Environmental Science: Atmospheres

PAPER

View Article Online
View Journal | View Issue

Cite this: *Environ. Sci.: Atmos.*, 2023, **3**, 230

Precursor apportionment of atmospheric oxygenated organic molecules using a machine learning method†

Xiaohui Qiao,[a] Xiaoxiao Li,[a] Chao Yan,[bcd] Nina Sarnela,[c] Rujing Yin,[a] Yishuo Guo,[d] Lei Yao,[cd] Wei Nie,[b] Dandan Huang,[e] Zhe Wang,[f] Federico Bianchi,[cd] Yongchun Liu,[d] Neil M. Donahue,[gh] Markku Kulmala[cd] and Jingkun Jiang *[a]

Gas-phase oxygenated organic molecules (OOMs) can contribute significantly to both atmospheric new particle growth and secondary organic aerosol formation. Precursor apportionment of atmospheric OOMs connects them with volatile organic compounds (VOCs). Since atmospheric OOMs are often highly functionalized products of multistep reactions, it is challenging to reveal the complete mapping relationships between OOMs and their precursors. In this study, we demonstrate that the machine learning method is useful in attributing atmospheric OOMs to their precursors using several chemical indicators, such as O/C ratio and H/C ratio. The model is trained and tested using data acquired in controlled laboratory experiments, covering the oxidation products of four main types of VOCs (isoprene, monoterpenes, aliphatics, and aromatics). Then, the model is used for analyzing atmospheric OOMs measured in both urban Beijing and a boreal forest environment in southern Finland. The results suggest that atmospheric OOMs in these two environments can be reasonably assigned to their precursors. Beijing is an anthropogenic VOC dominated environment with ~64% aromatic and aliphatic OOMs, and the other boreal forested area has ~76% monoterpene OOMs. This pilot study shows that machine learning can be a promising tool in atmospheric chemistry for connecting the dots.

Received 3rd October 2022
Accepted 30th November 2022

DOI: 10.1039/d2ea00128d

rsc.li/esatmospheres

[a]State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, 100084, Beijing, China. E-mail: jiangjk@tsinghua.edu.cn

[b]Joint International Research Laboratory of Atmospheric and Earth System Research, School of Atmospheric Sciences, Nanjing University, Nanjing, China

[c]Institute for Atmospheric and Earth System Research/Physics, Faculty of Science, University of Helsinki, 00014, Helsinki, Finland

[d]Aerosol and Haze Laboratory, Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing University of Chemical Technology, 100029, Beijing, China

[e]State Environmental Protection Key Laboratory of Formation and Prevention of Urban Air Pollution Complex, Shanghai Academy of Environmental Sciences, Shanghai, China

[f]Division of Environment and Sustainability, The Hong Kong University of Science and Technology, Hong Kong SAR, China

[g]Center for Atmospheric Particle Studies, Carnegie Mellon University, Pittsburgh, PA, USA

[h]Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA, USA

† Electronic supplementary information (ESI) available: Detailed information on the evaluation of the precursor apportionment model (Table S1); examples of the voting strategy (Table S2); examples of the expansion of aliphatic OOMs (Table S3); the Venn-plot of laboratory-generated OOMs used in this study (Fig. S1); the apportionment results of OOMs in urban Beijing using the workflow method (Fig. S2); the overall accuracy of the decision tree model using a laboratory dataset without the expansion of aliphatic OOMs (Fig. S3); the application results for the model trained with the dataset without the expansion of aliphatic OOMs (Fig. S4); and the performances of models consisting of different number of trees (Fig. S5). See DOI: https://doi.org/10.1039/d2ea00128d

Environmental significance

The formation of new particles and secondary organic aerosols has potential effects on the earth's radiation balance and air quality, and atmospheric oxygenated organic molecules (OOMs) are currently acknowledged to significantly contribute to them. Volatile organic compounds (VOCs) are important precursors for OOMs. Since OOMs are often highly functionalized products of multistep reactions in the complex atmospheric environments, it is challenging to reveal the complete mapping relationships between OOMs and their precursors. In this study, we demonstrate that the machine learning method is useful in attributing OOMs to their precursors using several chemical indicators, such as O/C ratio and H/C ratio. Based on controlled oxidation experiments of four main types of VOCs (isoprene, monoterpenes, aliphatics, and aromatics), we trained the machine learning model to extract the internal mapping relationships of OOMs and their precursors. It showed unique advantages in precursor apportionment over current methods. We then used it for analyzing atmospheric OOMs measured in both urban Beijing and forested Hyytiälä. It can well identify the differences in OOMs between anthropogenic and biogenic dominated atmospheric environments. This pilot study shows that machine learning can be a promising tool in atmospheric chemistry for connecting the dots and is worth further exploring.

1 Introduction

Gas-phase oxygenated organic molecules (OOMs) can condense onto particles and usually play dominant roles in atmospheric

new particle growth and even in secondary organic aerosol formation,[1–4] which is known to affect the earth's radiation balance and air quality. OOMs are formed in the atmosphere *via* complex oxidation of volatile organic compounds (VOCs), in which autoxidation and multi-generation oxidation are usually involved.[5,6] Precursor apportionment of atmospheric OOMs is the cornerstone of understanding their origins, which helps in regulating the emissions of their precursors.

Both the complex functionalization processes and the differences between controlled laboratory conditions and real atmospheric environments make precursor apportionment of atmospheric OOMs very challenging. The VOC precursors of OOMs can generally be divided into anthropogenic volatile organic compounds (*e.g.*, aromatics and aliphatics) and biogenic volatile organic compounds (*e.g.*, monoterpenes and isoprene).[7] Positive matrix factorization (PMF)[8] is a widely used source apportionment analysis method,[9,10] which classifies numerous species into several factors based on their similarities in temporal variations. However, PMF could not sufficiently attribute atmospheric OOMs to their precursors.[9,11] It is partially because the concentration of OOMs is not only affected by their precursors but also strongly affected by the oxidation processes under the given atmospheric conditions. Atmospheric OOMs from different precursors may share similar time series if they are oxidized by the same oxidants. For example, OOMs generated from the photo-oxidation of aromatics and monoterpenes cannot be separated by the PMF analysis.[3] Recently, Nie *et al.*[3] developed a workflow method, which performs precursor apportionment of atmospheric OOMs based on the up-to-date knowledge of the characteristics of the products from VOC oxidation processes. In that method, however, the identification of OOMs from monoterpene oxidation (monoterpene OOMs) still relies on PMF. This is known to underestimate the concentration of monoterpene OOMs because monoterpene oxidized by the OH radical, which exists under the given conditions, cannot easily be retrieved by PMF.[3,12] Controlled laboratory experiments provide important information on OOMs produced from different precursors. However, the chemical settings of controlled laboratory experiments cannot fully reflect the complexity of the real atmosphere, which makes it difficult to interpret atmospheric OOMs solely based on laboratory conditions.

Machine learning methods possess the advantage of mining the relationships among complex data. For instance, the decision tree model, one of the classical machine learning methods, has been successfully applied in predicting amino acid sequencing and in proteomics research due to its high interpretability and tolerance of data scale.[13,14] With the development and application of high-resolution mass spectrometry in the field of atmospheric chemistry, large datasets of atmospheric OOMs at the molecular level are obtained with high time resolution for long-term periods.[4,15] Thus, finding precursors for atmospheric OOMs appears to be a natural playground for machine learning methods. They have the potential to mine large datasets from controlled laboratory experiments with no reliance on the variation of OOM concentration and further attribute OOMs measured in ambient atmosphere to their likely precursors.

In this work, we use the decision tree model to test the feasibility of machine learning for precursor apportionment of atmospheric OOMs. This model is trained and tested using the datasets from controlled laboratory experiments using various VOCs as OOM precursors. It helps build a mapping relationship between OOMs and their precursors, including isoprene, monoterpenes, aliphatics, and aromatics. Finally, we apply the model to atmospheric datasets obtained in both urban Beijing and a remote forest environment of Hyytiälä.

2 Methods

2.1 Model training strategy

Decision tree was selected for precursor apportionment of OOMs, performed with the *Statistics and Machine Learning Toolbox* in MATLAB. It is a supervised machine learning method, which requires training data that are *pre-labeled* with known precursors (from controlled laboratory experiments). According to the standard training process of machine learning models, we randomly divided the pre-labeled data into the training dataset and the testing dataset, accounting for 70% and 30% of the data, respectively. As shown in Fig. 1, the training dataset was used to obtain the decision tree, and the testing dataset was used to evaluate the accuracy of the trained model. Cross-validation was also performed which is described in the ESI.†

In order to reduce the uncertainty caused by one single training model, we repeated the above process ten times to get ten independent decision trees. The outputs of all the ten decision trees vote together for the final apportionment result. Specifically, precursors with votes more than the upper limit, *i.e.*, the sum of the mean and standard deviation of the whole votes, will be retained, or, if all the precursors received votes no more than the upper limit, precursors with votes more than the lower limit will be retained, *i.e.*, the difference of the mean and standard deviation of the whole votes. There are examples in the ESI† to illustrate this rule.

This voting strategy also helps to reduce the uncertainty caused by the overlapping formulae of OOMs generated from different precursors. According to the laboratory experiments, there are overlaps between OOMs oxidized from those precursors (Fig. S1†). Therefore, in the pre-labeled dataset, the same descriptors of an OOM molecule may correspond to more than one label of precursors. The ten decision trees would take into account the overlaps and give a combination of the most likely answers. Table S2† shows two examples of the determination of overlapping molecules. The optimization process for the number of trees is described in the ESI.†

2.2 Pre-labeled dataset and atmospheric dataset

Pre-labeled data are obtained from single-precursor oxidation experiments performed under laboratory-controlled conditions. Precursors are marked with four labels, *i.e.*, isoprene, monoterpenes, aliphatics, and aromatics. Isoprene and

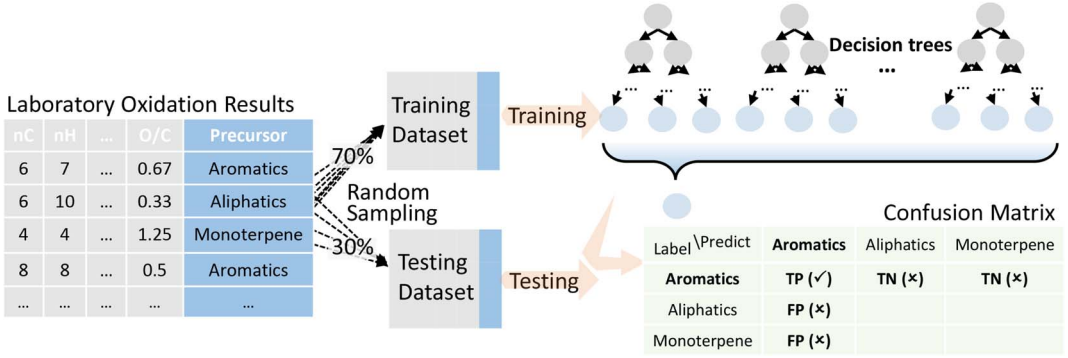

Fig. 1 Schematic of the decision tree in the precursor apportionment of atmospheric oxygenated organic molecules. In the confusion matrix, TP, FP, or TN is a description of the labels and predictions, in which T is for true, F is for false, P is for positive, and N is for negative. The above table takes aromatics as an example. The first letter is for reality and the latter is for prediction. Thus, when aromatics are predicted as aromatics, the result is TP; when aromatics are predicted as aliphatics and monoterpenes, the result is TN; when aliphatics and monoterpenes are predicted as aromatics, the result is FP.

monoterpenes such as α-pinene and limonene are representative biogenic precursors with high formation yields of OOMs. From previous monoterpene oxidation experiments, 872 products of monoterpene OOMs were reported (Table 1).[16,17] As the laboratory report for comprehensive isoprene OOMs is not available, we obtained 30 isoprene OOMs from a previous field measurement.[18] Aromatics are representative anthropogenic precursors and are considered to make a dominant contribution to the formation of OOMs in an urban atmospheric environment. According to previous studies of aromatic oxidation experiments, 485 oxidation products were used in the pre-labeled dataset.[6,19,20]

Aliphatics, including some endocyclic alkenes and straight-chain model compounds, also give considerable yields of OOMs which are present in relatively high atmospheric concentrations in an urban environment. Due to the limited laboratory experiments on aliphatics, only 63 oxidation products of aliphatics are available which are mainly C6 and C10 substances.[21] In order to reduce the possible biases due to the imbalanced number of input compounds for different precursors, we tested the case that artificially increases the number of aliphatic compounds by adding unstudied but likely existing ones that are homologous to the studied aliphatic compounds. The number of oxidation products of aliphatics was extended to 346. Details about the expansion of aliphatic OOMs and comparisons to the results without the expansion are provided in the ESI.† It should be noted that this cannot be done for isoprene OOMs, because isoprene itself has no homologous compounds.

As shown in Table 1, there are eight descriptors of oxidation products based on their chemical formula. The first four are the original information of elemental composition, *i.e.*, the number of C, H, O, and N. The other four are the processed information of the molecules: double bond equivalent (DBE) and H/C ratio which reflect the carbon saturation state of OOMs, and O/C ratio and OSc which reflect their carbon oxidation state. As

Table 1 Summary of the oxidation products of monoterpenes,[16,17] aliphatics,[21,31] aromatics,[6,19,20] and isoprene[18] under controlled laboratory conditions

		Descriptors[a,b,c,d,f]							
Precursors	Products	nC	nH	nO	nN	DBE	H/C	O/C	OSc
Monoterpenes									
α-Pinene, limonene	872	12.7	19.3	9.3	0.4	3.9	1.5	0.9	0.2
Aliphatics									
t-Decalin, decalin, cyclohexane, *n*-decane, generated[e]	346	11.4	20.5	5.8	0.2	2.0	1.8	0.6	−0.6
Aromatics									
Benzene, toluene, ethylbenzene, xylene, mesitylene	485	9.7	11.8	8.8	0.2	4.7	1.2	1.1	0.9
Isoprene	30	4.7	8.7	4.9	0.9	0.9	1.9	1.4	1.02

[a] nC, nH, nO, and nN: the number of carbon, hydrogen, oxygen, and nitrogen atoms. [b] DBE: double bond equivalent, which is calculated as $(2nC + 2nH - nN)/2$. [c] H/C and O/C: the ratio of nH over nC, and nO over nC. [d] OSc: carbon oxidation state, which is calculated as 2O/C–H/C.[32] [e] 63 reported aliphatic OOMs and 283 generated substances (adding an integer number of CH_2 groups to the reported aliphatic OOMs). [f] Values of nC, nH, nO, nN, DBE, H/C, O/C, and OSc in the table are averaged from each product with equal weight.

the VOC precursors are different in carbon number, DBE, and functional groups, their oxidation pathways and the generated OOMs are significantly different in these characteristics. For example, monoterpene OOMs and aromatic OOMs generally have higher DBE than aliphatic OOMs based on the current knowledge of their oxidation reactions.[3] We found that the minimum number of features is 3–4 including H/C and O/C. Increasing the number of features would slightly improve the performance of the models. So all the 8 features are selected.

The atmospheric OOM dataset was acquired from the measurements in Beijing (324 OOM species) and Hyytiälä (328 OOM species). The former was conducted at the BUCT-AHL site during 2018.12.26–2019.1.26, which is ~500 meters west of the Third Ring Road in Beijing with heavy traffic loading and surrounded by residential and commercial areas.[4,22] The latter was made at the SMEAR II station during 2018.3.9–2018.3.31.[9] Details of these two sites can be found in previous studies. Note that the atmospheric OOM dataset does not contain nitrated phenols since its concentration is significantly high in Beijing (~75%). In order to reduce the impact of extremely high signals,[3] we excluded nitrated phenols in this analysis.

2.3 Evaluation methods

Validation of the trained model was performed with the testing laboratory dataset. For the testing laboratory dataset, a confusion matrix was used to evaluate the consistency between the predicted results and their original labels (Fig. 1). Taking aromatics as an example, *precision* is used to characterize how many compounds that were predicted as aromatic OOMs are truly from aromatics, and *recall* is used to characterize how many aromatic oxidation products are predicted as aromatic OOMs. The overall accuracy of the decision tree is evaluated by *F1-score* (ranging from 0 to 1), which is calculated from *precision* and *recall*. Details of the evaluation metrics are given in the ESI.† Generally, we will train a series of decision trees with different number of bifurcations and then choose the most appropriate solution based on the variation of the above three indexes.

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

3 Results and discussion

3.1 Model testing with laboratory data

We first test the performance of the model with the resting laboratory data, and the testing diagnostics show that the decision tree method can effectively attribute OOMs to their likely precursors. The decision tree with bifurcations of 10 was selected as the optimal solution. The overall accuracy (*F1-score*) of the precursor apportionment using the decision tree model is 0.75, 0.73, 0.67, and 0.47 for OOMs oxidized from monoterpenes, aliphatics, aromatics, and isoprene, respectively. The

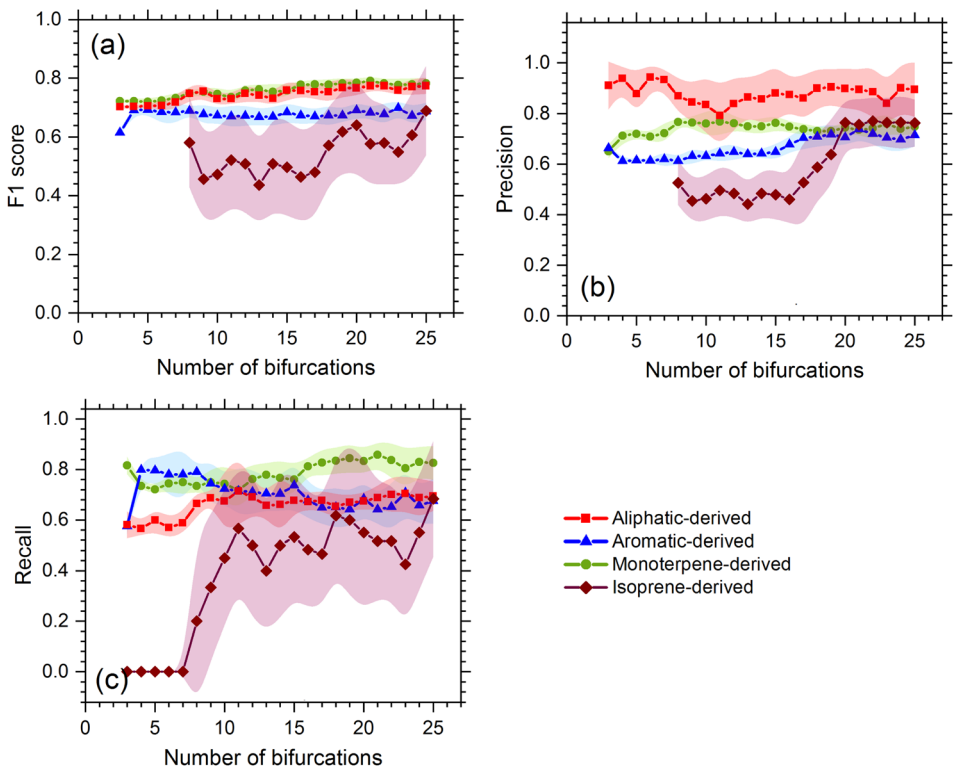

Fig. 2 The overall accuracy of the decision tree model in the precursor apportionment of atmospheric oxygenated organic molecules using laboratory testing data (a) is for *F1-score*, (b) is for *precision*, and (c) is for *recall*. The dotted lines are the mean value of repeated tests and the shaded area illustrates the standard deviation.

smallest *precision*, *recall*, and *F1-score* for isoprene are due to its small dataset, which are likely to be overcome once the dataset is expanded with more experimental results. As shown in Fig. 2, along with the increasing bifurcations, the *F1-score* of the former three precursors is relatively stable compared to isoprene OOMs. Considering *precision* and *recall* comprehensively, it is found that a bifurcation of 10 is a watershed of isoprene OOMs with the lowest cost. Decision trees with bifurcations lower than 10 may not be stable enough and those higher than 10 may be over-fitted.

3.2 Model application with atmospheric data

Fig. 3 shows apportionment results of atmospheric OOMs measured in urban Beijing and forested Hyytiälä. Since there were no labels for atmospheric data, two criteria are suggested for evaluating whether the model works for atmospheric OOMs measured in urban Beijing and forested Hyytiälä. One is that the trained model can identify the difference of atmospheric OOMs generated in these two environments with different dominant precursors, and the other is that it can identify the characteristics of atmospheric OOMs generated from the same precursor under different atmospheric conditions such as different NO_x levels. From the above two perspectives, the identified characteristics of atmospheric OOMs at these two sites by the decision tree model are satisfactory.

As suggested by the model, atmospheric OOMs in Beijing have a significant contribution from anthropogenic precursors, while those in Hyytiälä are mainly the oxidation products of biogenic monoterpenes. 37% and 27% of the OOM species in Beijing are from the oxidation of aromatics and aliphatics, respectively. As for the number concentration, the proportion of aromatic OOMs and aliphatic OOMs in Beijing is 43% and 23%, respectively. The observed dominance of aromatic and aliphatic OOMs in urban Beijing is consistent with those obtained using the workflow method[9] (Fig. S2†). The contribution from isoprene OOMs (13%) is also similar to those predicted by the workflow (10%). The non-negligible urban isoprene could come from traffic, biomass burning, and a minor contribution from vegetation.[23–25] The predicted monoterpene OOMs (21%) is much higher than those by the workflow (4%). This is partly due to the missing identification of OH-oxidized monoterpene oxidation products from the workflow.[3] In contrast, monoterpene OOMs in forested Hyytiälä dominate both in species (76%) and number concentration (62%). In addition, even for forested Hyytiälä, there are still ~33% aromatic OOMs and ~5% aliphatic OOMs in total number concentrations. Unnegligible proportions of anthropogenic OOMs in Hyytiälä have also been reported in previous PMF results on OOMs by Yan *et al.*[9] They may come from distant anthropogenic sources or from local wood combustion sources.[26–28]

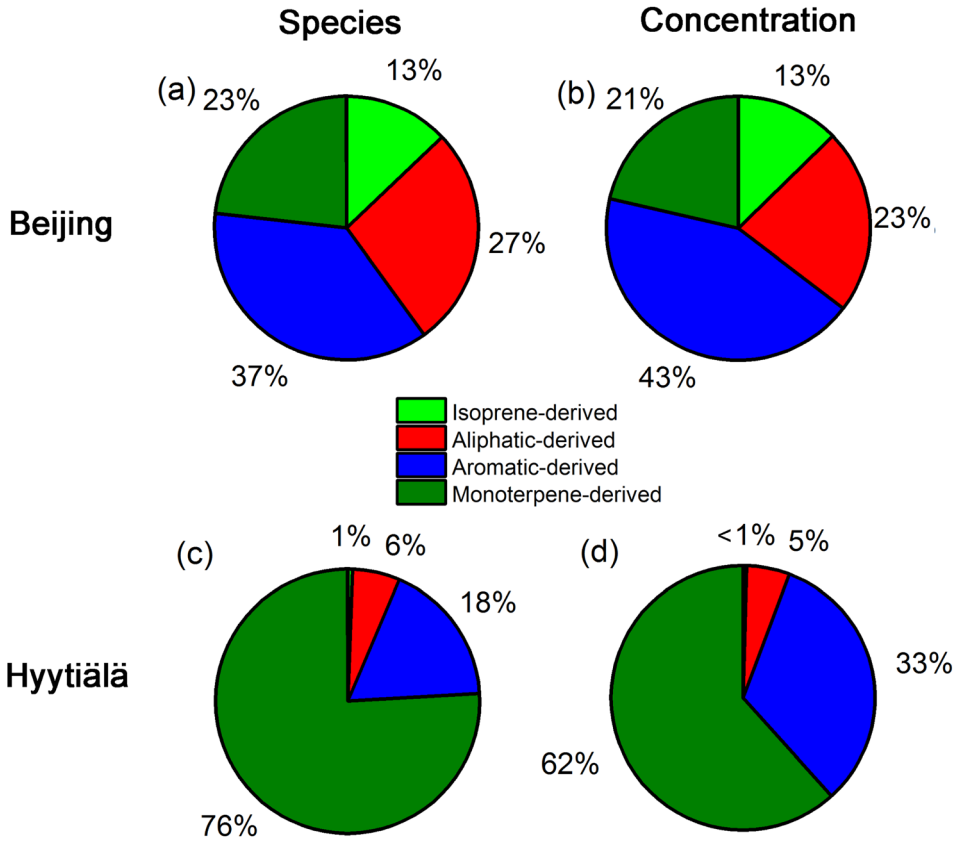

Fig. 3 The proportion of the identified oxygenated organic molecules that are oxidized from monoterpenes, aliphatics, aromatics, and isoprene in Beijing and Hyytiälä: (a) and (c) are the proportions of OOM species; (b) and (d) are the proportions of OOM number concentration.

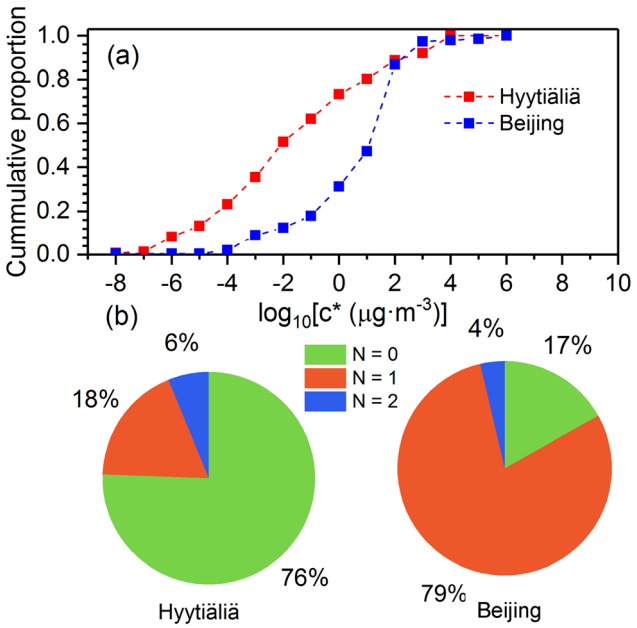

Fig. 4 The characteristics of monoterpene-derived oxygenated organic molecules (monoterpene OOMs) in Beijing and Hyytiälä: (a) the cumulative volatility distribution of the identified monoterpene OOMs; (b) the proportions of monoterpene OOMs with different numbers of nitrogen.

As shown in Fig. 4, atmospheric OOMs oxidized from monoterpenes in Beijing and Hyytiälä have different characteristics. In Hyytiälä, monoterpene OOMs consist of 76% non-nitrogen compounds, 18% compounds with one nitrogen, and 6% compounds with two nitrogens. In Beijing, 79% of monoterpene OOMs are with one nitrogen, 17% are non-nitrogen ones, and 4% have two nitrogens. Moreover, monoterpene OOMs in Hyytiälä show a lower volatility distribution compared to those in Beijing (Fig. 4a). In Hyytiälä, ~22% of monoterpene OOMs are extremely low-volatility organic compounds (ELVOCs, $C^* \leq 10^{-4.5}$ μg m^{-3}). In Beijing, only ~1% of monoterpene OOMs are ELVOCs. This could be caused by the large difference of NO_x concentrations in Beijing and Hyytiälä, which are in averages of ~30 ppb and ~2 ppb, respectively.[17,29,30] Laboratory experiments showed that high NO_x concentrations can reduce the formation of dimer products and inhibit consecutive oxygen addition in the auto-oxidation of monoterpenes, and consequently shift the volatility distribution of OOMs to the higher range.

4 Uncertainties

The main uncertainties of the current decision tree model come from the imbalanced laboratory datasets and the overlapping chemical formulae of different precursor OOMs. Currently, the laboratory dataset for isoprene oxidation is relatively small (only 30). This leads to the lowest accuracy (0.47) for the prediction of isoprene OOMs and may cause unknown uncertainties to the predicted isoprene contribution in Beijing and Hyytiälä. For reference, the datasets from aliphatics were artificially expanded from 63 to 346; this has increased the predicted aliphatic OOM contribution from 19% to 23% in Beijing, and from 2% to 5% in Hyytiälä (Fig. S4†). This indicates that the scale of the dataset did have an influence on the results, but here 63 species are already sufficient to separate the majority of alipahtic OOMs from others. As for the overlapping chemical formulae, monoterpenes and aromatics have the largest overlaps of 151 OOM formulae. This may cause an overestimation of monoterpene OOMs in Beijing where aromatic OOMs dominate, and cause an overestimation of aromatic OOMs in Hyytiälä where monoterpene OOMs dominate.

Improvement of model precision requires more laboratory results to train the model, including experiments with diverse precursors and oxidation conditions, and more information about the oxidation products, *e.g.*, the relative concentration of the oxidation products to the fingerprint molecules. As shown by the expansion example of aliphatic OOMs, a more balanced and comprehensive dataset would decrease the uncertainty of the machine learning method (Fig. S3 & S4 in the ESI†).

5 Implications

This pilot study demonstrated that the decision tree model can differentiate OOMs of different precursors under complex atmospheric conditions. This method is based on the molecular features of OOMs and has no reliance on the temporal variation of OOM concentrations. Therefore, it has advantages over the PMF method when the temporal variation of OOM concentration is not available and when the time series of OOMs from different precursors cannot be separated due to similar oxidation processes. Moreover, it's superior in processing large amounts of data, which is especially important in the era of big data obtained with the increasingly high temporal and chemical resolution mass spectrometry technology.

This work shows a vivid example of connecting laboratory data under various experimental conditions to measurements in the complex atmospheric environment. With the rapid development of analytical technologies, a large amount of multi-dimensional data with higher temporal resolution, higher spatial resolution, and higher chemical or physical resolutions are obtained. Developing data analysis methods using machine learning will largely improve data interpretation and provide an effective way of comparing and utilizing data from various collectors.

Conflicts of interest

The authors declare no competing financial interest.

Acknowledgements

Financial support from the National Science Foundation of China (92044301 and 22106083) and Samsung $PM_{2.5}$ SRP is acknowledged.

References

1 I. Riipinen, T. Yli-Juuti, J. R. Pierce, T. Petäjä, D. R. Worsnop, M. Kulmala and N. M. Donahue, The contribution of organics to atmospheric nanoparticle growth, *Nat. Geosci.*, 2012, **5**, 453–458.

2 M. Ehn, J. A. Thornton, E. Kleist, M. Sipila, H. Junninen, I. Pullinen, M. Springer, F. Rubach, R. Tillmann, B. Lee, F. Lopez-Hilfiker, S. Andres, I. H. Acir, M. Rissanen, T. Jokinen, S. Schobesberger, J. Kangasluoma, J. Kontkanen, T. Nieminen, T. Kurten, L. B. Nielsen, S. Jorgensen, H. G. Kjaergaard, M. Canagaratna, M. D. Maso, T. Berndt, T. Petaja, A. Wahner, V. M. Kerminen, M. Kulmala, D. R. Worsnop, J. Wildt and T. F. Mentel, A large source of low-volatility secondary organic aerosol, *Nature*, 2014, **506**, 476–479.

3 W. Nie, C. Yan, D. D. Huang, Z. Wang, Y. Liu, X. Qiao, Y. Guo, L. Tian, P. Zheng, Z. Xu, Y. Li, Z. Xu, X. Qi, P. Sun, J. Wang, F. Zheng, X. Li, R. Yin, K. R. Dallenbach, F. Bianchi, T. Petäjä, Y. Zhang, M. Wang, M. Schervish, S. Wang, L. Qiao, Q. Wang, M. Zhou, H. Wang, C. Yu, D. Yao, H. Guo, P. Ye, S. Lee, Y. J. Li, Y. Liu, X. Chi, V.-M. Kerminen, M. Ehn, N. M. Donahue, T. Wang, C. Huang, M. Kulmala, D. Worsnop, J. Jiang and A. Ding, Secondary organic aerosol formed by condensing anthropogenic vapours over China's megacities, *Nat. Geosci.*, 2022, **15**, 255–261.

4 X. Qiao, C. Yan, X. Li, Y. Guo, R. Yin, C. Deng, C. Li, W. Nie, M. Wang, R. Cai, D. Huang, Z. Wang, L. Yao, D. R. Worsnop, F. Bianchi, Y. Liu, N. M. Donahue, M. Kulmala and J. Jiang, Contribution of Atmospheric Oxygenated Organic Compounds to Particle Growth in an Urban Environment, *Environ. Sci. Technol.*, 2021, **55**, 13646–13656.

5 J. D. Crounse, L. B. Nielsen, S. Jørgensen, H. G. Kjaergaard and P. O. Wennberg, Autoxidation of organic compounds in the atmosphere, *J. Phys. Chem. Lett.*, 2013, **4**, 3513–3520.

6 O. Garmash, M. P. Rissanen, I. Pullinen, S. Schmitt, O. Kausiala, R. Tillmann, D. Zhao, C. Percival, T. J. Bannan, M. Priestley, A. M. Hallquist, E. Kleist, A. Kiendler-Scharr, M. Hallquist, T. Berndt, G. McFiggans, J. Wildt, T. F. Mentel and M. Ehn, Multi-generation OH oxidation as a source for highly oxygenated organic molecules from aromatics, *Atmos. Chem. Phys.*, 2020, **20**, 515–537.

7 F. Bianchi, T. Kurten, M. Riva, C. Mohr, M. P. Rissanen, P. Roldin, T. Berndt, J. D. Crounse, P. O. Wennberg, T. F. Mentel, J. Wildt, H. Junninen, T. Jokinen, M. Kulmala, D. R. Worsnop, J. A. Thornton, N. Donahue, H. G. Kjaergaard and M. Ehn, Highly Oxygenated Organic Molecules (HOM) from Gas-Phase Autoxidation Involving Peroxy Radicals: A Key Contributor to Atmospheric Aerosol, *Chem. Rev.*, 2019, **119**(6), 3472–3509.

8 P. Paatero and U. Tapper, Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics*, 1994, **5**(2), 111–126.

9 C. Yan, W. Nie, M. Äijälä, M. P. Rissanen, M. R. Canagaratna, P. Massoli, H. Junninen, T. Jokinen, N. Sarnela, S. A. K. Häme, S. Schobesberger, F. Canonaco, L. Yao, A. S. H. Prévôt, T. Petäjä, M. Kulmala, M. Sipilä, D. R. Worsnop and M. Ehn, Source characterization of highly oxidized multifunctional compounds in a boreal forest environment using positive matrix factorization, *Atmos. Chem. Phys.*, 2016, **16**, 12715–12731.

10 Q. Zhang, J. L. Jimenez, M. R. Canagaratna, I. M. Ulbrich, N. L. Ng, D. R. Worsnop and Y. Sun, Understanding atmospheric organic aerosols *via* factor analysis of aerosol mass spectrometry: A review, *Anal. Bioanal. Chem.*, 2011, **401**, 3045–3067.

11 B. Yuan, M. Shao, J. de Gouw, D. D. Parrish, S. Lu, M. Wang, L. Zeng, Q. Zhang, Y. Song, J. Zhang and M. Hu, Volatile organic compounds (VOCs) in urban air: How chemistry affects the interpretation of positive matrix factorization (PMF) analysis, *J. Geophys. Res.: Atmos.*, 2012, **117**, D24302.

12 Y. Liu, W. Nie, Y. Li, D. Ge, C. Liu, Z. Xu, L. Chen, T. Wang, L. Wang, P. Sun, X. Qi, J. Wang, Z. Xu, J. Yuan, C. Yan, Y. Zhang, D. Huang, Z. Wang, N. M. Donahue, D. Worsnop, X. Chi, M. Ehn and A. Ding, Formation of condensable organic vapors from anthropogenic and biogenic volatile organic compounds (VOCs) is strongly perturbed by NOx in eastern China, *Atmos. Chem. Phys.*, 2021, **21**(19), 14789–14814.

13 M. Mann, C. Kumar, W. F. Zeng and M. T. Strauss, Artificial intelligence for proteomics and biomarker discovery, *Cell Syst.*, 2021, **12**(8), 759–770.

14 L.-C. Wu, J.-X. Lee, H.-D. Huang, B.-J. Liu and J.-T. Horng, An expert system to predict protein thermostability using decision tree, *Expert Syst. Appl.*, 2009, **36**(5), 9007–9014.

15 T. H. Bertram, J. R. Kimmel, T. A. Crisp, O. S. Ryder, R. L. N. Yatavelli, J. A. Thornton, M. J. Cubison, M. Gonin and D. R. Worsnop, A field-deployable, chemical ionization time-of-flight mass spectrometer, *Atmos. Meas. Tech.*, 2011, **4**, 1471–1479.

16 X. Li, S. Chee, J. Hao, J. P. D. Abbatt, J. Jiang and J. N. Smith, Relative humidity effect on the formation of highly oxidized molecules and new particles during monoterpene oxidation, *Atmos. Chem. Phys.*, 2019, **19**(3), 1555–1570.

17 C. Yan, W. Nie, A. L. Vogel, L. Dada, K. Lehtipalo, D. Stolzenburg, R. Wagner, M. P. Rissanen, M. Xiao, L. Ahonen, L. Fischer, C. Rose, F. Bianchi, H. Gordon, M. Simon, M. Heinritzi, O. Garmash, P. Roldin, A. Dias, P. Ye, V. Hofbauer, A. Amorim, P. S. Bauer, A. Bergen, A. K. Bernhammer, M. Breitenlechner, S. Brilke, A. Buchholz, S. B. Mazon, M. R. Canagaratna, X. Chen, A. Ding, J. Dommen, D. C. Draper, J. Duplissy, C. Frege, C. Heyn, R. Guida, J. Hakala, L. Heikkinen, C. R. Hoyle, T. Jokinen, J. Kangasluoma, J. Kirkby, J. Kontkanen, A. Kurten, M. J. Lawler, H. Mai, S. Mathot, R. L. Mauldin III, U. Molteni, L. Nichman, T. Nieminen, J. Nowak, A. Ojdanic, A. Onnela, A. Pajunoja, T. Petaja, F. Piel, L. L. J. Quelever, N. Sarnela, S. Schallhart, K. Sengupta, M. Sipila, A. Tome, J. Trostl, O. Vaisanen, A. C. Wagner, A. Ylisirnio, Q. Zha, U. Baltensperger, K. S. Carslaw, J. Curtius, R. C. Flagan, A. Hansel, I. Riipinen, J. N. Smith, A. Virtanen, P. M. Winkler, N. M. Donahue,

V. M. Kerminen, M. Kulmala, M. Ehn and D. R. Worsnop, Size-dependent influence of NOx on the growth rates of organic aerosol particles, *Sci. Adv.*, 2020, **6**(22), 4945.

18 Z. N. Xu, W. Nie, Y. L. Liu, P. Sun, D. D. Huang, C. Yan, J. Krechmer, P. L. Ye, Z. Xu, X. M. Qi, C. J. Zhu, Y. Y. Li, T. Y. Wang, L. Wang, X. Huang, R. Z. Tang, S. Guo, G. L. Xiu, Q. Y. Fu, D. Worsnop, X. G. Chi and A. J. Ding, Multifunctional Products of Isoprene Oxidation in Polluted Atmosphere and Their Contribution to SOA, *Geophys. Res. Lett.*, 2021, **48**, 1–10.

19 U. Molteni, F. Bianchi, F. Klein, I. El Haddad, C. Frege, M. J. Rossi, J. Dommen and U. Baltensperger, Formation of highly oxygenated organic molecules from aromatic compounds, *Atmos. Chem. Phys.*, 2018, **18**, 1909–1921.

20 A. Mehra, Y. Wang, J. E. Krechmer, A. Lambe, F. Majluf, M. A. Morris, M. Priestley, T. J. Bannan, D. J. Bryant, K. L. Pereira, J. F. Hamilton, A. R. Rickard, M. J. Newland, H. Stark, P. Croteau, J. T. Jayne, D. R. Worsnop, M. R. Canagaratna, L. Wang and H. Coe, Evaluation of the chemical composition of gas- And particle-phase products of aromatic oxidation, *Atmos. Chem. Phys.*, 2020, **20**, 9783–9803.

21 Z. Wang, M. Ehn, M. P. Rissanen, O. Garmash, L. Quéléver, L. Xing, M. Monge-Palacios, P. Rantala, N. M. Donahue, T. Berndt and S. M. Sarathy, Efficient alkane oxidation under combustion engine and atmospheric conditions, *Commun. Chem.*, 2021, **4**, 1–8.

22 Y. Guo, C. Yan, Y. Liu, X. Qiao, F. Zheng, Y. Zhang, Y. Zhou, C. Li, X. Fan, Z. Lin, Z. Feng, Y. Zhang, P. Zheng, L. Tian, W. Nie, Z. Wang, D. Huang, K. R. Daellenbach, L. Yao, L. Dada, F. Bianchi, J. Jiang, Y. Liu, V. M. Kerminen and M. Kulmala, Seasonal Variation of Oxygenated Organic Molecules in Urban Beijing and their Contribution to Secondary Organic Aerosol, *Atmos. Chem. Phys. Discuss.*, 2022, **2022**, 1–33.

23 Y. Liu, M. Shao, L. Fu, S. Lu, L. Zeng and D. Tang, Source profiles of volatile organic compounds (VOCs) measured in China: Part I, *Atmos. Environ.*, 2008, **42**(25), 6247–6260.

24 B. Barletta, S. Meinardi, F. Sherwood Rowland, C.-Y. Chan, X. Wang, S. Zou, L. Yin Chan and D. R. Blake, Volatile organic compounds in 43 Chinese cities, *Atmos. Environ.*, 2005, **39**(32), 5979–5990.

25 A. Borbon, H. Fontaine, M. Veillerot, N. Locoge, J. C. Galloo and R. Guillermo, An investigation into the traffic-relatedfraction of isoprene at an urban location, *Atmos. Environ.*, 2001, **35**, 3749–3760.

26 J. Patokoski, T. M. Ruuskanen, H. Hellen, R. Taipale, T. Gronholm, M. K. Kajos, T. Petaja, H. Hakola, M. Kulmala and J. Rinne, Winter to spring transition and diurnal variation of VOCs in Finland at an urban background site and a rural site, *Boreal Environ. Res.*, 2014, **19**, 79–103.

27 J. Patokoski, T. M. Ruuskanen, M. K. Kajos, R. Taipale, P. Rantala, J. Aalto, T. Ryyppö, T. Nieminen, H. Hakola and J. Rinne, Sources of long-lived atmospheric VOCs at the rural boreal forest site, SMEAR II, *Atmos. Chem. Phys.*, 2015, **15**(23), 13413–13432.

28 H. Hellen, J. Kukkonen, M. Kauhaniemi, H. Hakola, T. Laurila and H. Pietarila, Evaluation of atmospheric benzene concentration in the Helsinki Metropolitan Area in 2000-2003 using diffusive sampling and atmospheric dispersion modelling, *Atmos. Environ.*, 2005, **39**, 4003–4014.

29 Y. Wang, M. Hu, Y. Wang, J. Zheng, D. Shang, Y. Yang, Y. Liu, X. Li, R. Tang, W. Zhu, Z. Du, Y. Wu, S. Guo, Z. Wu, S. Lou, M. Hallquist and J. Z. Yu, The formation of nitro-aromatic compounds under high NO_x and anthropogenic VOC conditions in urban Beijing, China, *Atmos. Chem. Phys.*, 2019, **19**(11), 7649–7665.

30 X. Li, Y. Li, R. Cai, C. Yan, X. Qiao, Y. Guo, C. Deng, R. Yin, Y. Chen, Y. Li, L. Yao, N. Sarnela, Y. Zhang, T. Petäjä, F. Bianchi, Y. Liu, M. Kulmala, J. Hao, J. N. Smith and J. Jiang, Insufficient condensable organic vapors lead to slow growth of new particles in an urban environment, *Environ. Sci. Technol.*, 2022, **56**(14), 9936–9946.

31 S. Wang, M. Riva, C. Yan, M. Ehn and L. Wang, Primary Formation of Highly Oxidized Multifunctional Products in the OH-Initiated Oxidation of Isoprene: A Combined Theoretical and Experimental Study, *Environ. Sci. Technol.*, 2018, **52**, 12255–12264.

32 J. H. Kroll, N. M. Donahue, J. L. Jimenez, S. H. Kessler, M. R. Canagaratna, K. R. Wilson, K. E. Altieri, L. R. Mazzoleni, A. S. Wozniak, H. Bluhm, E. R. Mysak, J. D. Smith, C. E. Kolb and D. R. Worsnop, Carbon oxidation state as a metric for describing the chemistry of atmospheric organic aerosol, *Nat. Chem.*, 2011, **3**(2), 133–139.