Supplementary Information

1. Estimation of uncertainty from laboratory experiments

The detection uncertainty of data collecting from the CI-APi-TOF need to be studied or re-examined, as the ion detection of this instrument is more complicated than other instruments such as AMS. The wide detection range CI-APi-TOF (normally ~3000 Th) on the one hand allows us to detect larger ion clusters; on the other hand, it is hard to guarantee an optimized transmission all over the detection range. This can lead to a significant change of the ion detection efficiency, which may in turn influence the signal background as well as the signal intensity. Apart from that, the *a* value (Eq. 6) used for AMS data may not be applicable for data from CI-APi-TOF. Thus, a set of laboratory experiments were conducted to find out a proper equation to describe the detection uncertainty.

The schematic of the experiment setting was as shown in Fig. S1. A temperature controlled permeation sources were connected to the CI-inlet. Nitrogen gas (N_2) was used as both the carrier gas and the dilution air. The optimized flow rate for N_2 flowing through the permeation source was found to be 100 slpm, which ensured that there were enough permeated chemicals being carried out without generating large turbulence that may cause additional loss. The outflow of the permeation source was then mixed with another N_2 flow, which is controlled by vacuum line (30 lpm) together with synthetic air (20 lpm).

The setting of the instrument mentioned above was kept identical throughout the experiments, however, two different chemicals ($CF_3(CF_2)_2COOH$ and $CF_3(CF_2)_7COOH$) were used as the permeation source. The temperature range for them were 20~60 °C and 30~85 °C, depending on the volatility of the chemicals. Moreover, experiment with the same chemical was repeated twice with different instrument tunings, which were tuned to have optimized transmission in low-mass range (<200 Th) and high-mass range (>800 Th), respectively.

Fig. S2 shows an example of signal variation caused by temperature change, with all peaks in the spectra. Firstly, based on Eq. 5 the background need to be fitted. Different from AMS measurement, CI-APi-TOF was running without a routine background measurement. As an alternative, we used the "blank masses", where few peaks are located, to estimate the background, with an assumption the net noise is independent of the transmission (detection efficiency) and thus remains the same over the whole mass range. Fig. S3 shows the background estimations with low-mass setting, high-mass setting, and the setting we used in the ambient measurement. The background for all tunings in the 'blank mass' (800~1000 Th) is estimated to be 0.035, except that some large molecules or clusters can still be observed in the high-mass tuning, resulting in some discrete outliers. Note that, the constancy of background in different tunings also confirms the validity of the pre-assumption.

The *a* value in Eq. 6 can be then fitted from the analytical uncertainty to the signal strength. Fig. S4 shows the correlation between uncertainty and signal intensity, counting only major peaks in the spectra with different permeation sources and different instrument tunings. In general, the fitting of uncertainties in all experiments follows

the same trend, implying an independence of the uncertainty on both chemical species and instrument conditions, over a large range of signal intensities between 0.1~10000 cps (count per second). Since the strongest signal in the ambient measurement is about 20 cps, we fitted the uncertainty only with peaks below this value including isotope peaks. As shown in Fig. S5, the best fitted value for $a/\sqrt{t_s}$ is found to be 0.074±0.005 (corresponding to the upper and lower bounds of 95% confidence), and corresponding a value for 5 min averaged data ($t_s = 300$) is 1.3±0.1.

2. **Estimation of uncertainty from ambient measurement data**

To assess if the uncertainty derived from the laboratory experiment agrees with what we observe in actual measurement of ambient air, we devised a simple technique to estimate the instrumental noise based on the ambient air data independently. We tested the technique on the same data input for PMF, containing 9084 measured time steps and 450 variables from 201 – 650 Th.

The basis of the method is approximating instrument noise as the difference of measured signal (in unit cps) relative to the signal's moving median over a short period of time (5 data points). Assuming changes in the chemical composition happen generally in a longer timescale than the timescale of measurement (5 minutes), we can consider the deviation from the moving median to result mostly from the uncertainty of the measurement rather than actual chemical changes in the aerosol. However, as some of the deviation undoubtedly arises from actual variation in the sample, we consider this estimate to represent the upper limit of instrument noise. To avoid possible contamination peaks, that would yield very high positive difference, from being interpreted as sigh instrument noise, causing potential overestimation of instrument uncertainty, we filtered out highest 10% (in cps) of observations for each ion separately.

We would expect all the signals at various different m/z ratios behave similarly, but as the selection of a specific signal in the ambient air to represent variability of all the data may be problematic, due to very different dynamic ranges of the signals, we decided to perform the test for all available m/z. This also allows us the broadest set of observations to work with and should minimize any conceivable biasing effects of using a potentially non-representative signal.

We chose to study the noise dependence on signal level, by dividing the "noise estimate" (i.e. signal minus trend) data into bins, each bin representing a part of the ambient-air-relevant signal range. I.e. a bin containing the "noises" observed, for ion "*i*", when the ion's signal is between the bins limiting values S_L and S_H [cps]. To cover the entire ambient air relevant dynamic range of signal, we defined the upper and lower signal limits for the bins dynamically. S_L and S_H of each bin was derived by dividing the observations for each ion to signal deciles. Now for each ion *i* we would have bins $S_1 – S_9$, S_1 corresponding to the noise when signal is within the lowest decile (0-10%) and S_9 to the noise associated with the very highest signals (decile 9; 90-100%).

By this we reduced data dimensionality from the original 1000x9084 data matrix to a 1000x10 matrix, now corresponding to 10 bins for each of the 1000 ions. Each bin yielding approximately 940 observations of the instrument noise. We then quantified the 'instrument noise' or 'uncertainty' related to each of the 10,000 bins, individually, by

assuming the deviations are normally distributed, and fitted for each bin a normal distribution, extracting the fit parameters, mean μ and standard deviation σ with their 95% confidence intervals (see example in Fig. S6). To reduce data, we henceforth use the standard deviation σ as a single parameter measure of the noise, effectively representing the bin contents of over 900 observations with the fitted distributions, represented by the two parameters and their confidence limits. We would expect the distribution mean μ to be zero, which it generally conforms to within the limits of uncertainty.

[Note on the mathematics: Strictly speaking the distribution would be a superposition of a normal distribution (electronic noise) and a Poisson distribution (counting error). Unfortunately resolving this would be mathematically and computationally exceedingly complex, and we instead take advantage of the fact that the shape of a Poisson distribution closely approaches that of a bell curve for sufficiently large number of occurrences (here: ion counts, signal intensity), while for very low counts (signal) the normal distribution (electronic noise) anyway dominates the superposition, as the of counting error magnitude is negligible when number of counts is close to zero. With the approximation we deal with superposition of two bell curves instead, a summation which actually is normally distributed – hence the fit should be well justified.]

Having much simplified the situation, we are now left with ten standard deviation values σ per ion *i*, one for each signal range decile. We then want to parametrize the noise's dependence on signal, which we do by constructing a (weighted non-linear) least squares fit, modeling the observed noise with a two parameter (constant electronic noise *e* and the square root function of signal $a\sqrt{s}$) function,

$$f(a,e) = a\sqrt{s} + e \quad \text{(Eq. S1)}$$

where *e* and *a* are constant parameters, s is the signal at the bin middle point). We also supply for the fitting algorithm (Matlab curve fitting toolbox) the uncertainty associated with σ, obtained from the Gaussian fit, to be used as a (inverse) weight when determining the best fit. Three examples of such fits for ions (339 Th, 340 Th, and 555Th) are given in Fig. S7. From this second fit the parameters *e* and *a* are determined again with their uncertainties at 95% confidence level.

Parameter *e* can now be directly understood as the electronic noise of the instrument, assumed to be constant (relative to signal variation). Parameter *a* is similar the *a* in the Allan et al. (2003) equation (eq. 6), and defines the square root dependence constant. Any fits with clearly non-physical outcome (such as negative *a* or *e*, or clear outliers outside of two standard deviations from the mean) were excluded. Taking the mean (weighted by the inverses of their uncertainties) of the parameters *e* and *a*, over all the ions, we obtain the final e (DL) and a values to be used as in the parameterization of the total uncertainty, as derived from the ambient data. The upper and lower limits for the estimate are obtained for the estimate using propagation of error, after which the uncertainty associated with the final error estimate be written:

$$\Delta f(a,e)\sqrt{(\sqrt{s}\,\Delta a)^2 + (\Delta e)^2} \quad \text{(Eq. S2)}$$

where $f\ (a,e)$ is the error estimate parameterization function from Eq. S1, and Δa and Δe are the respective 95% confidence level uncertainties for a and e. The final result depicted in Fig. 1 for the ambient air data noise estimate was thus:

$$f = a \pm x\sqrt{s} + e \pm x \quad \text{(Eq. S3)}$$

with the total error calculated from Eq. S.2.

3. Examining Q distribution of time and variables

Fig. S8a shows the Q distribution over variables in 6-factor PMF solution (the optimal solution, see Section 4.1&4.3), together with the average signal to noise ratio (SNR). The mean value of Q on all variables was well below 4, the threshold in robust-mode PMF. This suggests that all variables are well described by the model. Fig. S8b illustrates the Q distribution over samples in 6-factorial solution, where Q distribution in 2-factor solution is also plotted as a reference. The shaded area denotes the period when the location was influenced by continuous transported pollution. In both solutions, Q does not exhibit an elevation in transported pollution period, suggesting that this transported pollution event can be equally described by the model. However, comparing to the result in 2-factor PMF, Q/Q_{exp} in 6-factor solution is systematically lower in all samples. Especially for the high Q/Q_{exp} value shown in 2-factor PMF, using 6 factors significantly reduce the error, showing an improvement of the model performance.

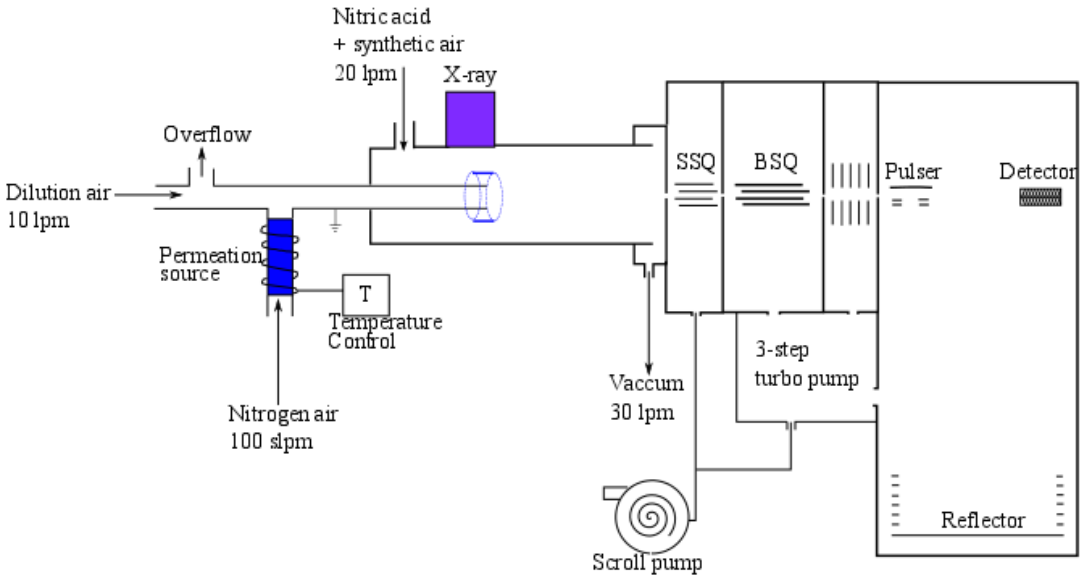


Fig. S1. The schematic of the laboratory experiment assembly. All the flows were set identical throughout the experiments, while different chemical, temperature and instrument tuning were tested.

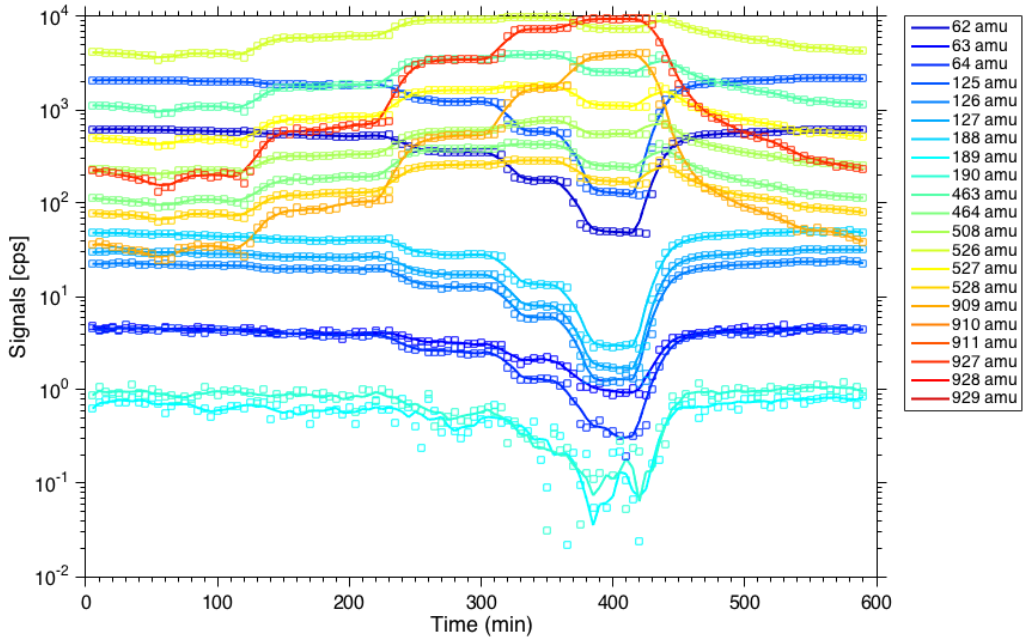


Fig. S2. An example of signal variations at different temperatures in the experiment using $CF_3(CF_2)_7COOH$ and high-mass tuning. The temperatures increased stepwise (i.e. 30, 40, 50, 60, 70, and 85 °C), and the signals showed stepwise change simultaneously. For further error fitting (Fig. S4 and Fig. S5), only steady-state data were used.

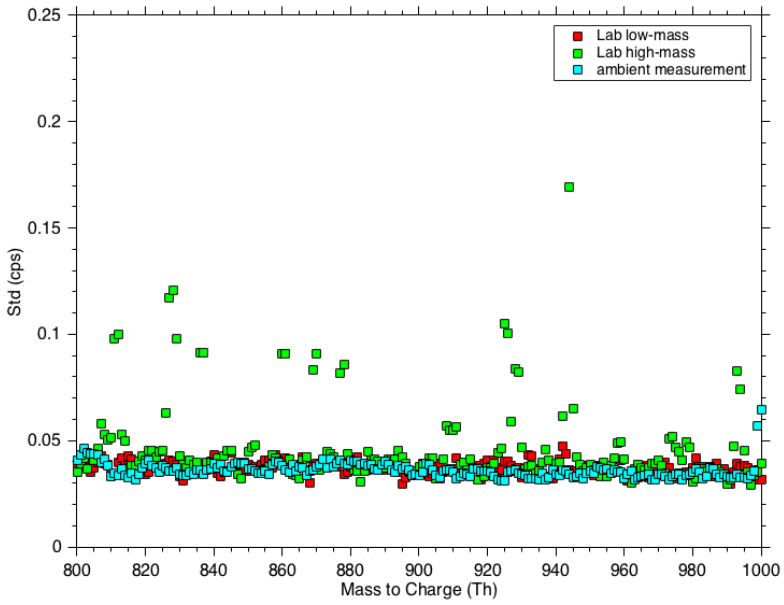
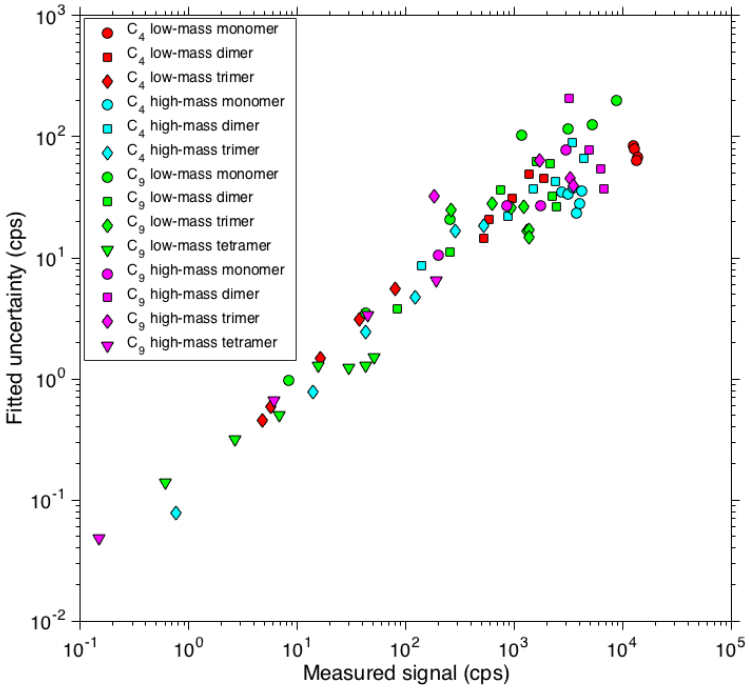


Fig. S3. Background estimation for data from low-mass tuning (red), high-mass tuning (green), and tuning for ambient measurement. 800~1000 amu was selected as the 'blank mass' though some peaks can be observed in high mass tuning. The background for all tunings shows a good agreement, indicating that the net noise level is that same for the whole mass range.



141

142 **Fig. S4**. The analytical uncertainty versus signal strength for different chemicals and instrument tunings. Different
143 combinations of a certain chemical and a certain tuning are marked with different color. Within each combination,
144 different shapes are used to mark different chemical oligomers or the reagent ions.

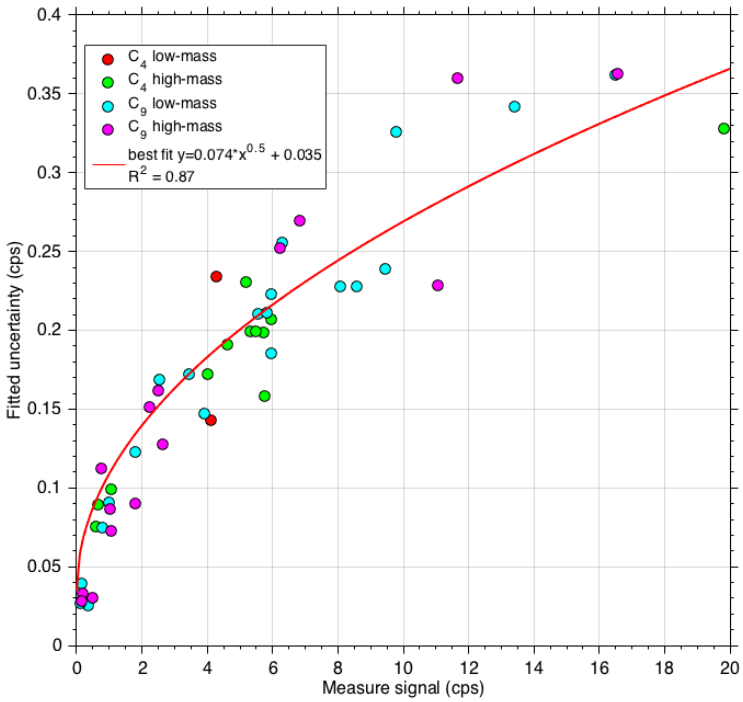


Fig. S5. Fitting of uncertainty versus signal based on Eq. 6. Only signals smaller than 20 cps (in the typical atmospheric level) were used. The same color code was used as in Figure S4.

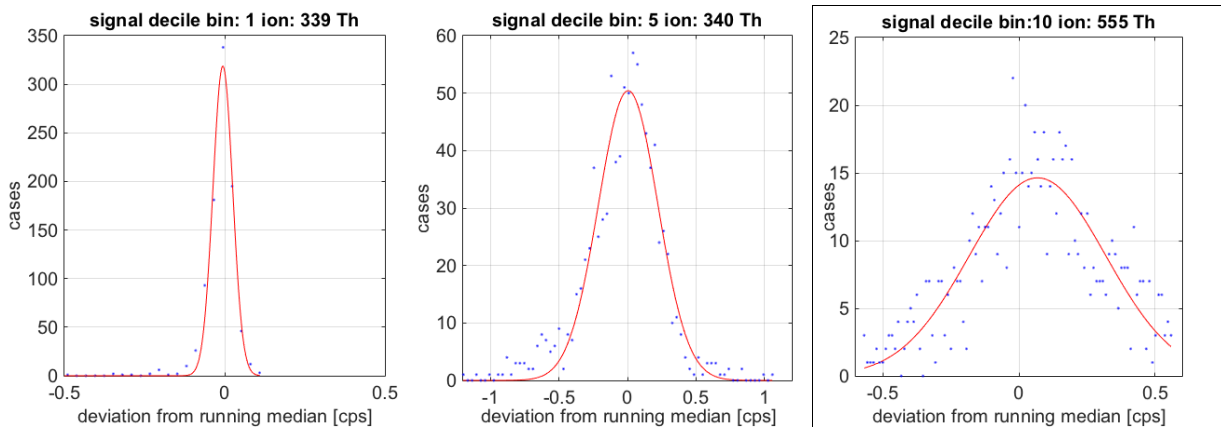


Fig S6. Examples of a histogram of the deviations between ion signal and the five point moving median for ions at 339 (1st signal decile) , 340 (5th decile) and 555 (10th decile) Th . The median points (difference = 0) are excluded. Also shown are the least squares Gaussian fits, from which the standard deviation σ (along with its uncertainty) is extracted.

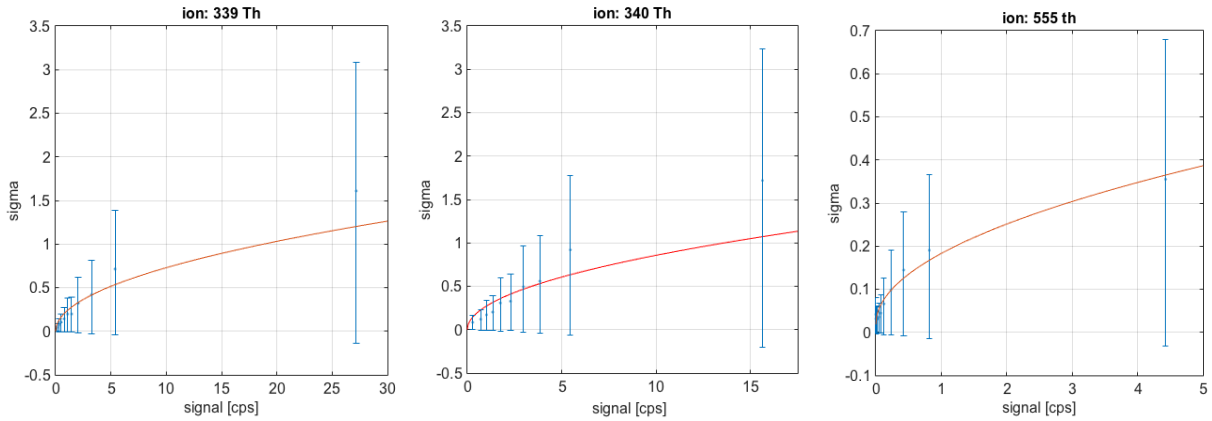


Fig. S7. The normal distribution (see Figure S6) standard deviations and their 95% confidence limits associated with the nine signal bins of ions at 339, 340 and 555 Th. The best (weighted non-linear least squares) fit for $a\sqrt{s} + e$ is shown in red, depicting our model for the error's (σ) signal dependence.

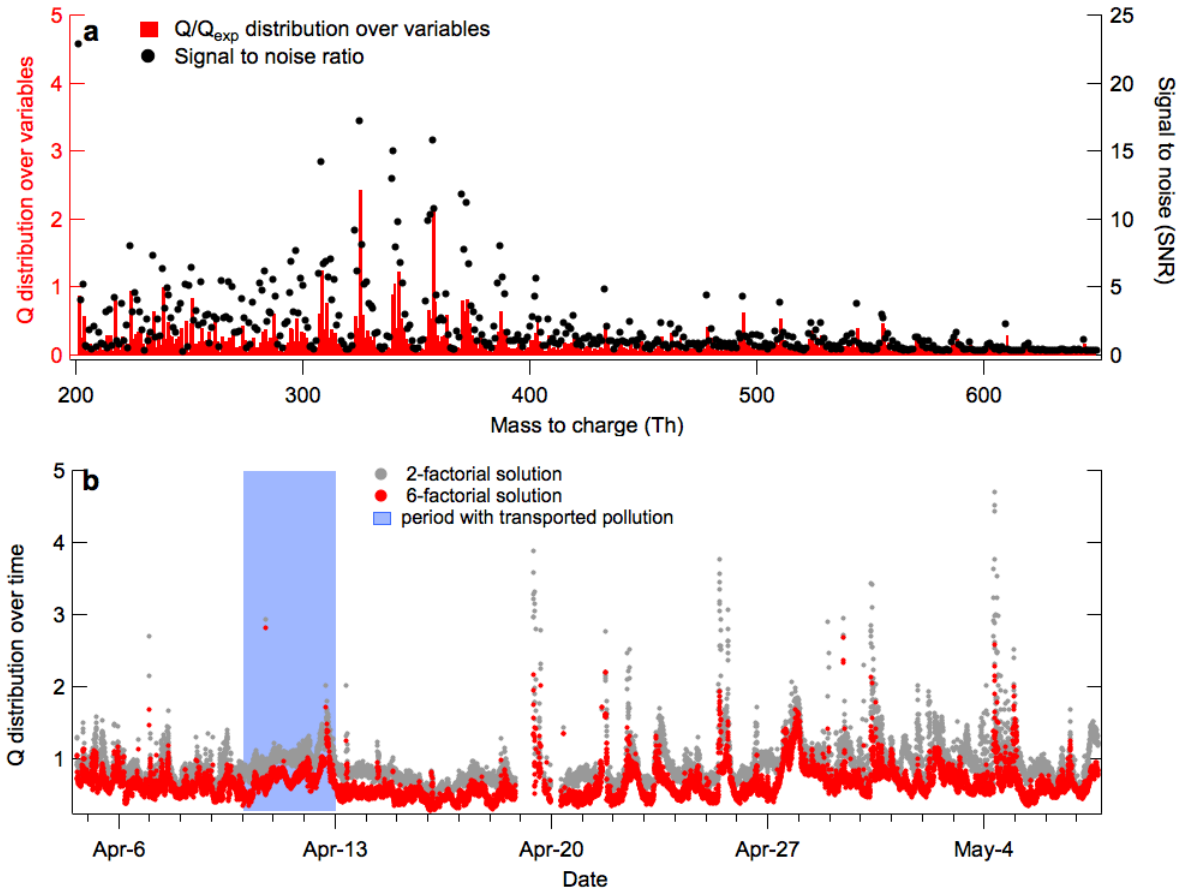


Fig. S8. (a) Distribution of Q/Q_{exp} on variables (m/z, red bars) and average signal to noise ratio (SNR, black dots) of those variables. (b) Distribution of Q/Q_{exp} on samples: in 2-factor and 6-factor solutions. Blue shaded area denote the period with transported pollution.

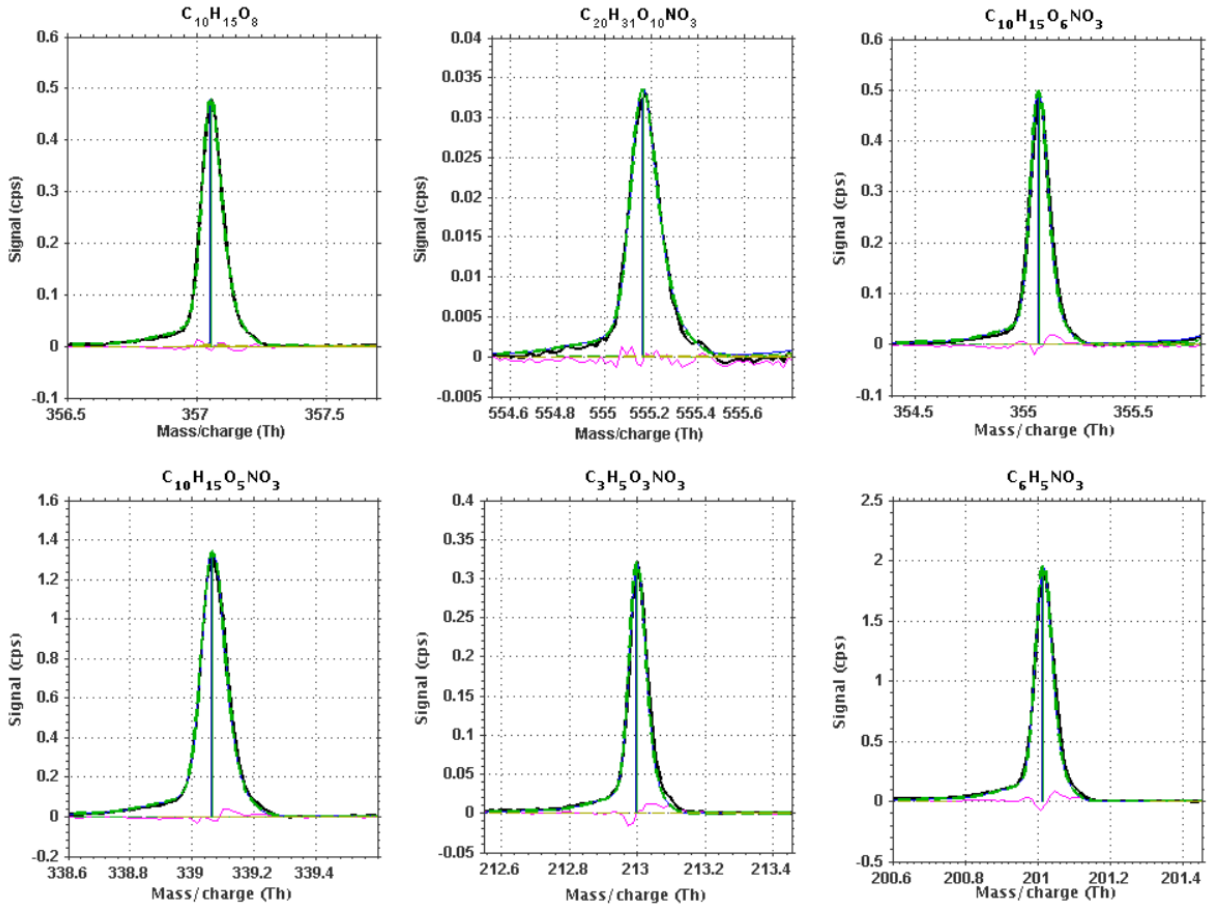


Fig. S9. Examples of peak fitting. The black solid line is the measured signal, the green dashed line denotes the fitted peak, and the purple one is the residue. The six examples correspond to the fingerprint molecules chosen from the 6 factors (marked with * in Table 1).

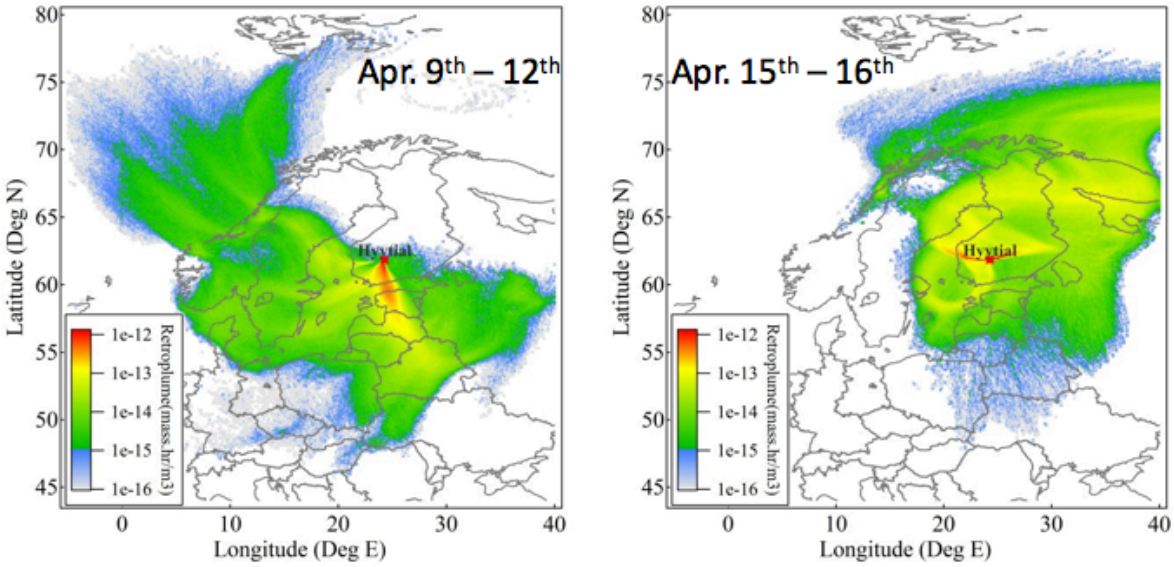


Fig. S10. Air mass analysis using backward Lagrangian particle dispersion model (LPDM). The shown results are based on 500m altitude calculation. The plot on the left shows that air masses were mainly from Eastern Europe on Apr. 9th – Apr.12th, while the plot on the right shows that air masses were from Northern Europe on most other days, for example Apr. 15th – Apr. 16th.

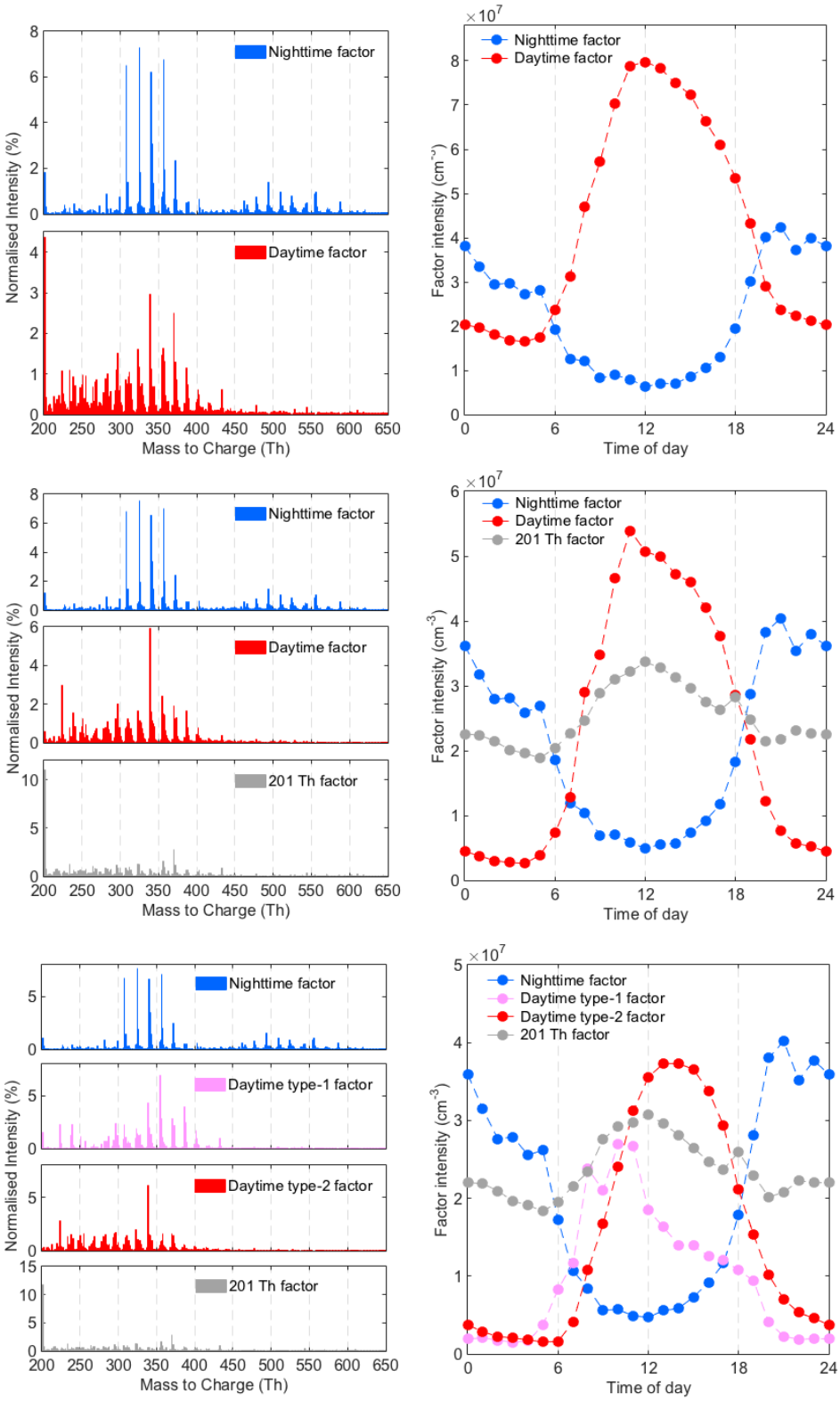


Fig. S11. Profile (left panels) and diurnal variation (right panels) of PMF factors. The top panels show the 2-factor case, the mid panels denote the 3-factor case, and the bottom panels demonstrate the 4-factor case.

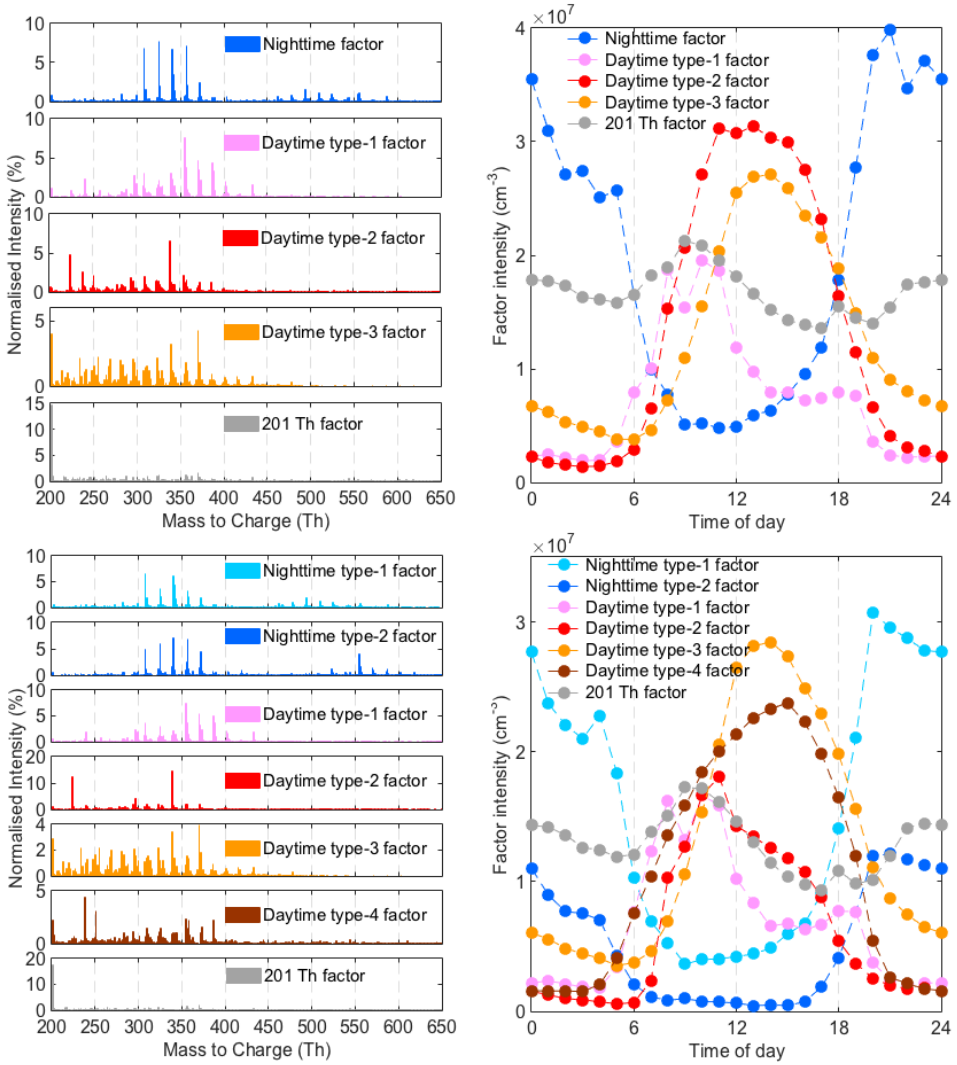


Fig. S11 (continued). Profile (left panels) and diurnal variation (right panels) of PMF factors. The top panels show the 5-factor case, and the bottom panels demonstrate the 7-factor case. Note that the optimal solution with 6 factors are shown in Fig. 5 and Fig. 6.