}H_{15}O_3N_2$), and 6-(3-acetoxypropyl)-8-tert-butyl-2-oxo-2H-chromen-3-carboxylic acid ($C_{19}H_{21}O_6$), respectively. No spectra of $C_4H_6O_4NS$ have been identified till now and therefore we attempted to reconstruct the spectra of $C_4H_6O_4NS$ by the ML approach (Text S10 in Supporting Information S1). The result showed that the isomers of $C_4H_6O_4NS$ might be 2-cyanoethylmethanesulfonate, 2-cyanopropane-1-sulfonate, etc., which their light-absorbing bands ranging from 230 to 497 nm with the major absorbed peak of 287–409 nm (Table S4 of the Supporting Information S1). Such an ML model has been validated through another 873 compounds within the light-absorbing molecules (Figure S11 in Supporting Information S1).

3.3. The Potential Light-Absorbing Chromophores of $HULIS_{WS}$

Section 3.2 discussed the possible structure of the main light-absorbing molecules. However, since ESI-FT-ICR-MS only provided the mass-to-charge ratio of molecular clusters derived from isobarically resolved ions, the molecules obtained from Reaxys might not reflect the real chemical composition in the atmosphere. Functional groups, as the composed unit of molecular absorption (Higashiguchi et al., 2005; Zhao et al., 2021), could be used to reflect the composition of light-absorbing chromophores. In this research, 18 functional groups that have been reported in previous atmospheric studies, such as C-N bound (mainly from vehicles) and -CO-NH- (primarily from biomass burnings, Table S5 in Supporting Information S1) were considered in the RF_{Mol} model (Table S5 in Supporting Information S1) (Cheng et al., 2006). Except for the kind and position of the functional group (Higashiguchi et al., 2005; Zhao et al., 2021), some previous research reported that the molecular absorption spectra could also be affected by the properties and pH value of solutions (Phillips et al., 2017), the water-soluble metallic ions (e.g., Fe(III)) (Li et al., 2022), as well as the interaction of chromophores inside the mixture components in solution (Boyle et al., 2009; Yakimov et al., 2022).

To quantify the potential contribution of functional groups to the Abs_{365} value, another ML approach (RF_{Mol}) was built based on two important prediction variables, including the number of functional groups and the kind of solutions. This model was trained by 296 compounds of the 5 main light-absorbing molecules selected by the FSL_{Abs} in the winter and validated by other 873 molecular absorption spectra from Reaxys (Text S11 in Supporting Information S1). Note that although the position of substitution and water-soluble metallic ions was not considered here, the RF_{Mol} model could be used to predict molecular absorption beyond the training molecules with an acceptable error ($R^2 = 0.77$, MAE = 0.27, RMSE = 0.75). Therefore, even for the mixture of species with the same elements in the detached ions, this model could also predict its molecular absorption under an acceptable error.

Then, the contribution of functional groups to the molecular absorption coefficient in the winter could be estimated using the ES3, and the results are shown in Figure 4. Our results indicated that the –C≡N, –CO–NH–, and phenyl radical (C_6H_5–) dominated the molecular absorption coefficient near the 365 nm of the spectrum (Figure 4), with average contributions of 31% (±4%), 30% (±4%), and 12% (±5%), respectively.

In all, the CHON and CHONS components, which mainly contained –C≡N and –CO–NH–groups, dominated the light-absorbing abilities of $HULIS_{WS}$ near 365 nm in the winter (44%). Moreover, nitrogen-containing

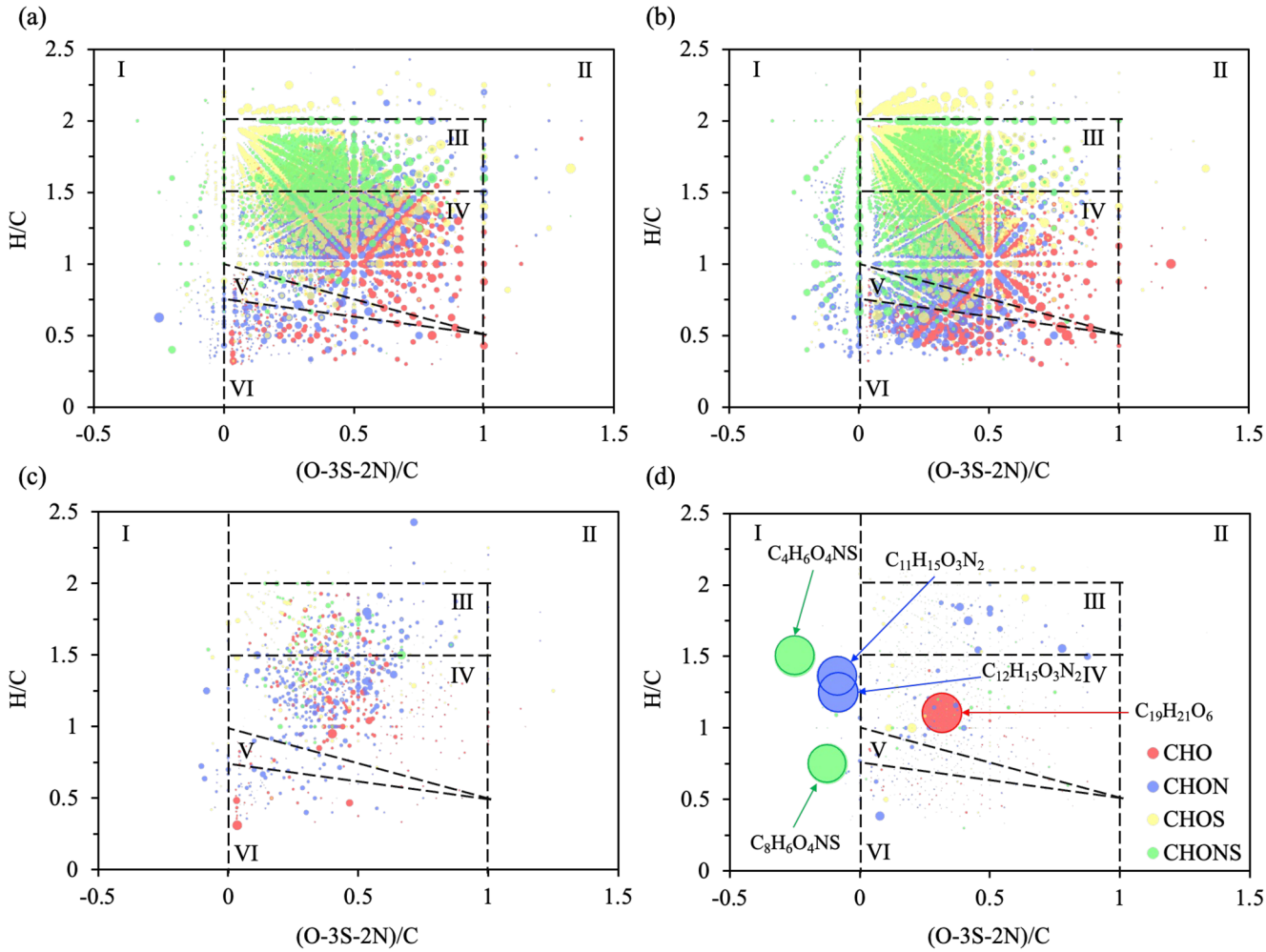

Figure 3. The Van-Krevelen diagrams of molecules detected by ESI-FT-ICR-MS analysis. The size of the circle represents the molecular intensity in the summer (a) and winter (b) and light absorption (Abs_{365}) values in summer (c) and winter (d), respectively. In subfigure (d), the molecular formula with a higher contribution to Abs_{365} is labeled with its molecular formula. Lines separating compound categories here were for better visualization, and accurate categorization could be found elsewhere (Jiang, Li, Tang, Cui, et al., 2022). The categories I, II, III, IV, V, and VI represent the less oxidation state sulfur-containing, saturated or highly oxidized, unsaturated aliphatic, highly unsaturated, polyphenolic aromatics, and polycyclic aromatics molecules, respectively (Jiang, Li, Tang, Cui, et al., 2022). After removing the point with 0 molecular intensity and absorption, the number of exhibited chemical formulas in the figure was (a) n = 5,622 and (c) n = 2,250 in summer, (b) n = 6,121 and (d) n = 2,085 in winter, respectively.

light-absorbing functional groups (i.e., –C≡N and –CO–NH–) contributed 61% (±8%) to the Abs in the winter, while the oxygen-containing functional groups also contributed ~27% (±5%) to the Abs. As the most light-absorbing group, the nitrogen-containing unsaturated bond was mainly contained by the aerosols emitted from fossil fuel combustion (including both vehicle (Lyu et al., 2019; Zhang, Ren, et al., 2022) and cooking (Lyu et al., 2019; Wang et al., 2020)). –CO–NH– was profoundly emitted from cooking and biomass burning (Cheng et al., 2006). The oxygen-containing functional groups widely came from the photochemical formation (Chhantyal-Pun et al., 2018; Frossard et al., 2014; Millet et al., 2015), biomass and fossil fuel combustion (Chhantyal-Pun et al., 2018; Humes et al., 2022; Millet et al., 2015), and marine sources (Frossard et al., 2014).

4. Conclusion

In this study, we proposed a few-shot learning approach that can successfully allocate the light absorption of $HULIS_{WS}$ into its functional groups. Our results indicated that the nitrogen-containing functional groups were a crucial contributor to $HULIS_{WS}$ light absorption. This approach considers all the existing molecules and their functional groups which may influence the light absorption of $HULIS_{WS}$ from chemical databases and may be helpful to allocate other aerosol properties (e.g., toxicology and pathology) into the molecular or even functional

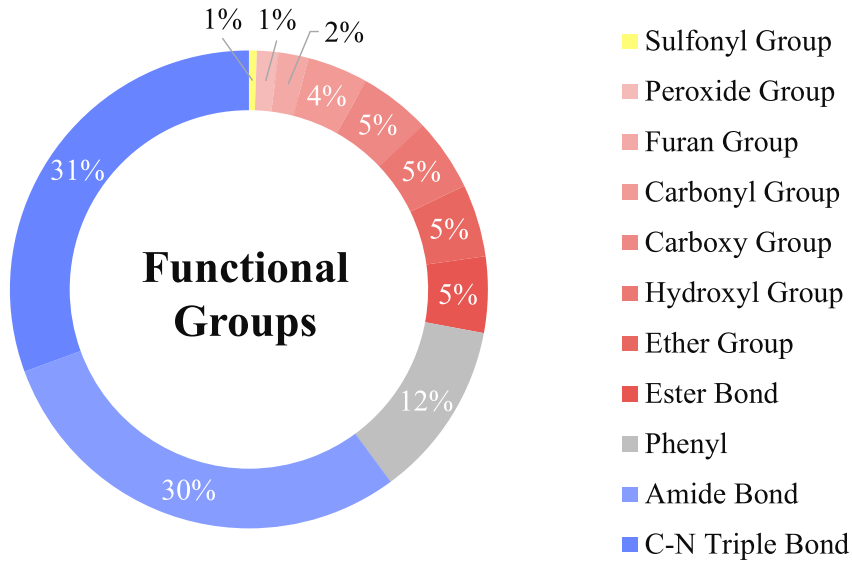

Figure 4. The results of relative contribution derived from certain functional groups in the modeling work of the light absorption coefficient (Abs) by using the RF_{Mol} model (see details in Text S11 of the Supporting Information S1) with a permutation mode in winter. Here, CH represents the functional groups only containing C and H elements; O represents the functional groups containing O but without N and S; N represents the functional groups containing N; S represents the functional groups containing S. The functional groups contributed lower than 0.1% were not displayed in this figure.

group levels. However, since some molecular structures have not been determined or listed in chemical databases, the bias from isomerism was hard to remove. Although this is the most likely estimated result in mathematics with 95% confidence, we do not rule out that other light-absorbing molecules may also contribute to the light-absorbing results in the real atmosphere. In the future, more chemical structures and their UV-Vis spectra were recommended to consider in this model to improve the accuracy of our results.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

Data: The observation data of ESI-FT-ICR-MS and the collected UV-Vis spectra data from the chemical database in this work are available in Tables S8 and S9 of the Supporting Information S1 which can be found via the link https://doi.org/10.17605/OSF.IO/BKX9H (Hong et al., 2023). *Data*: The observation data of ESI-FT-ICR-MS and Abs_{365} value of dissolved organic matters (DOM) used in this work was collected from Jiang, Li, et al. (2020). *Software*: The random forest algorithm used in this research was built by the ranger model (version 0.15.1), which is available for public use at https://cran.r-project.org/web/packages/ranger/index.html (Wright et al., 2023). *Software*: The support vector regression (SVR) model was built by the e1071 model (version 1.7-13), which is available for public use at https://cran.r-project.org/web/packages/e1071/index.html (Meyer et al., 2023). *Software*: The partial least squares regression (PLSR) was built by pls model (version 2.8-2), which is available for public use at https://cran.r-project.org/web/packages/pls/index.html (Liland et al., 2023). *Software*: The artificial neural network (ANN) was built by neuralnet model (version 1.44.2), which is available for public use at https://cran.r-project.org/web/packages/neuralnet/index.html (Fritsch et al., 2019). *Software*: The aerosol liquid water content (ALWC) was calculated using the ISORROPIA-II model, which can be obtained at http://wiki.seas.harvard.edu/geos-chem/index.php?title=ISORROPIA_II (Fountoukis & Nenes, 2007).

Acknowledgments
This research was financially supported by the National Natural Science Foundation of China (Grants 42192512 and 41977305). The license for the Reaxys database came from the University of Reading Library, United Kingdom.

References

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, *34*(6), 26–38. https://doi.org/10.1109/MSP.2017.2743240

Arya, M., Supriya, D., Archita, R., Shiguo, J., Supratim, B., & Sayantan, S. (2020). Sources and atmospheric processing of brown carbon and HULIS in the Indo-Gangetic Plain: Insights from compositional analysis. *Environmental Pollution*, *267*, 115440. https://doi.org/10.1016/j.envpol.2020.115440

Baduel, C., Voisin, D., & Jaffrezo, J. L. (2010). Seasonal variations of concentrations and optical properties of water soluble HULIS collected in urban environments. *Atmospheric Chemistry and Physics*, *10*(9), 4085–4095. https://doi.org/10.5194/acp-10-4085-2010

Bai, Z., Wen, W., Zhang, W., Li, L., Wang, L., & Chen, J. (2023). The light absorbing and molecule characteristic of PM2.5 brown carbon observed in urban Shanghai. *Environmental Pollution*, *318*, 120874. https://doi.org/10.1016/j.envpol.2022.120874

Bao, M., Zhang, Y. L., Cao, F., Hong, Y., Lin, Y. C., Yu, M., et al. (2023). Impact of fossil and non-fossil fuel sources on the molecular compositions of water-soluble humic-like substances in $PM_{2.5}$ at a suburban site of Yangtze River Delta, China. *Atmospheric Chemistry and Physics*, *23*(14), 8305–8324. https://doi.org/10.5194/acp-23-8305-2023

Bao, M., Zhang, Y. L., Cao, F., Lin, Y. C., Hong, Y., Fan, M., et al. (2022). Light absorption and source apportionment of water soluble humic-like substances (HULIS) in $PM_{2.5}$ at Nanjing, China. *Environmental Research*, *206*, 112554. https://doi.org/10.1016/j.envres.2021.112554

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Boyle, E. S., Guerriero, N., Thiallet, A., Vecchio, R. D., & Blough, N. V. (2009). Optical properties of humic substances and CDOM: Relation to structure. *Environmental Science & Technology*, *43*(7), 2262–2268. https://doi.org/10.1021/es803264g

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.5555/1622407.1622416

Chen, J., Wang, X., Zhang, J., Li, M., Li, H., Liu, Z., et al. (2022). Particulate organic nitrates at Mount Tai in winter and spring: Variation characteristics and effects of mountain-valley breezes and elevated emission sources. *Environmental Research*, *212*, 113182. https://doi.org/10.1016/j.envres.2022.113182

Chen, K., Chen, G., Li, J., Huang, Y., Wang, E., Hou, T., & Heng, P.-A. (2023). MetaRF: Attention-based random forest for reaction yield prediction with a few trails. *Journal of Cheminformatics*, *15*(1), 43. https://doi.org/10.1186/s13321-023-00715-x

Chen, Q., Ikemori, F., Higo, H., Asakawa, D., & Mochida, M. (2016). Chemical structural characteristics of HULIS and other fractionated organic matter in urban aerosols: Results from mass spectral and FT-IR analysis. *Environmental Science & Technology*, *50*(4), 1721–1730. https://doi.org/10.1021/acs.est.5b05277

Cheng, Y., Li, S.-M., & Leithead, A. (2006). Chemical characteristics and origins of nitrogen-containing organic compounds in $PM_{2.5}$ aerosols in the lower Fraser valley. *Environmental Science & Technology*, *40*(19), 5846–5852. https://doi.org/10.1021/es0603857

Chhantyal-Pun, R., Rotavera, B., McGillen, M. R., Khan, M. A. H., Eskola, A. J., Caravan, R. L., et al. (2018). Criegee intermediate reactions with carboxylic acids: A potential source of secondary organic aerosol in the atmosphere. *ACS Earth and Space Chemistry*, *2*(8), 833–842. https://doi.org/10.1021/acsearthspacechem.8b00069

Chow, K. S., Huang, X. H. H., & Yu, J. Z. (2016). Quantification of nitroaromatic compounds in atmospheric fine particulate matter in Hong Kong over 3 years: Field measurement evidence for secondary formation derived from biomass burning emissions. *Environmental Chemistry*, *13*(4), 665–673. https://doi.org/10.1071/EN15174

Chung, C. E., Ramanathan, V., & Decremer, D. (2012). Observationally constrained estimates of carbonaceous aerosol radiative forcing. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(29), 11624–11629. https://doi.org/10.1073/pnas.1203707109

Cortes-Ciriano, I., & Bender, A. (2015). Improved chemical structure–activity modeling through data augmentation. *Journal of Chemical Information and Modeling*, *55*(12), 2682–2692. https://doi.org/10.1021/acs.jcim.5b00570

Dallemagne, M. A., Huang, X. Y., & Eddingsaas, N. C. (2016). Variation in pH of model secondary organic aerosol during liquid–liquid phase separation. *The Journal of Physical Chemistry A*, *120*(18), 2868–2876. https://doi.org/10.1021/acs.jpca.6b00275

Davies, A., Velickovic, P., Buesing, L., Blackwell, S., Zheng, D., Tomasev, N., et al. (2021). Advancing mathematics by guiding human intuition with AI. *Nature*, *600*(7887), 70–74. https://doi.org/10.1038/s41586-021-04086-x

Fan, M.-Y., Hong, Y., Zhang, Y.-L., Sha, T., Lin, Y.-C., Cao, F., & Guo, H. (2023). Increasing nonfossil fuel contributions to atmospheric nitrate in urban China from observation to prediction. *Environmental Science & Technology*. https://doi.org/10.1021/acs.est.3c01651

Fan, X., Song, J., & Peng, P. A. (2016). Temporal variations of the abundance and optical properties of water soluble Humic-Like Substances (HULIS) in $PM_{2.5}$ at Guangzhou, China. *Atmospheric Research*, *172–173*, 8–15. https://doi.org/10.1016/j.atmosres.2015.12.024

Fan, X., Wei, S., Zhu, M., Song, J., & Peng, P. A. (2016). Comprehensive characterization of humic-like substances in smoke $PM_{2.5}$ emitted from the combustion of biomass materials and fossil fuels. *Atmospheric Chemistry and Physics*, *16*(20), 13321–13340. https://doi.org/10.5194/acp-16-13321-2016

Fan, X., Wei, S., Zhu, M., Song, J., & Peng, P. A. (2018). Molecular characterization of primary humic-like substances in fine smoke particles by thermochemolysis–gas chromatography–mass spectrometry. *Atmospheric Environment*, *180*, 1–10. https://doi.org/10.1016/j.atmosenv.2018.02.033

Fountoukis, C., & Nenes, A. (2007). ISORROPIA II: A computationally efficient thermodynamic equilibrium model for K^+-Ca^{2+}-Mg^{2+}-NH_4^+-Na^+-SO_4^{2-}-NO_3^--Cl^--H_2O aerosols. *Atmospheric Chemistry and Physics*, *7*(17), 4639–4659. https://doi.org/10.5194/acp-7-4639-2007

Fritsch, S., Guenther, F., Wright, M. N., Suling, M., & Mueller, S. M. (2019). Neuralnet: Training of neural networks [Software]. https://cran.r-project.org/web/packages/neuralnet/

Frossard, A. A., Russell, L. M., Burrows, S. M., Elliott, S. M., Bates, T. S., & Quinn, P. K. (2014). Sources and composition of submicron organic mass in marine aerosol particles. *Journal of Geophysical Research: Atmospheres*, *119*(22), 12977–13003. https://doi.org/10.1002/2014jd021913

Goodman, J. (2009). Computer software review: Reaxys. *Journal of Chemical Information and Modeling*, *49*(12), 2897–2898. https://doi.org/10.1021/ci900437n

Guo, Z., Yang, Y., Hu, X., Peng, X., Fu, Y., Sun, W., et al. (2022). The optical properties and in-situ observational evidence for the formation of brown carbon in clouds. *Atmospheric Chemistry and Physics*, *22*(7), 4827–4839. https://doi.org/10.5194/acp-22-4827-2022

Hems, R. F., & Abbatt, J. P. D. (2018). Aqueous phase photo-oxidation of brown carbon nitrophenols: Reaction kinetics, mechanism, and evolution of light absorption. *ACS Earth and Space Chemistry*, *2*(3), 225–234. https://doi.org/10.1021/acsearthspacechem.7b00123

Higashiguchi, K., Matsuda, K., Asano, Y., Murakami, A., Nakamura, S., & Irie, M. (2005). Photochromism of dithienylethenes containing fluorinated thiophene rings. *European Journal of Organic Chemistry*, *2005*(1), 91–97. https://doi.org/10.1002/ejoc.200400441

Hong, Y., Cao, F., Fan, M.-Y., Lin, Y.-C., Bao, M., Xue, Y., et al. (2022). Using machine learning to quantify sources of light-absorbing water-soluble humic-like substances (HULISws) in Northeast China. *Atmospheric Environment*, *291*, 119371. https://doi.org/10.1016/j.atmosenv.2022.119371

Hong, Y., Zhang, Y.-L., Bao, M., Fan, M.-Y., Lin, Y.-C., Xu, R., et al. (2023). The ESI(-) FT-ICR-MS datasets of water-soluble humic-like substances in Nanjing, China [Dataset]. OSF. https://doi.org/10.17605/OSF.IO/BKX9H

Huang, Z.-J., Li, H., Luo, J.-Y., Li, S., & Liu, F. (2023). Few-shot learning-based, long-term stable, sensitive chemosensor for on-site colorimetric detection of Cr(VI). *Analytical Chemistry*, *95*(14), 6156–6162. https://doi.org/10.1021/acs.analchem.3c00604

Humes, M. B., Wang, M., Kim, S., Machesky, J. E., Gentner, D. R., Robinson, A. L., et al. (2022). Limited secondary organic aerosol production from acyclic oxygenated volatile chemical products. *Environmental Science & Technology*, *56*(8), 4806–4815. https://doi.org/10.1021/acs.est.1c07354

Jablonka, K. M., Ongari, D., Moosavi, S. M., & Smit, B. (2020). Big-data science in porous materials: Materials genomics and machine learning. *Chemical Reviews*, *120*(16), 8066–8129. https://doi.org/10.1021/acs.chemrev.0c00004

Jiang, H., Li, J., Chen, D., Tang, J., Cheng, Z., Mo, Y., et al. (2020). Biomass burning organic aerosols significantly influence the light absorption properties of polarity-dependent organic compounds in the Pearl River Delta Region, China. *Environment International*, *144*, 106079. https://doi.org/10.1016/j.envint.2020.106079

Jiang, H., Li, J., Sun, R., Liu, G., Tian, C., Tang, J., et al. (2021). Determining the sources and transport of brown carbon using radionuclide tracers and modeling. *Journal of Geophysical Research: Atmospheres*, *126*(9), e2021JD034616. https://doi.org/10.1029/2021jd034616

Jiang, H., Li, J., Sun, R., Tian, C., Tang, J., Jiang, B., et al. (2021). Molecular dynamics and light absorption properties of atmospheric dissolved organic matter. *Environmental Science & Technology*, *55*(15), 10268–10279. https://doi.org/10.1021/acs.est.1c01770

Jiang, H., Li, J., Tang, J., Cui, M., Zhao, S., Mo, Y., et al. (2022). Molecular characteristics, sources, and formation pathways of organosulfur compounds in ambient aerosol in Guangzhou, South China. *Atmospheric Chemistry and Physics*, *22*(10), 6919–6935. https://doi.org/10.5194/acp-22-6919-2022

Jiang, H., Li, J., Tang, J., Zhao, S., Chen, Y., Tian, C., et al. (2022). Factors influencing the molecular compositions and distributions of atmospheric nitrogen-containing compounds. *Journal of Geophysical Research: Atmospheres*, *127*(7), e2021JD036284. https://doi.org/10.1029/2021JD036284

Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., & Cho, K. (2019). Augmentation for small object detection. *ArXiv, abs/1902.07296*. https://doi.org/10.48550/arXiv.1902.07296

Kucinski, T. M., Ott, E.-J. E., & Freedman, M. A. (2021). Dynamics of liquid–liquid phase separation in submicrometer aerosol. *The Journal of Physical Chemistry A*, *125*(20), 4446–4453. https://doi.org/10.1021/acs.jpca.1c01985

Laskin, J., Laskin, A., & Nizkorodov, S. A. (2018). Mass spectrometry analysis in atmospheric chemistry. *Analytical Chemistry*, *90*(1), 166–189. https://doi.org/10.1021/acs.analchem.7b04249

Li, J., Chen, Q., Sha, T., & Liu, Y. (2022). Significant promotion of light absorption ability and formation of triplet organics and reactive oxygen species in atmospheric HULIS by Fe(III) ions. *Environmental Science & Technology*, *56*(23), 16652–16664. https://doi.org/10.1021/acs.est.2c05137

Li, X., Yang, K., Han, J., Ying, Q., & Hopke, P. K. (2019). Sources of humic-like substances (HULIS) in $PM_{2.5}$ in Beijing: Receptor modeling approach. *Science of the Total Environment*, *671*, 765–775. https://doi.org/10.1016/j.scitotenv.2019.03.333

Li, X., Yang, Y., Liu, S., Zhao, Q., Wang, G., & Wang, Y. (2020). Light absorption properties of brown carbon (BrC) in autumn and winter in Beijing: Composition, formation and contribution of nitrated aromatic compounds. *Atmospheric Environment*, *223*, 117289. https://doi.org/10.1016/j.atmosenv.2020.117289

Liland, K. H., Mevik, B.-H., Wehrens, R., & Hiemstra, P. (2023). Pls: Partial least squares and principal component regression [Software]. https://cran.r-project.org/web/packages/pls/

Lin, P., Bluvshtein, N., Rudich, Y., Nizkorodov, S. A., Laskin, J., & Laskin, A. (2017). Molecular chemistry of atmospheric brown carbon inferred from a nationwide biomass burning event. *Environmental Science & Technology*, *51*(20), 11561–11570. https://doi.org/10.1021/acs.est.7b02276

Lin, P., Huang, X.-F., He, L.-Y., & Zhen Yu, J. (2010). Abundance and size distribution of HULIS in ambient aerosols at a rural site in South China. *Journal of Aerosol Science*, *41*(1), 74–87. https://doi.org/10.1016/j.jaerosci.2009.09.001

Lin, P., Rincon, A. G., Kalberer, M., & Yu, J. Z. (2012). Elemental composition of HULIS in the Pearl River Delta Region, China: Results inferred from positive and negative electrospray high resolution mass spectrometric data. *Environmental Science & Technology*, *46*(14), 7454–7462. https://doi.org/10.1021/es300285d

Lin, Y. C., Zhang, Y. L., Fan, M. Y., & Bao, M. (2020). Heterogeneous formation of particulate nitrate under ammonium-rich regimes during the high-PM2.5 events in Nanjing, China. *Atmospheric Chemistry and Physics*, *20*(6), 3999–4011. https://doi.org/10.5194/acp-20-3999-2020

Lyu, R., Alam, M. S., Stark, C., Xu, R., Shi, Z., Feng, Y., & Harrison, R. M. (2019). Aliphatic carbonyl compounds (C8–C26) in wintertime atmospheric aerosol in London, UK. *Atmospheric Chemistry and Physics*, *19*(4), 2233–2246. https://doi.org/10.5194/acp-19-2233-2019

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., & Lin, C.-C. (2023). Misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien. [Software]. E1071. https://cran.r-project.org/web/packages/e1071/

Millet, D. B., Baasandorj, M., Farmer, D. K., Thornton, J. A., Baumann, K., Brophy, P., et al. (2015). A large and ubiquitous source of atmospheric formic acid. *Atmospheric Chemistry and Physics*, *15*(11), 6283–6304. https://doi.org/10.5194/acp-15-6283-2015

Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, *16*, 100258. https://doi.org/10.1016/j.array.2022.100258

Noziere, B., Kalberer, M., Claeys, M., Allan, J., D'Anna, B., Decesari, S., et al. (2015). The molecular identification of organic compounds in the atmosphere: State of the art and challenges. *Chemical Reviews*, *115*(10), 3919–3983. https://doi.org/10.1021/cr5003485

Parnami, A., & Lee, M. (2022). Learning from few examples: A summary of approaches to few-shot learning. ArXiv, abs/2203.04291. https://doi.org/10.48550/arXiv.2203.04291

Phillips, S. M., Bellcross, A. D., & Smith, G. D. (2017). Light absorption by brown carbon in the southeastern United States is pH-dependent. *Environmental Science & Technology*, *51*(12), 6782–6790. https://doi.org/10.1021/acs.est.7b01116

Santos, P. S., Santos, E. B., & Duarte, A. C. (2012). First spectroscopic study on the structural features of dissolved organic matter isolated from rainwater in different seasons. *Science of the Total Environment*, *426*, 172–179. https://doi.org/10.1016/j.scitotenv.2012.03.023

Schmedding, R., Rasool, Q. Z., Zhang, Y., Pye, H. O. T., Zhang, H., Chen, Y., et al. (2020). Predicting secondary organic aerosol phase state and viscosity and its effect on multiphase chemistry in a regional-scale air quality model. *Atmospheric Chemistry and Physics*, *20*(13), 8201–8225. https://doi.org/10.5194/acp-20-8201-2020

Song, J., Li, M., Jiang, B., Wei, S., Fan, X., & Peng, P. A. (2018). Molecular characterization of water-soluble humic-like substances in smoke particles emitted from combustion of biomass materials and coal using ultrahigh-resolution electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Environmental Science & Technology*, *52*(5), 2575–2585. https://doi.org/10.1021/acs.est.7b06126

Tapavicza, E., von Rudorff, G. F., De Haan, D. O., Contin, M., George, C., Riva, M., & von Lilienfeld, O. A. (2021). Elucidating an atmospheric brown carbon species toward supplanting chemical intuition with exhaustive enumeration and machine learning. *Environmental Science & Technology*, *55*(12), 8447–8457. https://doi.org/10.1021/acs.est.1c00885

Wang, Q., He, X., Zhou, M., Huang, D. D., Qiao, L., Zhu, S., et al. (2020). Hourly measurements of organic molecular markers in urban Shanghai, China: Primary organic aerosol source identification and observation of cooking aerosol aging. *ACS Earth and Space Chemistry*, *4*(9), 1670–1685. https://doi.org/10.1021/acsearthspacechem.0c00205

Wang, X., Heald, C. L., Liu, J., Weber, R. J., Campuzano-Jost, P., Jimenez, J. L., et al. (2018). Exploring the observational constraints on the simulation of brown carbon. *Atmospheric Chemistry and Physics*, *18*(2), 635–653. https://doi.org/10.5194/acp-18-635-2018

Wang, Y., Hu, M., Lin, P., Tan, T., Li, M., Xu, N., et al. (2019). Enhancement in particulate organic nitrogen and light absorption of humic-like substances over Tibetan Plateau due to long-range transported biomass burning emissions. *Environmental Science & Technology*, *53*(24), 14222–14232. https://doi.org/10.1021/acs.est.9b06152

Wang, Y., Hu, M., Wang, Y., Zheng, J., Shang, D., Yang, Y., et al. (2019). The formation of nitro-aromatic compounds under high NOx and anthropogenic VOC conditions in urban Beijing, China. *Atmospheric Chemistry and Physics*, *19*(11), 7649–7665. https://doi.org/10.5194/acp-19-7649-2019

Wei, P., Lu, Z., & Song, J. (2015). A comprehensive comparison of two variable importance analysis techniques in high dimensions: Application to an environmental multi-indicators system. *Environmental Modelling & Software*, *70*, 178–190. https://doi.org/10.1016/j.envsoft.2015.04.015

Wright, M. N., Wager, S., & Probst, P. (2023). Ranger: A fast implementation of random forests [Software]. https://cran.r-project.org/web/packages/ranger/

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C^{++} and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Yakimov, B. P., Rubekina, A. A., Zherebker, A. Y., Budylin, G. S., Kompanets, V. O., Chekalin, S. V., et al. (2022). Oxidation of individual aromatic species gives rise to humic-like optical properties. *Environmental Science and Technology Letters*, *9*(5), 452–458. https://doi.org/10.1021/acs.estlett.2c00161

Zeng, Y., Shen, Z., Takahama, S., Zhang, L., Zhang, T., Lei, Y., et al. (2020). Molecular absorption and evolution mechanisms of $PM_{2.5}$ brown carbon revealed by electrospray ionization Fourier transform–ion cyclotron resonance mass spectrometry during a severe winter pollution episode in Xi'an, China. *Geophysical Research Letters*, *47*(10), e2020GL087977. https://doi.org/10.1029/2020gl087977

Zhang, A., Wang, Y., Zhang, Y., Weber, R. J., Song, Y., Ke, Z., & Zou, Y. (2020). Modeling the global radiative effect of brown carbon: A potentially larger heating source in the tropical free troposphere than black carbon. *Atmospheric Chemistry and Physics*, *20*(4), 1901–1920. https://doi.org/10.5194/acp-20-1901-2020

Zhang, D., Ren, H., Hu, W., Wu, L., Ren, L., Deng, J., et al. (2022). Latitudinal difference in the molecular distributions of lipid compounds in the forest atmosphere in China. *Environmental Pollution*, *294*, 118578. https://doi.org/10.1016/j.envpol.2021.118578

Zhang, T., Shen, Z., Zeng, Y., Cheng, C., Wang, D., Zhang, Q., et al. (2021). Light absorption properties and molecular profiles of HULIS in $PM_{2.5}$ emitted from biomass burning in traditional "Heated Kang" in Northwest China. *Science of the Total Environment*, *776*, 146014. https://doi.org/10.1016/j.scitotenv.2021.146014

Zhang, Y., Forrister, H., Liu, J., Dibb, J., Anderson, B., Schwarz, J. P., et al. (2017). Top-of-atmosphere radiative forcing affected by brown carbon in the upper troposphere. *Nature Geoscience*, *10*(7), 486–489. https://doi.org/10.1038/NGEO2960

Zhang, Y.-L., Zhang, W., Fan, M.-Y., Li, J., Fang, H., Cao, F., et al. (2022). A diurnal story of $\Delta^{17}O(NO_3^-)$ in urban Nanjing and its implication for nitrate aerosol formation. *Npj Climate and Atmospheric Science*, *5*(1), 50. https://doi.org/10.1038/s41612-022-00273-3

Zhao, J., Zhang, Y., Yan, X., Gan, L., & Wang, G. (2021). Investigation of UV-vis spectra of azobenzene containing carboxyl groups. *Journal of Molecular Modeling*, *27*(3), 79. https://doi.org/10.1007/s00894-021-04702-7

References From the Supporting Information

Almandoz, M. C., Sancho, M. I., Duchowicz, P. R., & Blanco, S. E. (2014). UV–Vis spectroscopic study and DFT calculation on the solvent effect of trimethoprim in neat solvents and aqueous mixtures. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *129*, 52–60. https://doi.org/10.1016/j.saa.2014.02.191

Capitaine, L., Genuer, R., & Thiébaut, R. (2020). Random forests for high-dimensional longitudinal data. *Statistical Methods in Medical Research*, *30*(1), 166–184. https://doi.org/10.1177/0962280220946080

Chan, A. W. H., Kreisberg, N. M., Hohaus, T., Campuzano-Jost, P., Zhao, Y., Day, D. A., et al. (2016). Speciated measurements of semivolatile and intermediate volatility organic compounds (S/IVOCs) in a pine forest during BEACHON-RoMBAS 2011. *Atmospheric Chemistry and Physics*, *16*(2), 1187–1205. https://doi.org/10.5194/acp-16-1187-2016

Cleland, D. M., Gordon, K. C., Officer, D. L., Wagner, P., & Walsh, P. J. (2009). Tuning the optical properties of ZnTPP using carbonyl ring fusion. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *74*(4), 931–935. https://doi.org/10.1016/j.saa.2009.08.040

Cui, M., Li, C., Chen, Y., Zhang, F., Li, J., Jiang, B., et al. (2019). Molecular characterization of polar organic aerosol constituents in off-road engine emissions using Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS): Implications for source apportionment. *Atmospheric Chemistry and Physics*, *19*(22), 13945–13956. https://doi.org/10.5194/acp-19-13945-2019

Du, J., Kodikara, M. S., Moxey, G. J., Morshedi, M., Barlow, A., Quintana, C., et al. (2018). Quadratic and cubic hyperpolarizabilities of nitro-phenyl/-naphthalenyl/-anthracenyl alkynyl complexes. *Dalton Transactions*, *47*(13), 4560–4571. https://doi.org/10.1039/C8DT00155C

Fan, X., Song, J., & Peng, P. A. (2012). Comparison of isolation and quantification methods to measure humic-like substances (HULIS) in atmospheric particles. *Atmospheric Environment*, *60*, 366–374. https://doi.org/10.1016/j.atmosenv.2012.06.063

Genuer, R., Poggi, J.-M., & Tuleau, C. (2008). Random forests: Some methodological insights. *arXiv: Machine Learning*. https://doi.org/10.48550/arXiv.0811.3619

Goldstein, B. A., Polley, E. C., & Briggs, F. B. S. (2011). Random forests for genetic association studies. *Random Forests for Genetic Association Studies*, *10*(1). https://doi.org/10.2202/1544-6115.1691

Grace, D. N., Sharp, J. R., Holappa, R. E., Lugos, E. N., Sebold, M. B., Griffith, D. R., et al. (2019). Heterocyclic product formation in aqueous brown carbon systems. *ACS Earth and Space Chemistry*, *3*(11), 2472–2481. https://doi.org/10.1021/acsearthspacechem.9b00235

Grange, S. K., & Carslaw, D. C. (2019). Using meteorological normalisation to detect interventions in air quality time series. *Science of the Total Environment*, *653*, 578–588. https://doi.org/10.1016/j.scitotenv.2018.10.344

Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., & Hueglin, C. (2018). Random forest meteorological normalisation models for Swiss PM10 trend analysis. *Atmospheric Chemistry and Physics*, *18*(9), 6223–6239. https://doi.org/10.5194/acp-18-6223-2018

Hays, M. D., Smith, N. D., & Dong, Y. (2004). Nature of unresolved complex mixture in size-distributed emissions from residential wood combustion as measured by thermal desorption-gas chromatography-mass spectrometry. *Journal of Geophysical Research*, *109*(D16). https://doi.org/10.1029/2003JD004051

Hong, Y., Cao, F., Fan, M. Y., Lin, Y. C., Gul, C., Yu, M., et al. (2022). Impacts of chemical degradation of levoglucosan on quantifying biomass burning contribution to carbonaceous aerosols: A case study in Northeast China. *Science of the Total Environment*, *819*, 152007. https://doi.org/10.1016/j.scitotenv.2021.152007

Jiang, J., Carter, W. P. L., Cocker, D. R., & Barsanti, K. C. (2020). Development and evaluation of a detailed mechanism for gas-phase atmospheric reactions of furans. *ACS Earth and Space Chemistry*, *4*(8), 1254–1268. https://doi.org/10.1021/acsearthspacechem.0c00058

Kautzman, K. E., Surratt, J. D., Chan, M. N., Chan, A. W. H., Hersey, S. P., Chhabra, P. S., et al. (2010). Chemical composition of gas- and aerosol-phase products from the photooxidation of Naphthalene. *The Journal of Physical Chemistry A*, *114*(2), 913–934. https://doi.org/10.1021/jp908530s

Khare, P., Machesky, J., Soto, R., He, M., Presto, A. A., & Gentner, D. R. (2020). Asphalt-related emissions are a major missing nontraditional source of secondary organic aerosol precursors. *Science Advances*, *6*(36), eabb9785. https://doi.org/10.1126/sciadv.abb9785

Koch, B. P., & Dittmar, T. (2006). From mass to structure: An aromaticity index for high-resolution mass data of natural organic matter. *Rapid Communications in Mass Spectrometry*, *20*(5), 926–932. https://doi.org/10.1002/rcm.2386

Kroll, J. H., Donahue, N. M., Jimenez, J. L., Kessler, S. H., Canagaratna, M. R., Wilson, K. R., et al. (2011). Carbon oxidation state as a metric for describing the chemistry of atmospheric organic aerosol. *Nature Chemistry*, *3*(2), 133–139. https://doi.org/10.1038/nchem.948

Lang, L., Tiancai, L., Shan, A., & Xiangyan, T. (2021). Erratum to: An improved random forest algorithm and its application to wind pressure prediction. *International Journal of Intelligent Systems*, *37*(2), 1802. https://doi.org/10.1002/int.22662

Liu, Q., Gao, Y., Huang, W., Ling, Z., Wang, Z., & Wang, X. (2022). Carbonyl compounds in the atmosphere: A review of abundance, source and their contributions to O_3 and SOA formation. *Atmospheric Research*, *274*, 106184. https://doi.org/10.1016/j.atmosres.2022.106184

Luo, Y., Zhou, X., Zhang, J., Xue, L., Chen, T., Zheng, P., et al. (2020). Characteristics of airborne water-soluble organic carbon (WSOC) at a background site of the North China Plain. *Atmospheric Research*, *231*, 104668. https://doi.org/10.1016/j.atmosres.2019.104668

Metreveli, N. O., Jariashvili, K. K., Namicheishvili, L. O., Svintradze, D. V., Chikvaidze, E. N., Sionkowska, A., & Skopinska, J. (2010). UV–vis and FT-IR spectra of ultraviolet irradiated collagen in the presence of antioxidant ascorbic acid. *Ecotoxicology and Environmental Safety*, *73*(3), 448–455. https://doi.org/10.1016/j.ecoenv.2009.12.005

Meyer, D. (2001). Support Vector Machines ∗ The Interface to libsvm in package e1071.

Mo, Y., Zhong, G., Li, J., Liu, X., Jiang, H., Tang, J., et al. (2022). The sources, molecular compositions, and light absorption properties of water-soluble organic carbon in marine aerosols from South China Sea to the Eastern Indian Ocean. *Journal of Geophysical Research: Atmospheres*, *127*(10). https://doi.org/10.1029/2021jd036168

Mousavi, M., Martis, V., & Fini, E. H. (2021). Inherently functionalized carbon from algae to adsorb precursors of secondary organic aerosols in noncombustion sources. *ACS Sustainable Chemistry & Engineering*, *9*(43), 14375–14384. https://doi.org/10.1021/acssuschemeng.1c03827

Ning, C., Gao, Y., Yu, H., Zhang, H., Geng, N., Cao, R., & Chen, J. (2022). FT-ICR mass spectrometry for molecular characterization of water-insoluble organic compounds in winter atmospheric fine particulate matters. *Journal of Environmental Sciences*, *111*, 51–60. https://doi.org/10.1016/j.jes.2020.12.017

Puri, A., Jang, H., Yavlovich, A., Masood, M. A., Veenstra, T. D., Luna, C., et al. (2011). Material properties of matrix lipids determine the conformation and intermolecular reactivity of diacetylenic phosphatidylcholine in the lipid bilayer. *Langmuir*, *27*(24), 15120–15128. https://doi.org/10.1021/la203453x

Rivera, W. A. (2017). Noise reduction a priori synthetic over-sampling for class imbalanced data sets. *Information Sciences*, *408*, 146–161. https://doi.org/10.1016/j.ins.2017.04.046

Song, J., Li, M., Fan, X., Zou, C., Zhu, M., Jiang, B., et al. (2019). Molecular characterization of water- and methanol-soluble organic compounds emitted from residential coal combustion using ultrahigh-resolution electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Environmental Science & Technology*, *53*(23), 13607–13617. https://doi.org/10.1021/acs.est.9b04331

Song, J., Li, M., Zou, C., Cao, T., Fan, X., Jiang, B., et al. (2022). Molecular characterization of nitrogen-containing compounds in humic-like substances emitted from biomass burning and coal combustion. *Environmental Science & Technology*, *56*(1), 119–130. https://doi.org/10.1021/acs.est.1c04451

Suhr, K. J., Bastatas, L. D., Shen, Y., Mitchell, L. A., Frazier, G. A., Taylor, D. W., et al. (2016). Phenyl substitution of cationic bis-cyclometalated iridium(iii) complexes for iTMC-LEECs. *Dalton Transactions*, *45*(44), 17807–17823. https://doi.org/10.1039/C6DT03415B

Tang, J., Li, J., Su, T., Han, Y., Mo, Y., Jiang, H., et al. (2020). Molecular compositions and optical properties of dissolved brown carbon in biomass burning, coal combustion, and vehicle emission aerosols illuminated by excitation-emission matrix spectroscopy and Fourier transform ion cyclotron resonance mass spectrometry analysis. *Atmospheric Chemistry and Physics*, *20*(4), 2513–2532. https://doi.org/10.5194/acp-20-2513-2020

Wang, S., Zhou, S., Tao, Y., Tsui, W. G., Ye, J., Yu, J. Z., et al. (2019). Organic peroxides and sulfur dioxide in aerosol: Source of particulate sulfate. *Environmental Science & Technology*, *53*(18), 10695–10704. https://doi.org/10.1021/acs.est.9b02591

Wang, W., Liu, M., Wang, T., Song, Y., Zhou, L., Cao, J., et al. (2021). Sulfate formation is dominated by manganese-catalyzed oxidation of SO_2 on aerosol surfaces during haze events. *Nature Communications*, *12*(1), 1993. https://doi.org/10.1038/s41467-021-22091-6

Wang, Y., Hu, M., Lin, P., Guo, Q., Wu, Z., Li, M., et al. (2017). Molecular characterization of nitrogen-containing organic compounds in humic-like substances emitted from straw residue burning. *Environmental Science & Technology*, *51*(11), 5951–5961. https://doi.org/10.1021/acs.est.7b00248

Wu, G., Wan, X., Gao, S., Fu, P., Yin, Y., Li, G., et al. (2018). Humic-like substances (HULIS) in aerosols of central Tibetan Plateau (Nam Co, 4730 masl): Abundance, light absorption properties, and sources. *Environmental Science & Technology*, *52*(13), 7203–7211. https://doi.org/10.1021/acs.est.8b01251

Xu, Q., Li, Z., Liu, N., Jia, J., Yang, J., & Song, Y. (2019). Third order nonlinear optical properties and transient dynamics of thiophene-contained pyrene derivatives: Effect of peripheral substituent group. *Optics & Laser Technology*, *109*, 666–670. https://doi.org/10.1016/j.optlastec.2018.08.048

Yazdani, A., Dudani, N., Takahama, S., Bertrand, A., Prévôt, A. S. H., El Haddad, I., & Dillner, A. M. (2021). Characterization of primary and aged wood burning and coal combustion organic aerosols in an environmental chamber and its implications for atmospheric aerosols. *Atmospheric Chemistry and Physics*, *21*(13), 10273–10293. https://doi.org/10.5194/acp-21-10273-2021

Zeng, Y., Ning, Y., Shen, Z., Zhang, L., Zhang, T., Lei, Y., et al. (2021). The roles of N, S, and O in molecular absorption features of brown carbon in $PM_{2.5}$ in a typical semi-arid megacity in Northwestern China. *Journal of Geophysical Research: Atmospheres*, *126*(16). https://doi.org/10.1029/2021jd034791

Zhang, Q., Li, Z., Shen, Z., Zhang, T., Zhang, Y., Sun, J., et al. (2022). Source profiles of molecular structure and light absorption of $PM_{2.5}$ brown carbon from residential coal combustion emission in Northwestern China. *Environmental Pollution*, *299*, 118866. https://doi.org/10.1016/j.envpol.2022.118866