Supplementary information for Dimensionality-reduction techniques for complex mass spectrometric datasets: application to laboratory atmospheric organic oxidation experiments

Abigail R. Koss[1], Manjula R. Canagaratna[2], Alexander Zaytsev[3], Jordan E. Krechmer[2], Martin Breitenlechner[3], Kevin Nihill[1], Christopher Lim[1], James C. Rowe[1], J. R. Roscioli[2], Frank N. Keutsch[3], Jesse H. Kroll[1]

[1] Massachusetts Institute of Technology, Department of Civil and Environmental Engineering, Cambridge, MA

[2] Aerodyne Research Incorporated, Billerica, MA

[3] Harvard University, Paulson School of Engineering and Applied Sciences, Cambridge, MA

Correspondence to: Abigail Koss (abigail.r.koss@gmail.com)

Figure S1

A. Kendrick mass defect plot of unambiguously identified ions (Vocus-2R-PTR instrument).

B. Kendrick mass defect plot of all ions.

Markers are sized and colored by peak area. In subplot A, a line has been drawn through large, unambiguously identified peaks $C_9H_{13}O_n^+$ with *n* between 1 and 4. In subplot B, the series has been extended to include *n*>4. The identities of other peaks with *m/z* >200 were suggested in a similar way, by identifying trends in ion formulas with m/z<200 and extending the series to larger *m/z*.

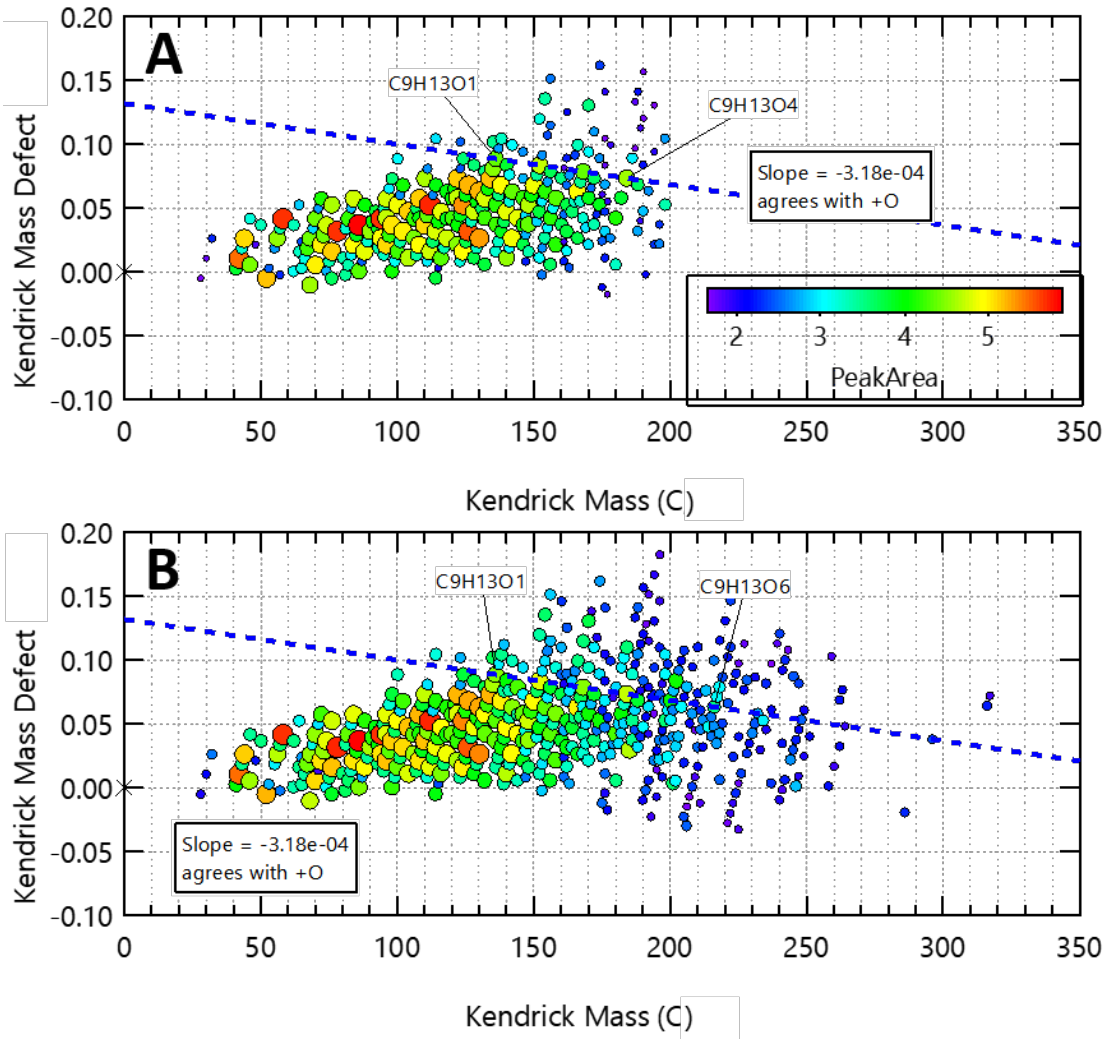


Figure S2

Signal-to-noise ratios for the synthetic data system (left) and chamber data (right). For PMF, species with SNR<2 are downweighted by a factor of 2, and species with SNR<0.2 are downweighted by a factor of 10.

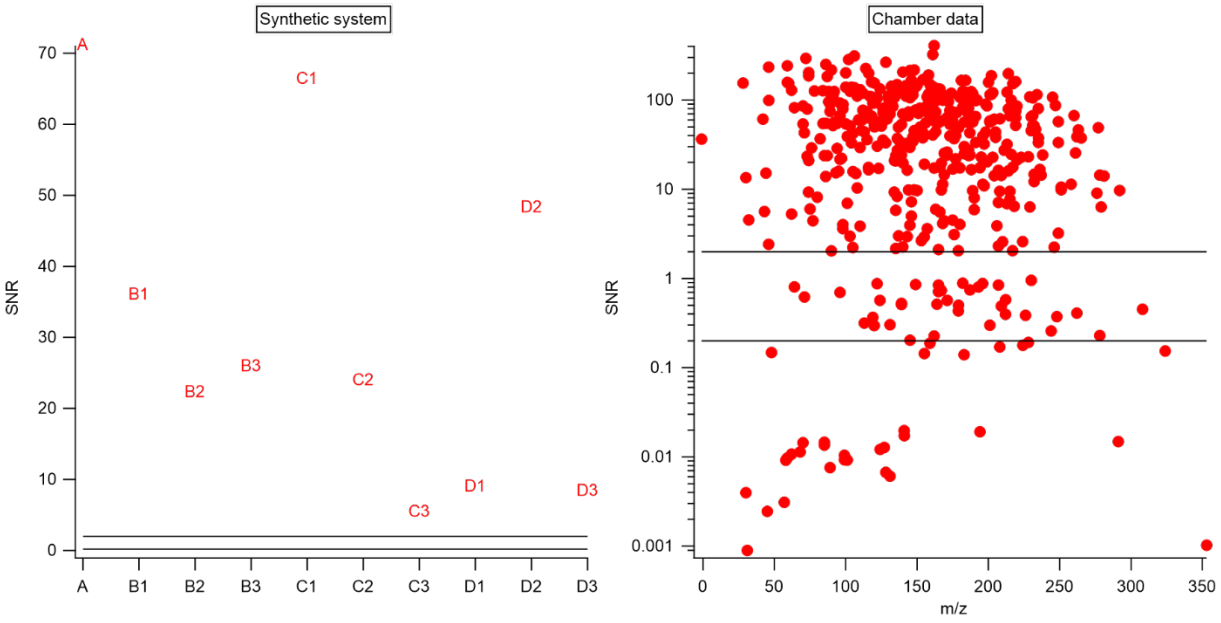


Figure S3

Relationship between standard deviation and signal for all chamber measurements.

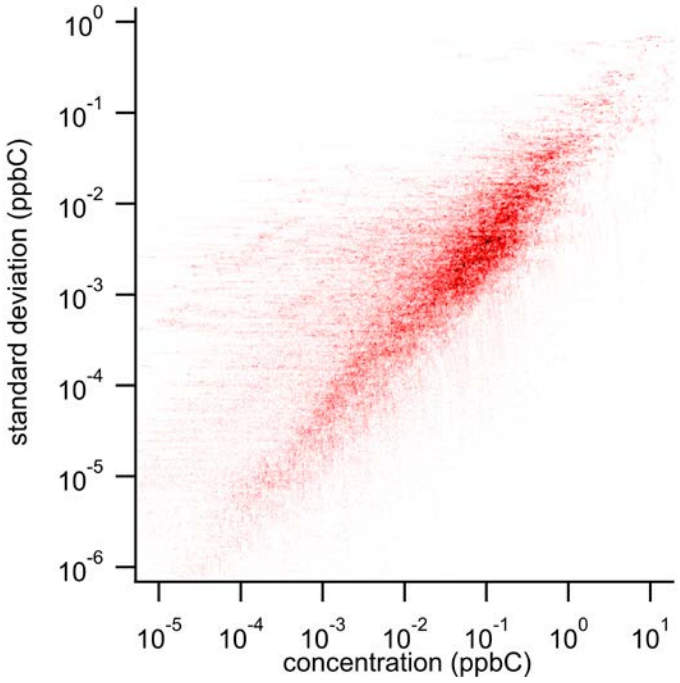


Figure S4

Time series and factor profiles of PMF analysis of synthetic data

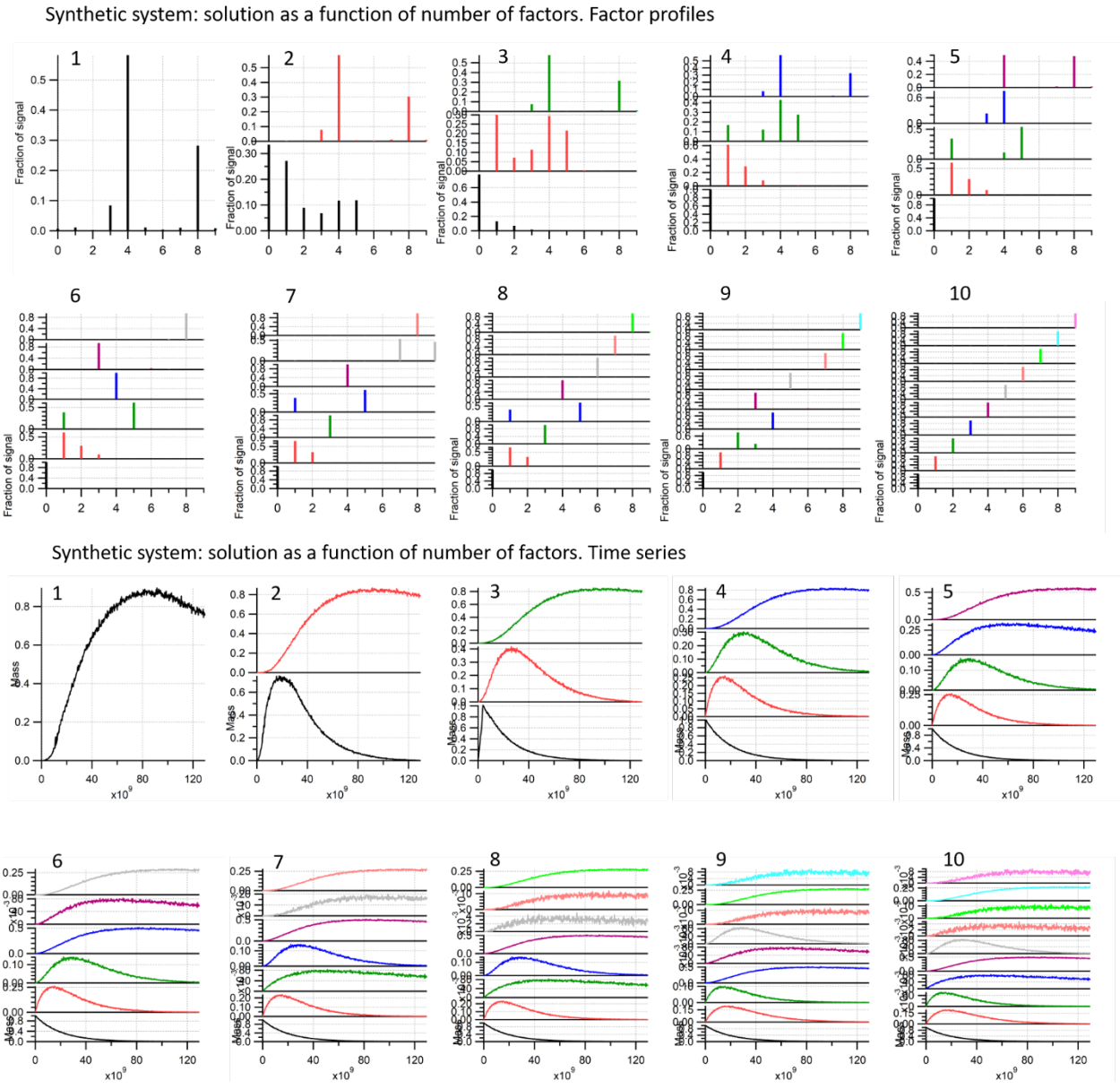


Figure S5

PMF results for chamber data. Three- to eight-factor solutions are shown. No solution was found for two factors.

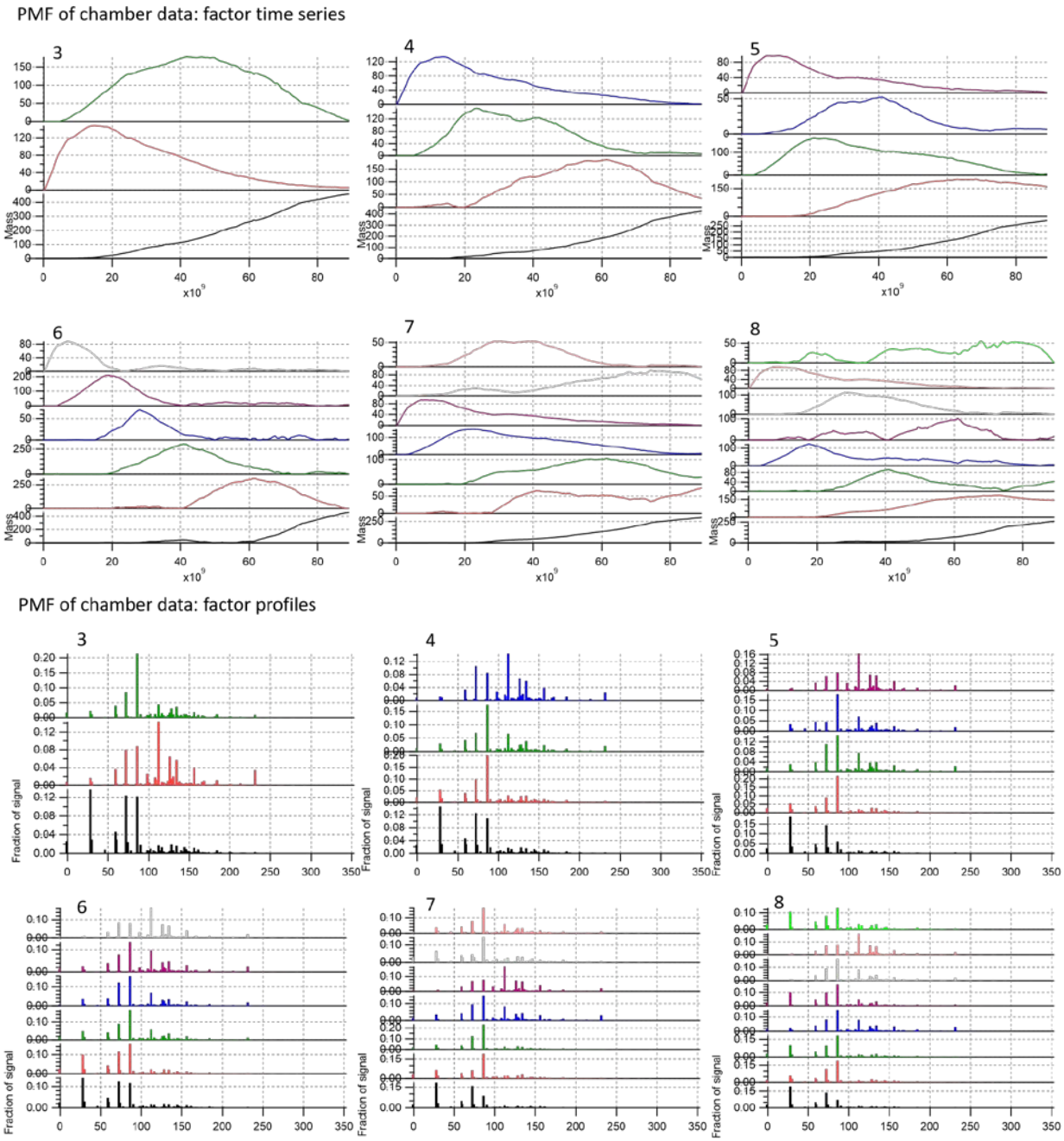


Figure S6

Time series of all clusters and individual species from HCA analysis. Individual species are shown as thin lines. Cluster averages are shown as thick lines, and the individual species contributing to that cluster are included as thin gray lines. In each plot, the y-axis is normalized intensity and the x-axis is OH exposure.

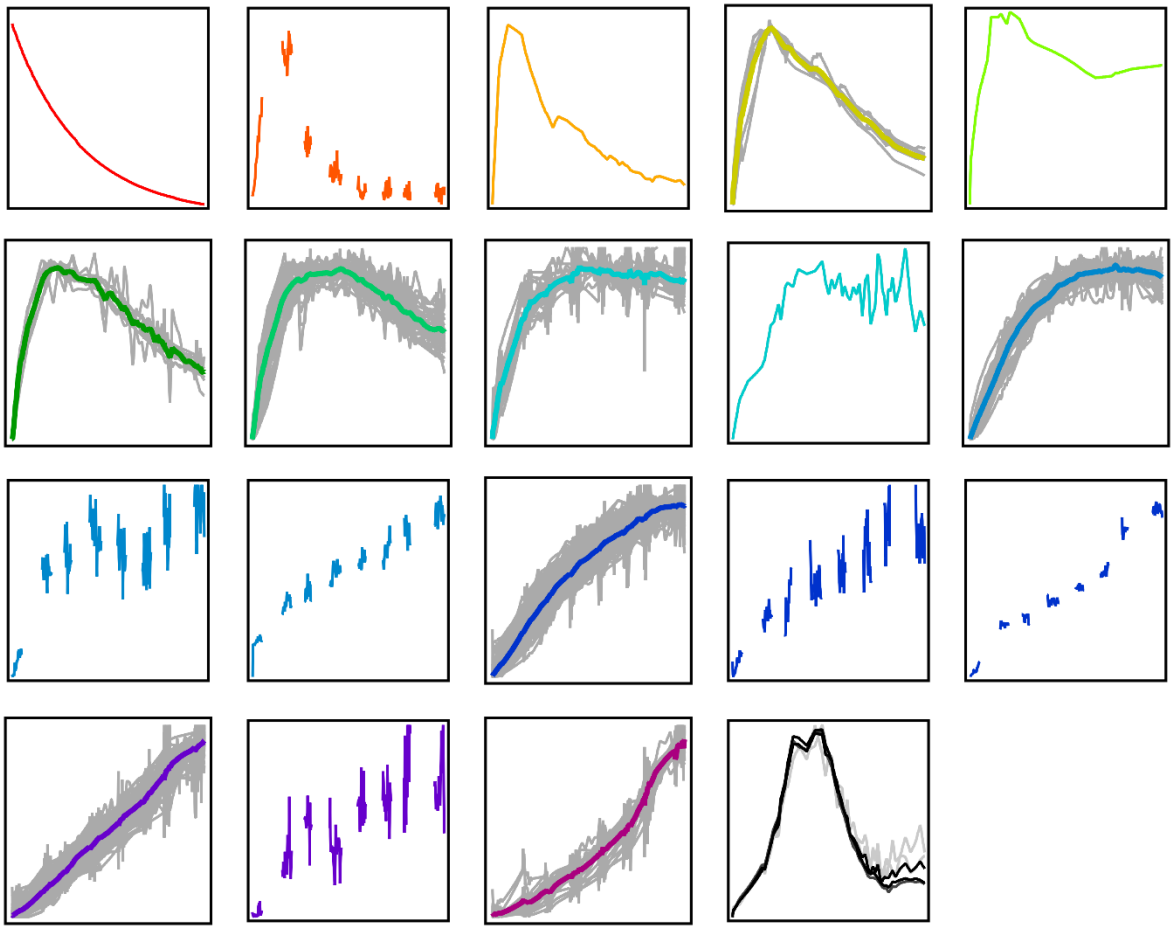


Figure S7

Clustering results at different relative distances.

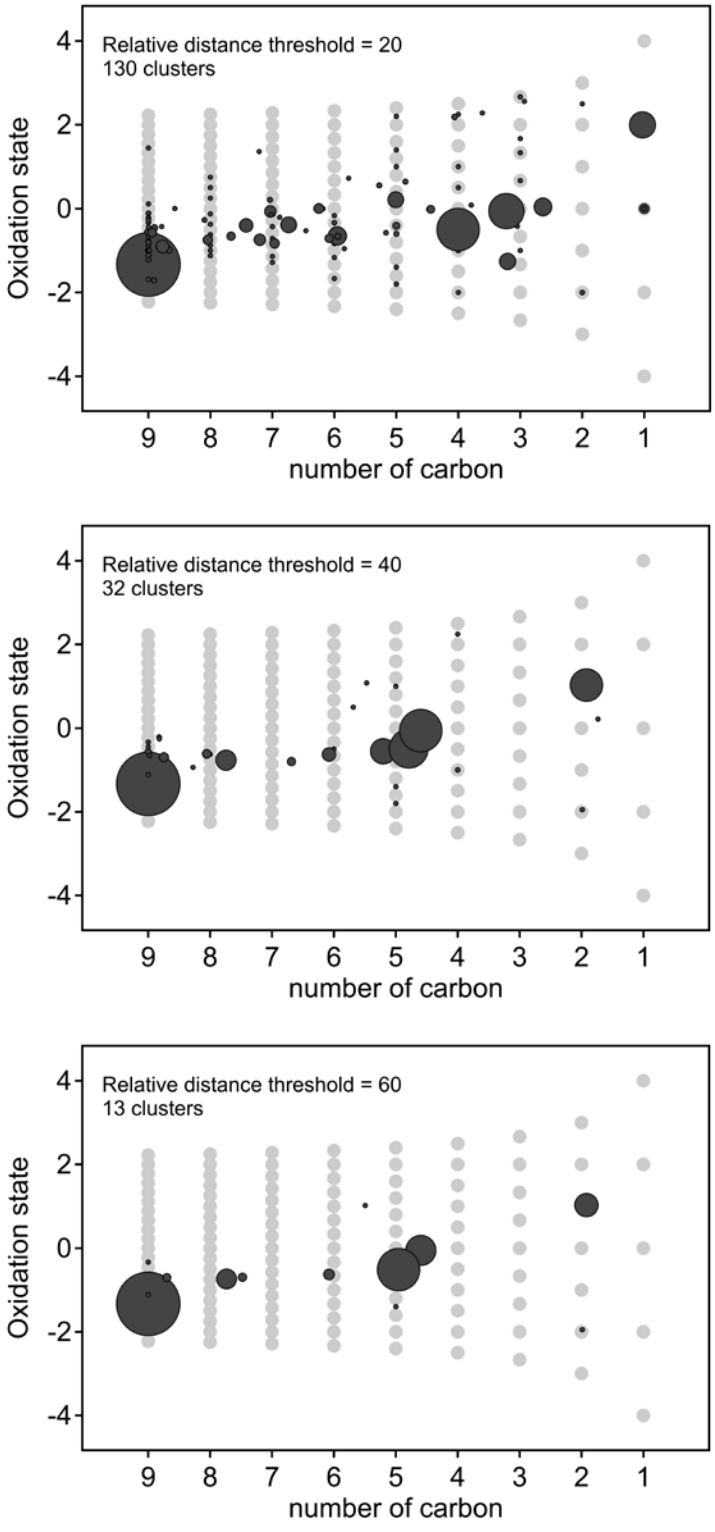


Figure S8

Standard deviation of fit parameter for *m* and *k.*

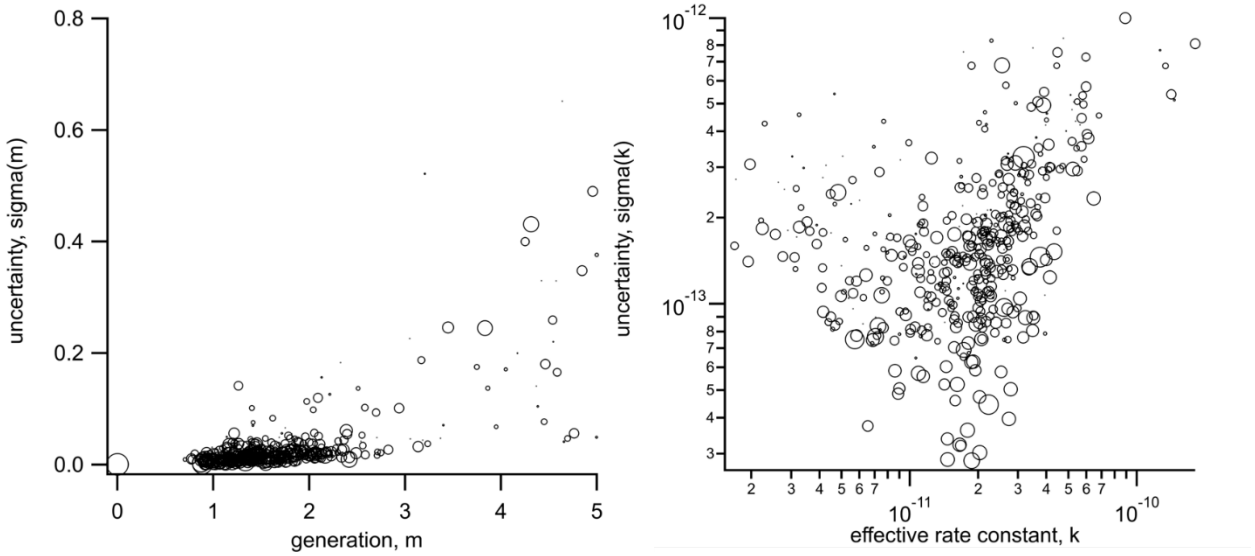
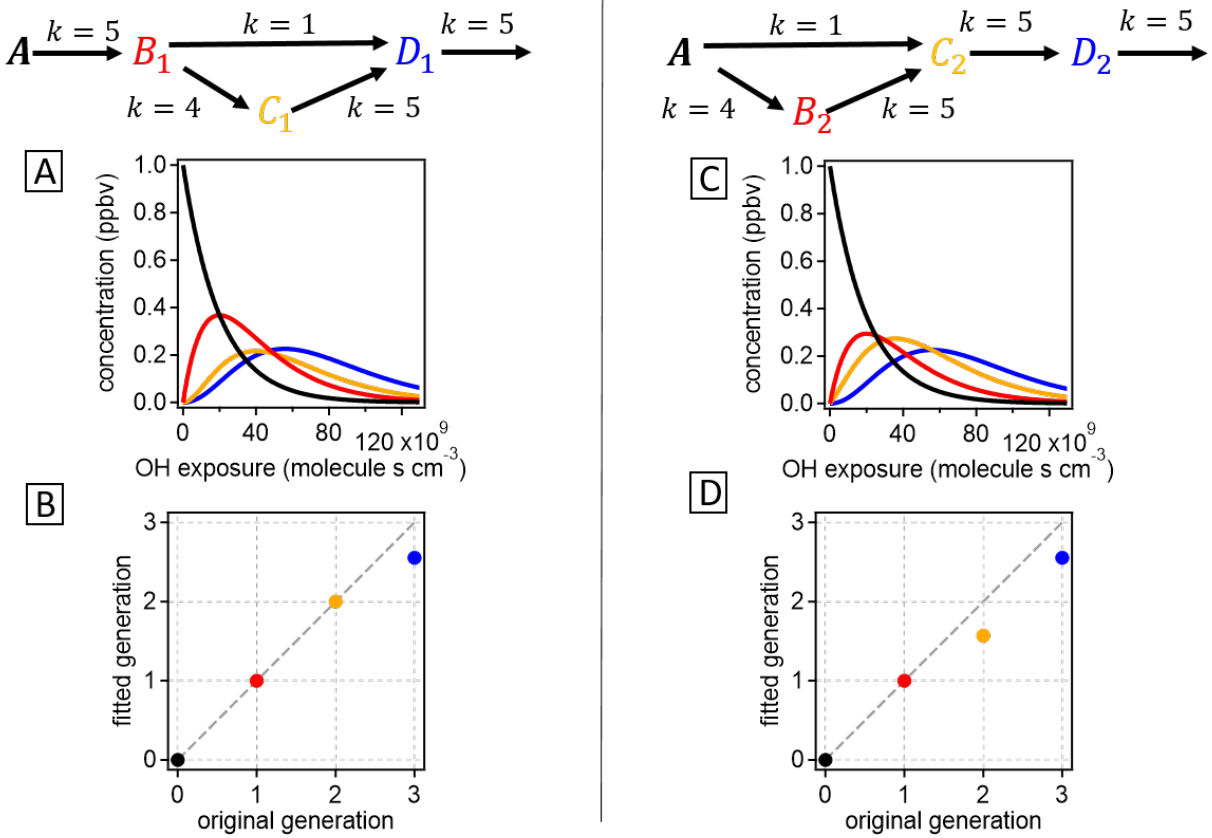


Figure S9

Parameterized generation for non-linear systems.

The reaction pathway for two different synthetic systems is shown at the top. The rate constants are in units of 10^{-11} cm^3 $molecule^{-3}$ s^{-1}.

A. Time series of reactant species in synthetic system 1. B. Parameterized generation numbers for synthetic system 1. C. Time series of reactant species in synthetic system 2. D. Parameterized generation numbers for synthetic system 2.



S1 Best methods for determining generation number

The best fit parameterization of *m* can be improved with two methods: one, by fitting to early data; and two, by reducing noise.

The generation number *m* is determined from the curvature of the initial growth of the product species (Figure 3). Based on the method of fitting, the curve fit algorithm can return an incorrect value of *m*. For example, the following two species were fit using least squares, which is the default method in many software packages.

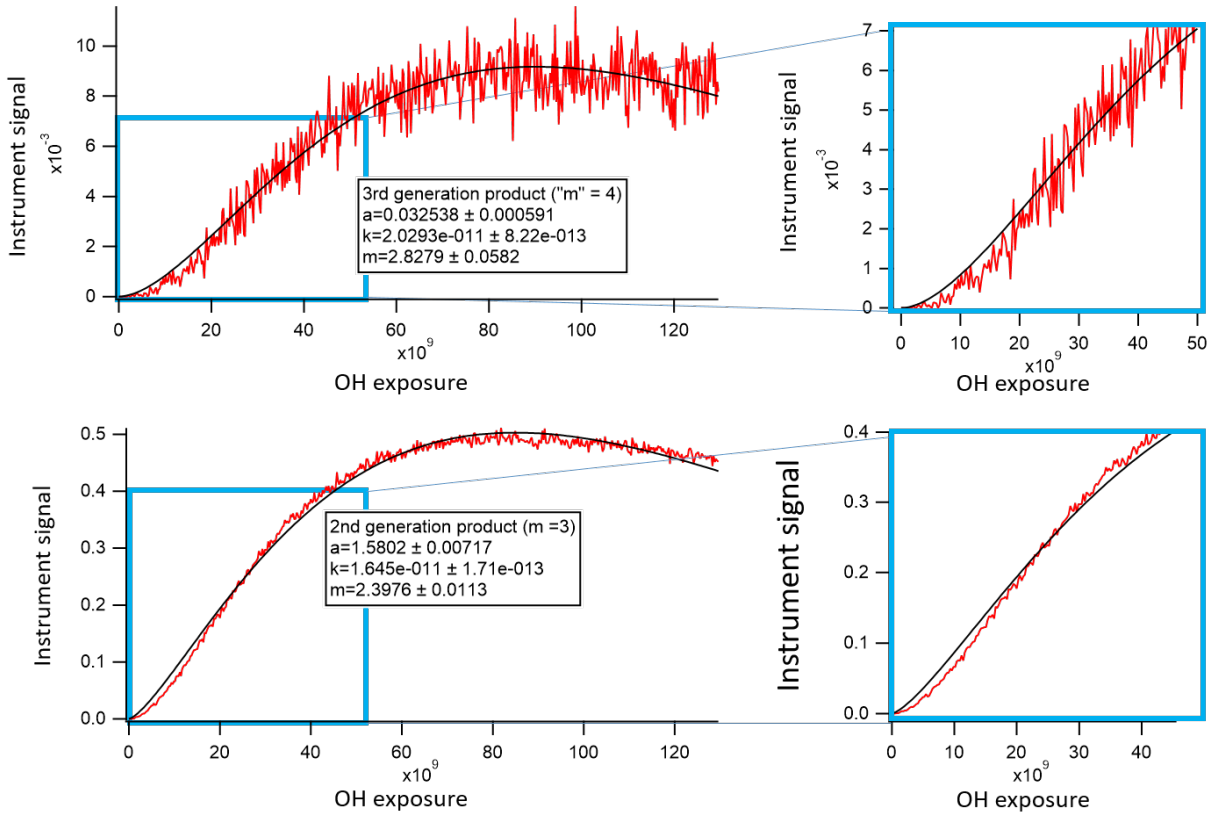


Figure S1.1

The left two panels of Figure S1.1 show the full time series, and the right two panels expand the boxed inset. This figure shows that the fit is poorer for early time points. Later data are fit better, because they have higher values and are therefore weighted more heavily in least-squares fitting. The result is an artificially low returned value of *m*. This issue can be solved by fitting to early data only. The optimal number of points to fit differs based on the *k* and *m* of the species in question. If two few points are fit, then no trend is discernable; if too many points are fit, then *m* is underpredicted.

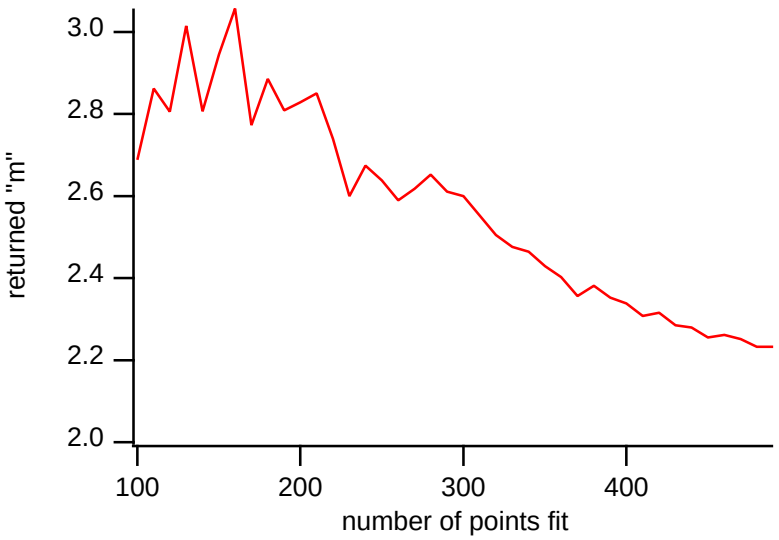


Figure S1.2

Figure S1.2 shows the returned value of *m* as a function of number of data points fit, using species "C3" from the synthetic system as an example. Based on the typical noise level of our data, we chose to exclude fits with fewer than 100 data points. The largest returned value of *m* is the most accurate.

The fit can be further improved by reducing noise. Mass spectrometers typically exhibit a Poisson noise distribution, where values are normally distributed about the actual signal. This noise should cancel out in the integration of a measurement, resulting in a smoother curve. The integral of Eq. 2 is:

$$\int [X]dt = \frac{a}{k}\left(1 - \frac{\Gamma(m,kt)}{\Gamma(m)}\right) \text{ (Eq. S1)}$$

where $\Gamma(m, kt)$ and $\Gamma(m)$ are the partial and complete gamma functions, respectively. Eq. S1 can be fit to integrated data, using for time *t* the OH exposure $OH\Delta t$. The returned values of *m* as a function of points fit, using integrated data, is shown in Figure S1.3. This returns a more accurate value of *m* using fewer data points.

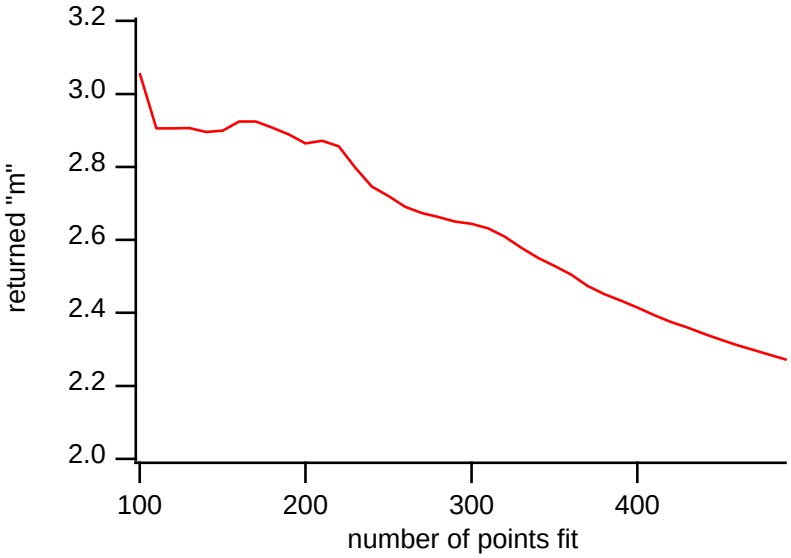


Figure S1.3