Supporting Information

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

Xiaohui Qiao[1], Xiaoxiao Li[1], Chao Yan[2,3,4], Nina Sarnela[3], Rujing Yin[1], Yishuo Guo[4], Lei Yao[3,4], Wei Nie[2], Dandan Huang[5], Zhe Wang[6], Federico Bianchi[3,4], Yongchun Liu[4], Neil M. Donahue[7,8], Markku Kulmala[3,4], Jingkun Jiang[1,]*

[1] State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, 100084 Beijing

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

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

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

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

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

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

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

Evaluation of precursor apportionment model

Table S1 is a confusion matrix for the prediction results. It is taken as an example to explain the evaluation of the precursor apportionment model. TP, FP, or TN is a description of the reality and predictions, in which T is for true, F is for false, P is for positive, and N is for negative. Aromatic was taken as an example in the table. Thus, when aromatics are predicted as aromatics, the result is TP; when aromatics are predicted as aliphatics and monoterpene, the result is TN; when aliphatics and monoterpene are predicted as aromatics, the result is FP. In one word, the first letter is for reality and the latter is for prediction.

Table S1. An example of the confusion matrix for a prediction.

Label\Prediction	Aromatics	Aliphatics	Monoterpenes
Aromatics	TP (🡺)	TN (🡻)	TN (🡻)
Aliphatics	NP (🡻)		
Monoterpenes	NP (🡻)		

Precision, *recall*, and *F1-score* are generally used to evaluate the performance of machine learning models. They are calculated as below,

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + TN} \quad (2)$$

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

Cross-validation

The model was cross-validated using 10-fold cross-validation, which randomly partitioning the training dataset into 10 sets, then reserving each set as validation data, and training the model using the other k-1 sets. Moreover, in this study, we evaluated the 10 models above with the same testing dataset, the 30% of original dataset that was reserved at the beginning, to present a more contrasting result (Figure 2 & S3).

Table S2 shows two examples of the determination of overlapping molecules. In example 1, ten trees give unbalanced votes, while in example 2, ten trees give balanced votes. According to the rules of voting, the molecule in example 1 is attributed to monoterpene, while the molecule in example 2 is attributed to an overlapping OOM oxidized from aromatic, aliphatic, or monoterpene.

Table S2. Examples of the voting strategy.

	Example 1	Example 2
Aromatic OOMs	3	3
Aliphatic OOMs	2	4
Monoterpene OOMs	4	3
Isoprene OOMs	1	0
Mean+Standard deviation	2.5+1.29=3.79	2.5+1.73=4.23
Mean-Standard deviation	2.5-1.29=1.21	2.5-1.73=0.77
Final	Monoterpene	Aromatic Aliphatic Monoterpene

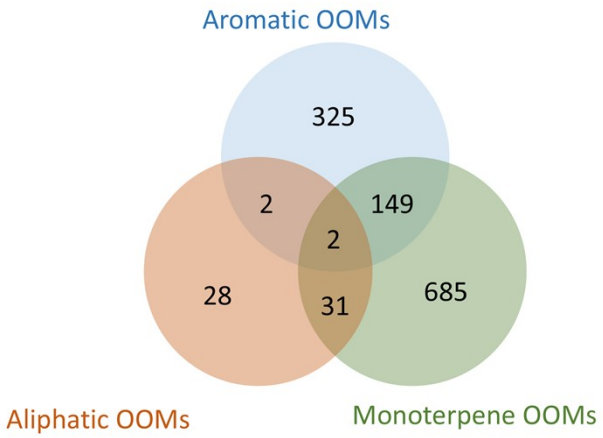

Figure S1. The Venn-plot of oxygenated organic molecules in Table 1. It should be noted that these aliphatic OOMs are without expansion.

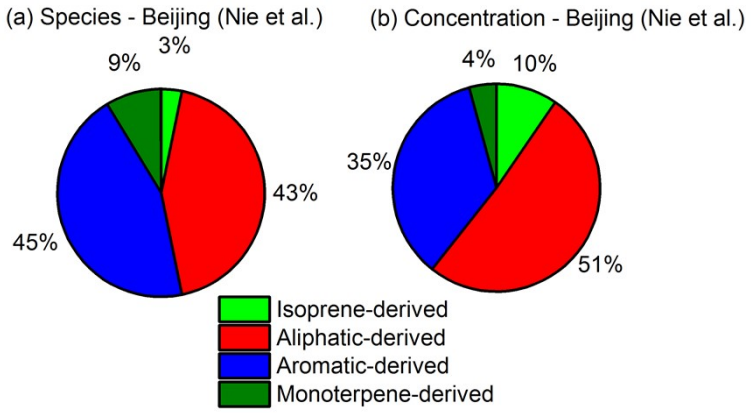

Figure S2. The apportionment result for oxygenated organic molecules in urban Beijing using the workflow method[1].

Expansion of aliphatic OOMs peaklist

In this study, we test the case that artifically construct molecules homogenous to the studied aliphatic OOMs. As shown in Table S2, we add or minus a integer number of -CH2 to those studied aliphatic OOMs (63 compounds) to obtain an enlarged peaklist of alipahtics (346 compounds). It should be mentioned that the reported aliphatic OOMs are mainly C6 and C10 compounds, we processed every molecule in the same way. Finally, there were some duplicate molecules, and only one of them was retained.

Table S3. Examples of aliphatic OOMs peaklist expansion

minus C_nH_{2n} (n=2,3)	minus CH_2 (C9)	Studied OOMs (C10)	add CH_2 (C11)	add C_nH_{2n} (n=2,3,4)
$C_{10-n}H_{20-2n}O_2$	$C_9H_{18}O_2$	$C_{10}H_{20}O_2$	$C_{11}H_{22}O_2$	$C_{10+n}H_{20+2n}O_2$
$C_{10-n}H_{21-2n}O_2$	$C_9H_{19}O_2$	$C_{10}H_{21}O_2$	$C_{11}H_{23}O_2$	$C_{10+n}H_{21+2n}O_2$
$C_{10-n}H_{20-2n}O_3$	$C_9H_{18}O_3$	$C_{10}H_{20}O_3$	$C_{11}H_{22}O_3$	$C_{10+n}H_{20+2n}O_3$
$C_{10-n}H_{21-2n}O_3$	$C_9H_{19}O_3$	$C_{10}H_{21}O_3$	$C_{11}H_{23}O_3$	$C_{10+n}H_{21+2n}O_3$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$C_{10-n}H_{21-2n}NO_3$	$C_9H_{19}NO_3$	$C_{10}H_{21}NO_3$	$C_{11}H_{23}NO_3$	$C_{10+n}H_{21+2n}NO_3$

The impact of data expansion

There is an increase in the overall accuracy of decision trees when comparing the evaluation result of the model trained with the original dataset (Figure S5) and that trained with the expanded dataset (Figure S2a). The *F1-score* of aliphatic OOMs increased from ~0.3 to ~0.5. For aromatic OOMs and monoterpene OOMs, however, it did not change evidently. Although the accuracy of identifying isoprene OOMs has decreased, the overall accuracy for these four precursors has improved generally.

Moreover, the application results showed that the proportion of monoterpene OOMs in Beijing decreased ~1% and that of aliphatic OOMs increased by ~4% when using the model trained with the expanded dataset in comparison to that trained with the original dataset. For Hyytiälä, monoterpene OOMs increased 3%, aromatic OOMs decreased 3% when using the model trained with the expanded dataset. The proportion of aliphatic OOMs increased and that of isoprene OOMs decreased.

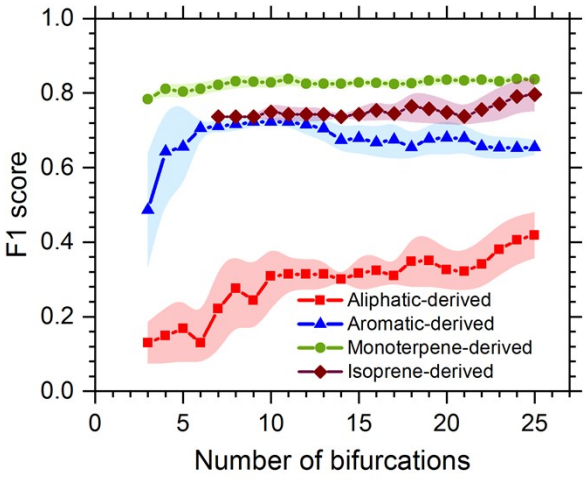

Figure S3. The overall accuracy of the decision tree model in precursor apportionment of atmospheric oxygenated organic molecules using laboratory dataset without the expansion of aliphatic OOMs.

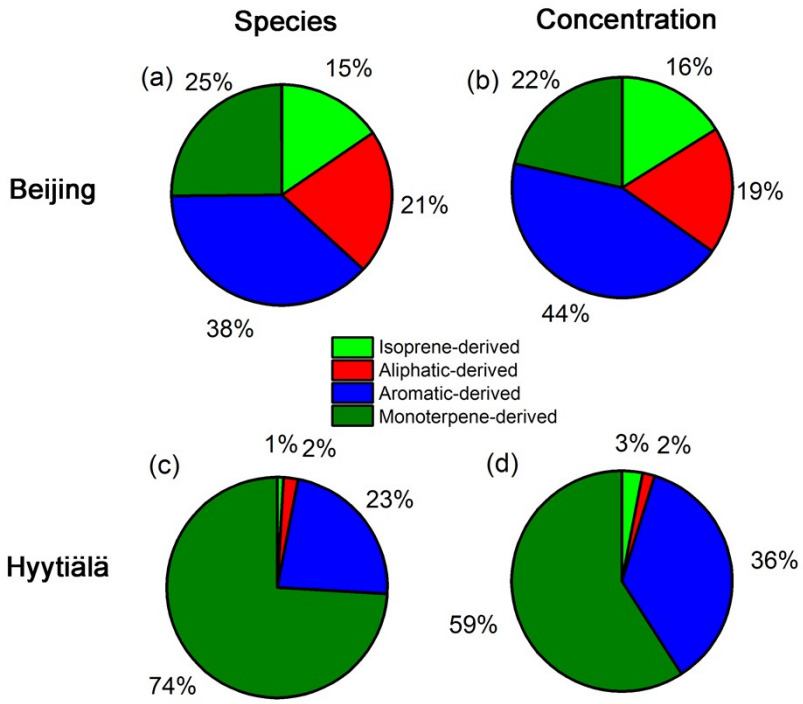

Figure S4. The proportion of the identified oxygenated organic molecules that are oxidized from monoterpenes, aliphatics, aromatics, and isoprene in Beijing and Hyytiälä predicted by model trained with laboratory dataset without the expansion of aliphatic OOMs: (a) and (c) are the proportions of OOM species; (b) and (d) are the proportions of OOM number concentration.

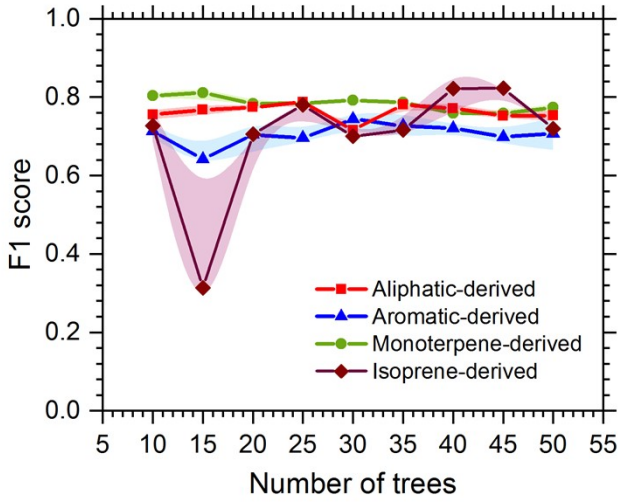

Figure S5. The *F1-score* of models consisting of different number of trees.

Along with the increasing number of trees, the *F1-score* of the trained model fluctuates without significant improvement. Thus, ten trees was adopted in this study. The performance of models with more trees may be limited by the imbalance training dataset, which can be improved by supplementing with more comprehensive laboratory data.