Supplementary Information to:

Mathematical Chromatography Deciphers the Molecular Fingerprints of Dissolved Organic Matter

Urban J. Wünsch[1*], Jeffrey A. Hawkes[2]

[1] Chalmers University of Technology, Architecture and Civil Engineering, Water Environment Technology, Sven Hultins Gata 6, 41296 Gothenburg, Sweden
[2] Analytical Chemistry, Department of Chemistry - BMC, Uppsala University, Uppsala, Sweden.

* *Correspondence to*: Urban J. Wünsch (wuensch@chalmers.se). Present address: Sven Hultins Gata 6, 41296 Gothenburg, Sweden.

This Supporting Information contains 9 pages and 14 figures.

Supplementary Information contents:

Supplementary Figures 2

Supplementary Figures

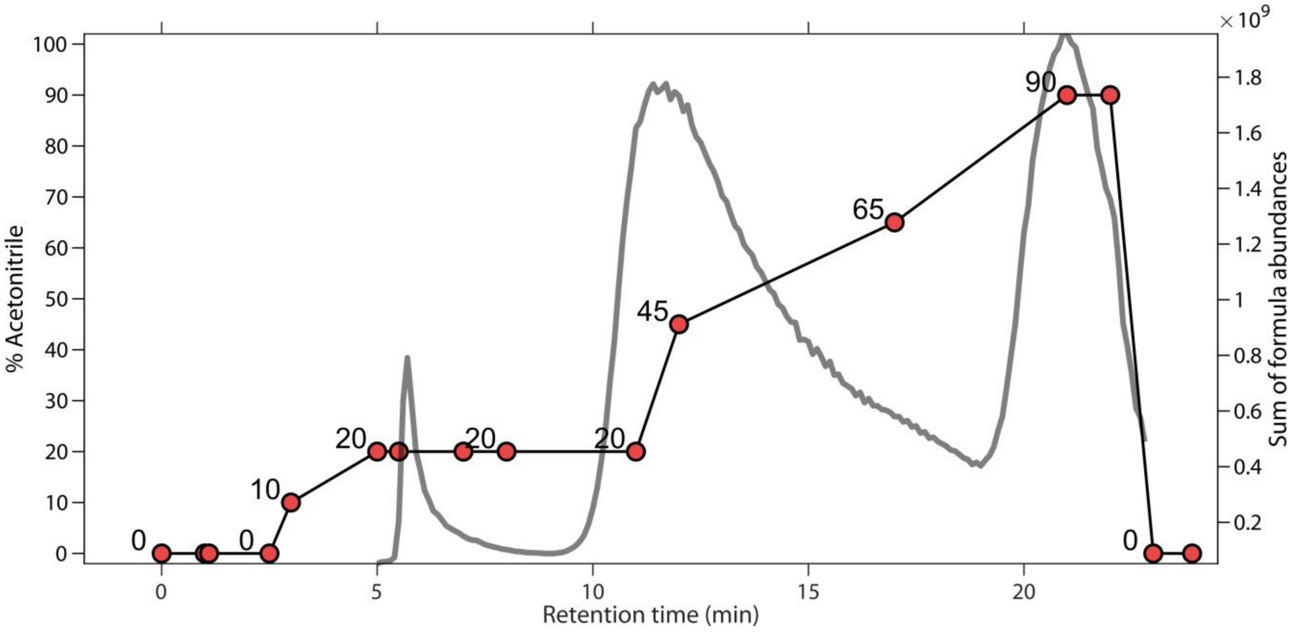

Figure S1: Reverse-phase elution gradient of acetonitrile. The solid line with red markers represents the acetonitrile gradient in each chromatographic separation, while the total sum of assigned molecular formula abundances (in all samples) is given for reference as thick, grey line.

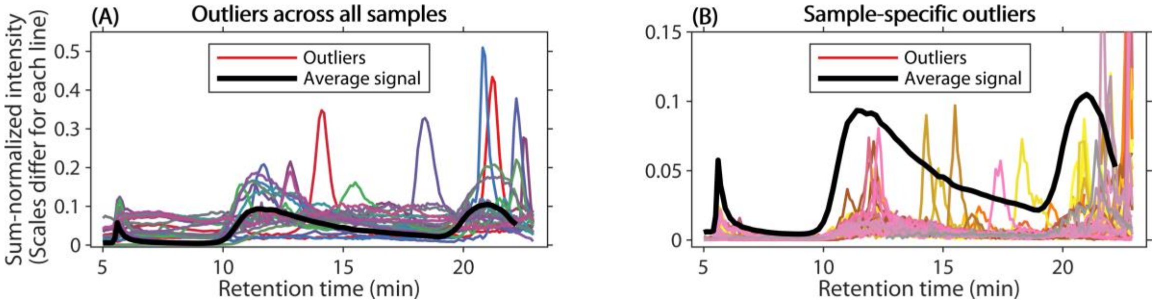

Figure S2: Outlier molecular formulas. (A): Outliers detected across all samples (N = 36). Coloured lines represent outliers, the bold black line is the average formula chromatogram. Outliers were removed from the dataset. **(B)**: Sample-specific outlier-formulas (N = 50). These were set to missing numbers.

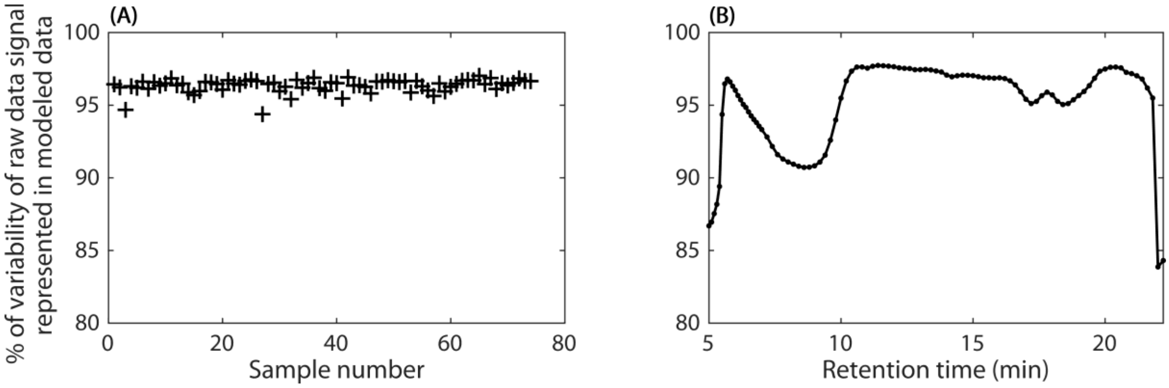

Figure S3: Percentage of the raw mass spectrum represented in the modelled dataset. The shown percentages represent the fraction "Discarded" relative to "Raw data" in Fig. S4. **(A):** Modelled fraction across the 74 samples (while summing the ion intensities across mass spectra and chromatograms). **(B)**: Modelled fraction across the LC chromatogram (while summing the ion intensities across mass spectra and samples).

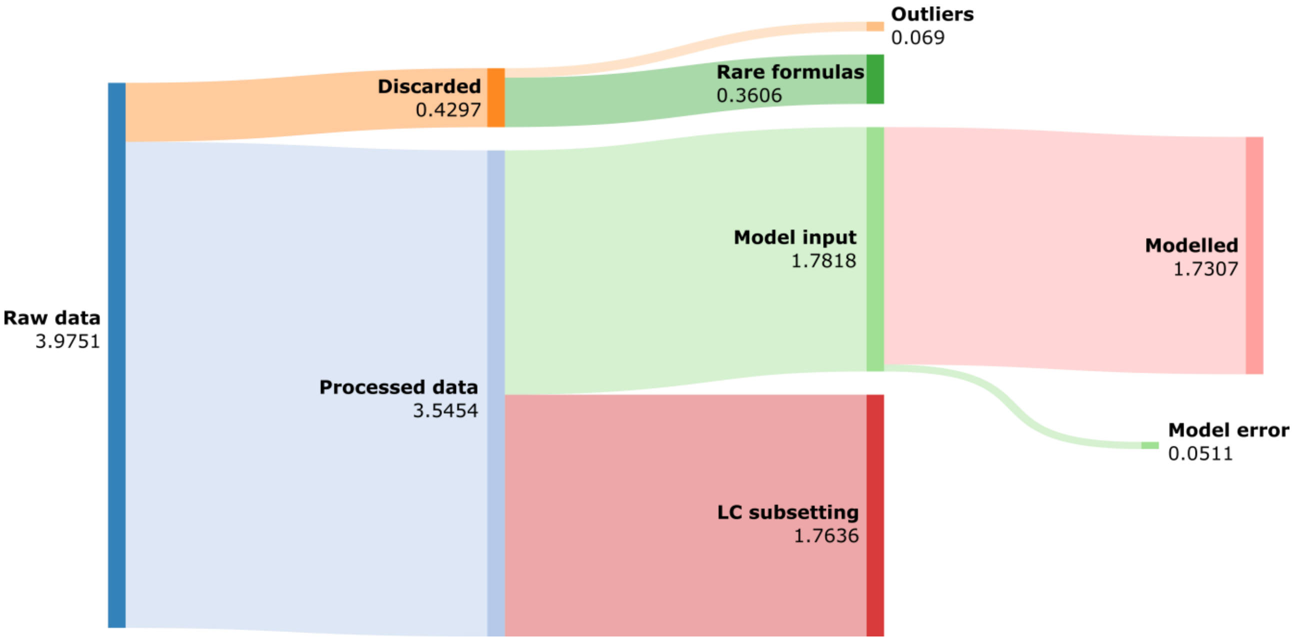

Figure S4: Flow diagram showing quantitative impact of processing steps and modelling error. The numbers shown are the sum of squared data at each processing steps divided by 1×10^{19}. In the step "LC subsetting", approximately every second retention time was omitted, which did not lead to loss of chemical information due to the broad elution patterns in the dataset. The figure was created with the software SankeyMATIC (https://github.com/nowthis/sankeymatic).

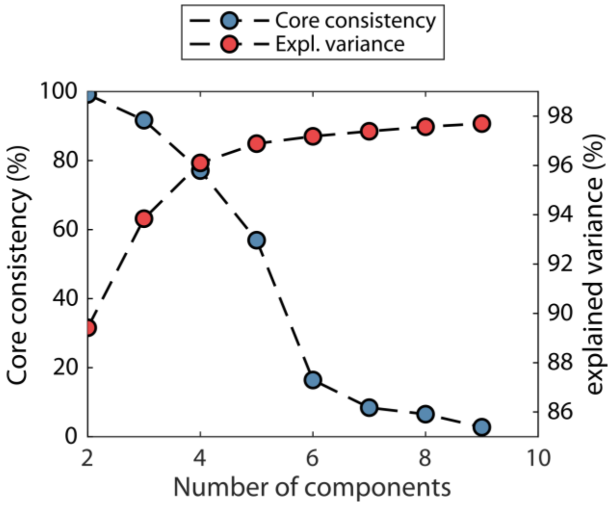

Figure S5: PARAFAC model performance for different number of components. Shown are core consistency (blue) and percentage of explained variance (red) for two to nine components.

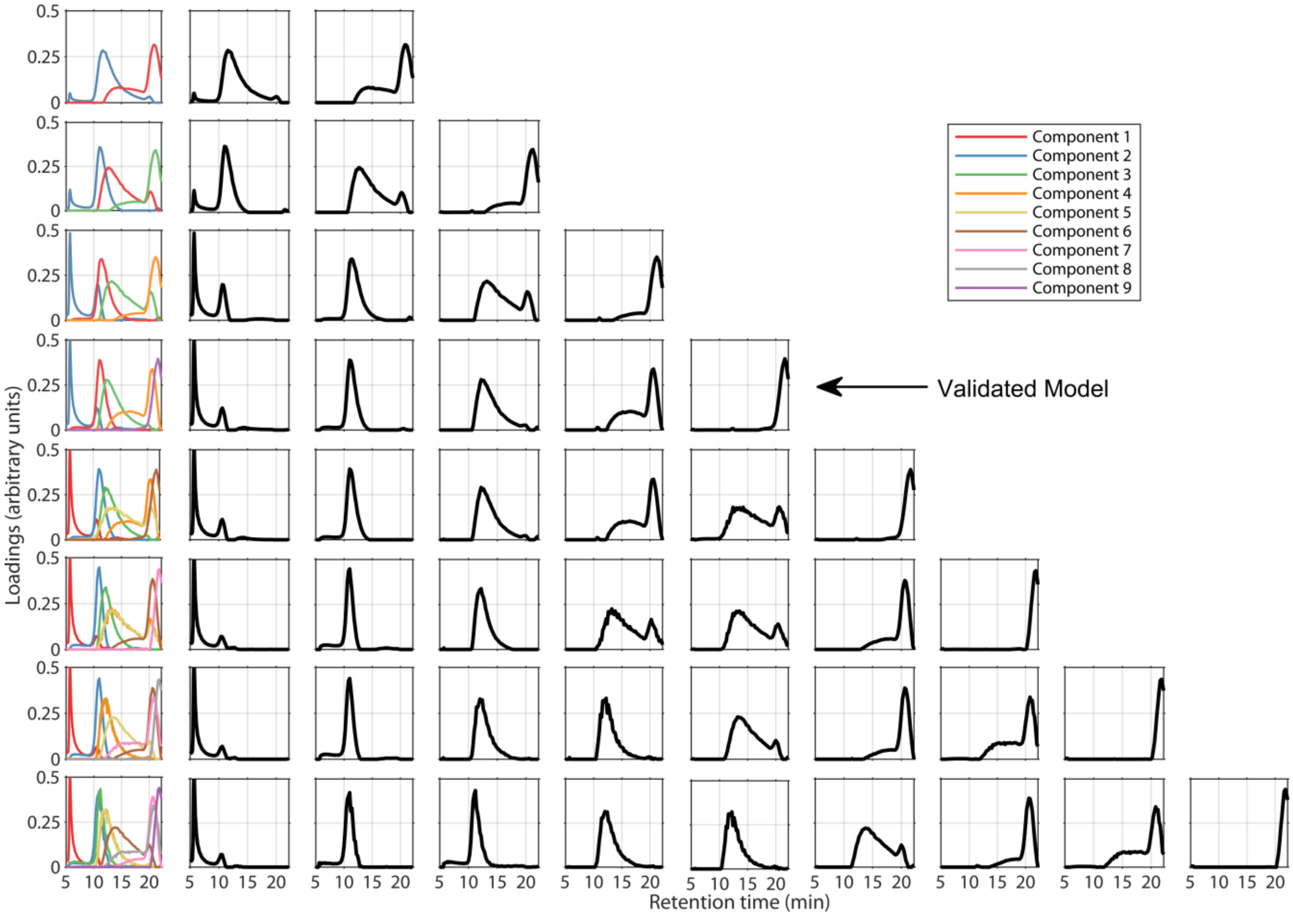

Figure S6: Elution profiles of PARAFAC models with two to nine components. The first plot in each row shows all components superimposed, while the following plots show elution profiles sorted by increasing retention time at which the maximum loading occurred. The validated model is highlighted for reference. The six- to nine-component models were not successfully split-half validated and had a noticeably lower core consistency without explaining much more variance (Fig. S5).

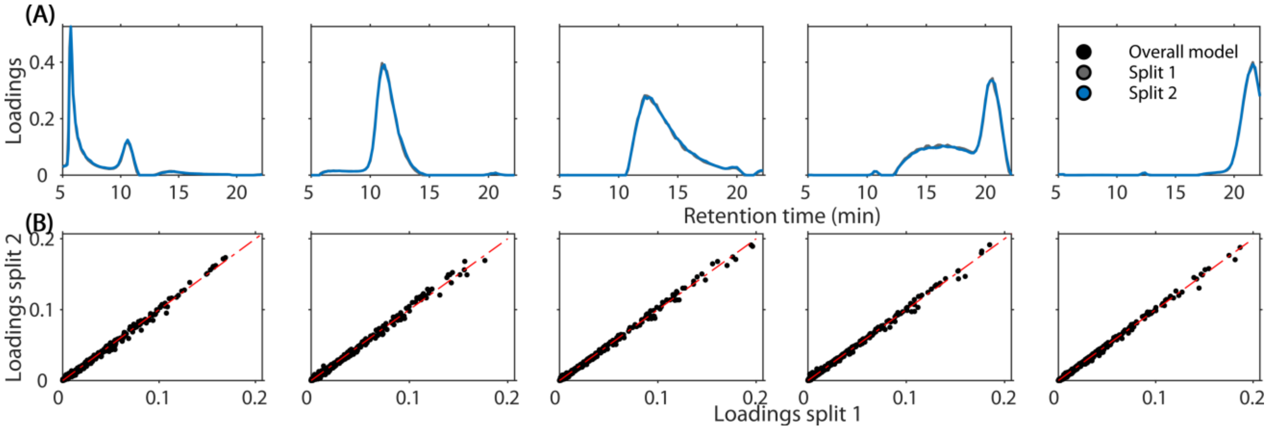

Figure S7: Split-half validation of the five-component PARAFAC model. (A): Superimposed comparison between loadings of the overall model (black bold line) retention time loadings with those of two random dataset halves (thin black and grey lines). **(B)**: Comparison between loadings of six components in two random halves of the overall dataset. Due to the large number of signals, a superimposed loadings plot was substituted by this direct comparison. The red line in each plot represents the 1:1 line that symbolizes a perfect match between components. Tucker congruence coefficients (TCC) for all comparisons was >0.99.

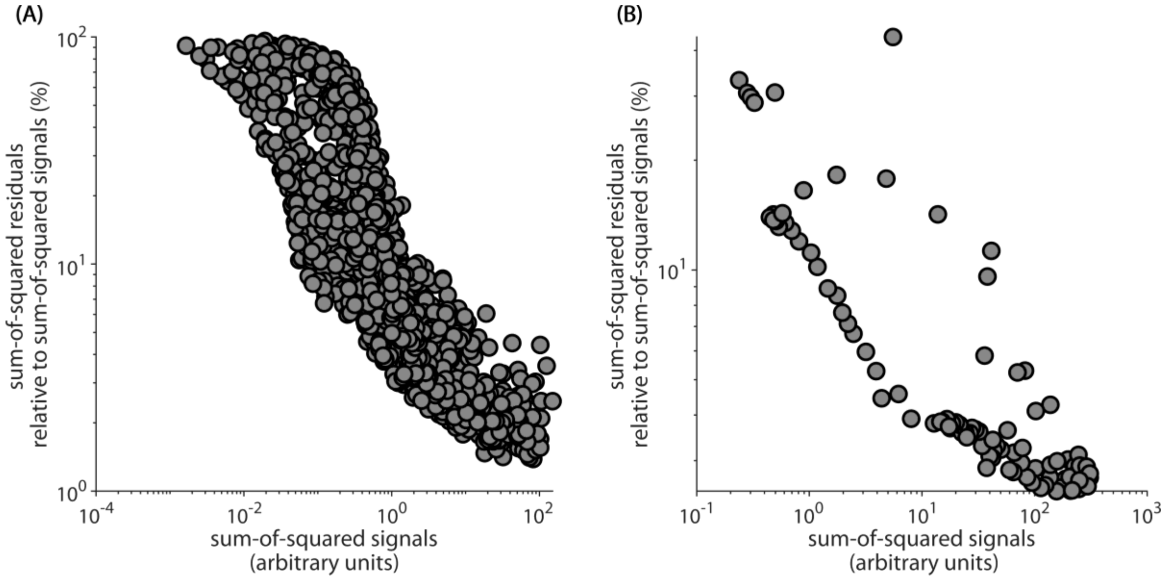

Figure S8: Comparison of sum-of-squared signal against sum-of-squared residuals. (A): Molecular formulas. (B): Retention times. Both x- and y-axes are log-scaled.

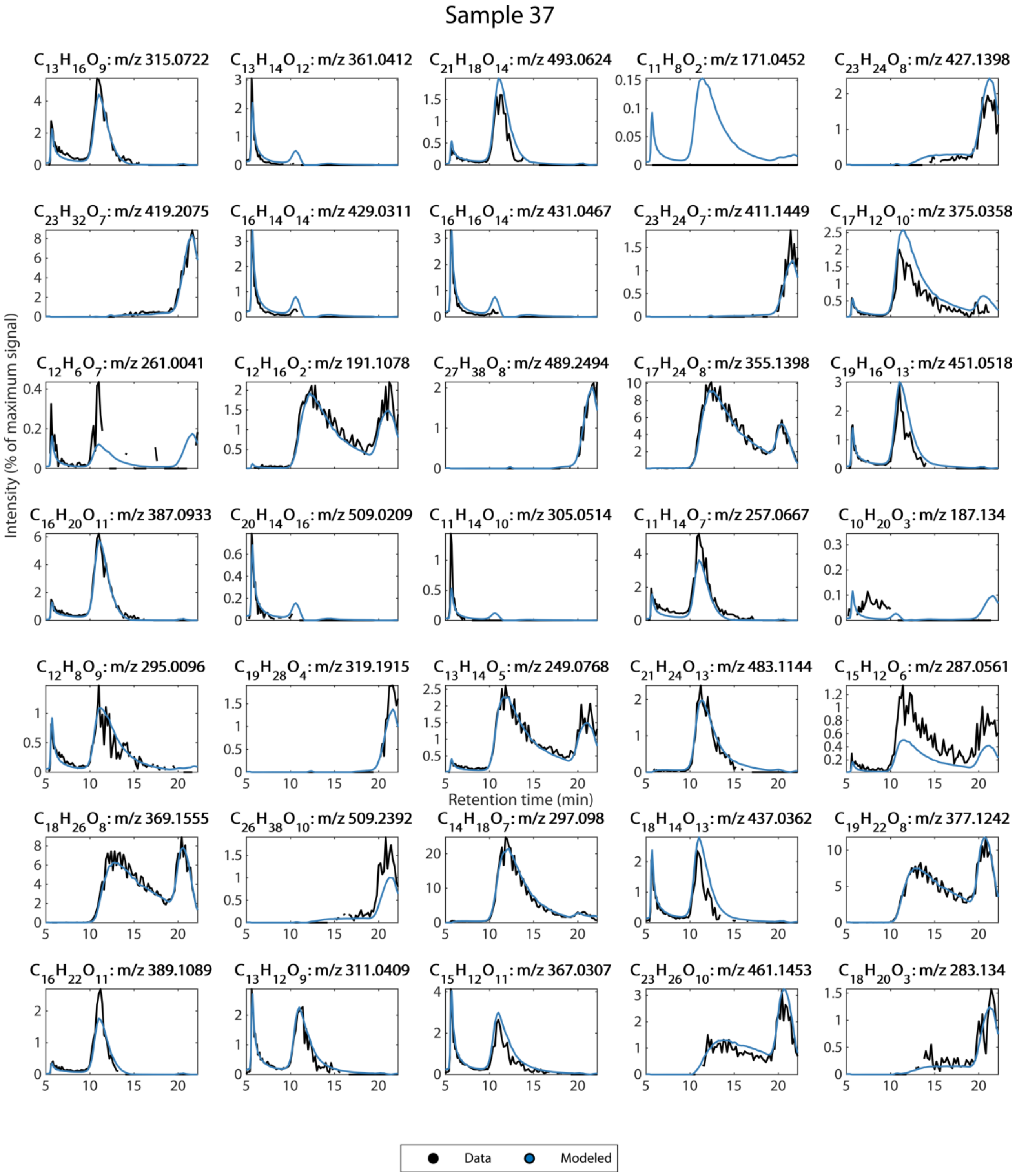

Figure S9: Data vs. modelled data for a randomly selected sample and randomly selected molecular formulas. Black: Chromatographic raw data for different molecular formulas (formula and m/z given as plot title). Blue: Corresponding modelled data. The intensities are expressed in percent, to indicate the distance to the highest observed signal in the dataset.

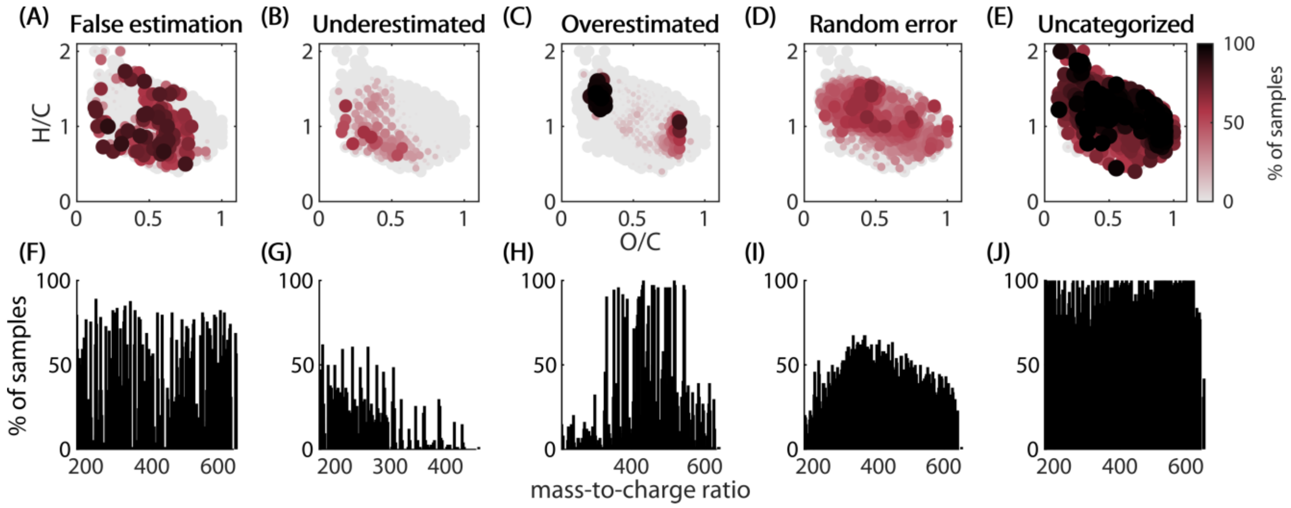

Figure S10: Frequencies of types of modelling error as a function of DOM composition. For 100 270 formula chromatograms (74 samples x 1355 formulas), residuals were analysed and categorized into false estimations of chromatograms when raw data only contained zeros **(A, F)**, systematic underestimations **(B, G)**, overestimations **(C, H)**, random modelling error **(D, I)**, and lastly uncategorizable modelling errors **(E, J)**. For reference, all detected formulas are given as grey background. Frequencies were counted for each formula and normalized by the number of samples (expressed in % of samples). The top row shows the composition in the van Krevelen space, the bottom row shows mass spectra.

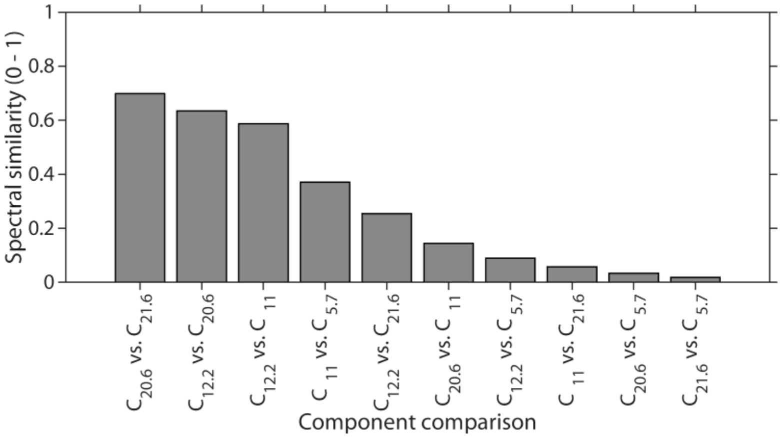

Figure S11: Similarity between the mass spectra of PARAFAC components. The similarity is measured on a scale from zero (no similarity) to one (identity), whereby 0.95 is commonly defined as significant similarity. Components were compared in all possible combinations (e.g. $C_{20.6}$ vs $C_{21.6}$: spectral similarity between components showing their elution maxima at 20.6 and 21.6 min). Comparisons are sorted from most to least similar.

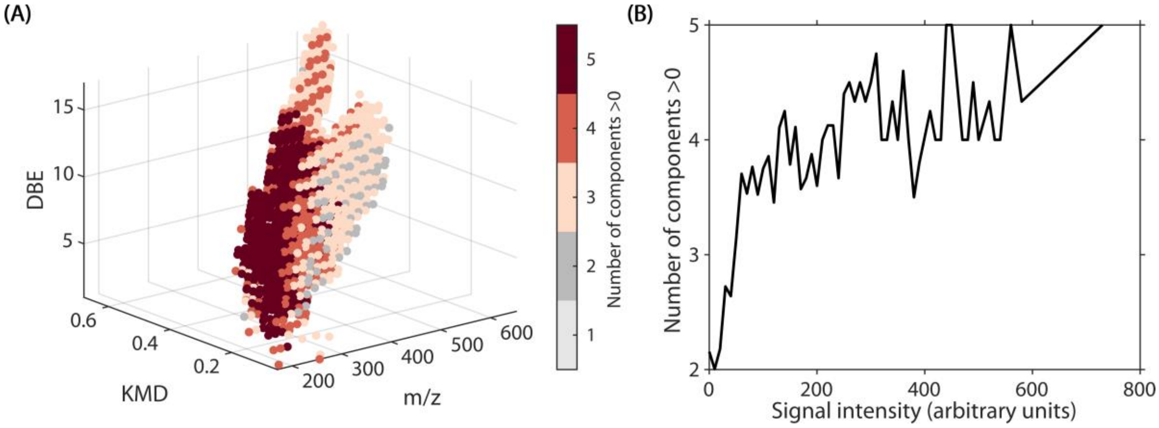

Figure S12: Statistical complexity of molecular formulas. (A): Number of components with loadings > 0 vs. mass-to-charge (m/z), Kendrick mass defect (KMD), and double bond equivalent (DBE). **(B):** Number of components with loadings > 0 compared to the signal intensity. For smooth visualization of broad trends, the data in (B) was rounded to the next integer divisible by ten.

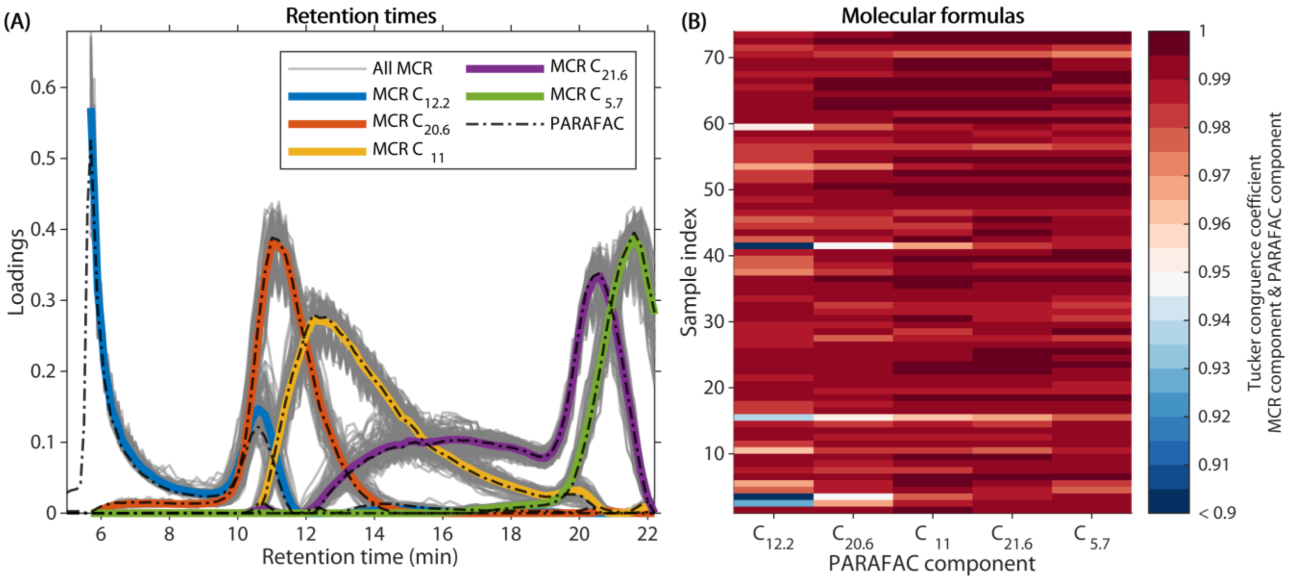

Figure S13: Comparison between nonnegative matrix factorization and parallel factor analysis. Five-component multivariate curve resolution (MCR) models were fit to each sample separately, while PARAFAC models were fit to all samples simultaneously. **(A):** Retention time loadings comparison between five-component MCR and PARAFAC models. Grey lines depict the MCR loadings for each sample, the coloured lines depict median loadings for each of the five components. The black dashed line depicts the validated PARAFAC model presented in the main text. **(B):** Similarity between the molecular formula loadings of MCR and PARAFAC models. Due to the difficulty of visualizing many mass spectra, a congruence analysis was carried out between the loadings of the global PARAFAC model and the sample-specific MCR model. A Tucker congruence coefficient exceeding 0.95 is commonly indicating spectrally indistinguishable components.

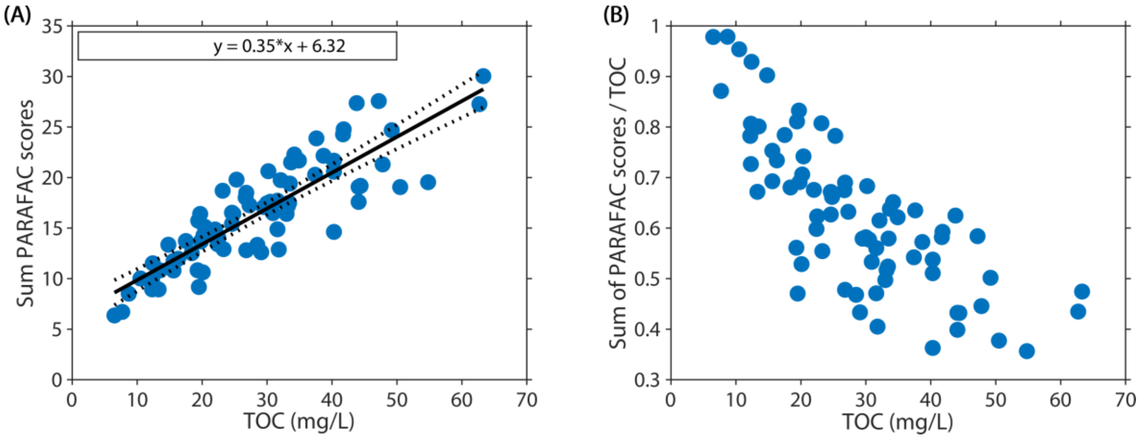

Figure S14: Relationship between the sum of PARAFAC scores and total organic carbon (TOC) (A): Sum of scores vs. TOC. $R^2 = 0.74$, $p < 0.001$. **(B):** Decrease of scores:TOC with increasing TOC.