

The molecular fingerprint of fluorescent natural organic matter offers insight into biogeochemical sources and diagenetic state

Wünsch, Urban; Acar, Evrim; Koch, Boris Peter; Murphy, Kathleen; Schmitt-Kopplin, Philippe; Stedmon, Colin Andrew

Published in: Analytical Chemistry

Link to article, DOI: 10.1021/acs.analchem.8b02863

Publication date: 2018

Document Version Version created as part of publication process; publisher's layout; not normally made publicly available

Link back to DTU Orbit

Citation (APA):

Wünsch, U., Ácar, E., Koch, B. P., Murphy, K., Schmitt-Kopplin, P., & Stedmon, C. A. (2018). The molecular fingerprint of fluorescent natural organic matter offers insight into biogeochemical sources and diagenetic state. *Analytical Chemistry*, *90*(24), 14188–14197. https://doi.org/10.1021/acs.analchem.8b02863

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.

- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

The molecular fingerprint of fluorescent natural organic matter offers insight into biogeochemical sources and diagenetic state

3 Urban J. Wünsch^{1,2*}, Evrim Acar³, Boris P. Koch^{4,5}, Kathleen R. Murphy¹, Philippe Schmitt-Kopplin⁶, Colin
 4 A. Stedmon²

- 6 Sven Hultins Gata 6, 41296 Gothenburg, Sweden
- ² National Institute of Aquatic Resources, Technical University of Denmark, Kemitorvet, Kgs. Lyngby 2800,
 Denmark
- 9³ Simula Metropolitan Center for Digital Engineering, Pilestredet 52, 0167 Oslo, Norway
- ⁴ Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Am Handelshafen 12, 27570
- 11 Bremerhaven, Germany
- 12 ⁵ University of Applied Sciences, An der Karlstadt 8, 27568 Bremerhaven, Germany
- 13 ⁶ Research Unit Analytical Biogeochemistry (BGC), Helmholtz Zentrum München, German Research Center
- 14 for Environmental Health, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany
- 15 * Correspondence to: Urban J. Wünsch (wuensch@chalmers.se). Present address: Sven Hultins Gata 6, 41296
- 16 Gothenburg, Sweden.
- 17

^{5&}lt;sup>1</sup> Chalmers University of Technology, Architecture and Civil Engineering, Water Environment Technology,

18 Abstract

19 Investigating the biogeochemistry of dissolved 20 organic matter (DOM) requires the synthesis of data 21 from several complementary analytical techniques. In 22 contrast to subjective post-hoc correlation analysis, a 23 robust integration requires data fusion, capable of 24 simultaneously decomposing data from multiple 25 instruments while identifying linked and unrelated 26 signals. Here, Advanced Coupled Matrix and Tensor 27 Factorization (ACMTF) was used to identify the 28 molecular fingerprint of DOM fluorescence fractions 29 in Arctic fjords. ACMTF explained 99.84 % of the



30 variability with six fully shared components. Individual molecular formulas were linked to multiple 31 fluorescence components and vice versa. Molecular fingerprints differed in diversity and oceanographic

32 patterns, suggesting a link to the biogeochemical sources and diagenetic state of DOM. The fingerprints

33 obtained through ACMTF were more specific compared to traditional correlation analysis and yielded greater

34 compositional insight. Multivariate data fusion aligns extremely complex, heterogeneous DOM datasets, and

35 thus facilitates a more holistic understanding of DOM biogeochemistry.

36 Introduction

The complex multifaceted interactions between dissolved organic matter (DOM) and biological¹ and physical 37 processes² cements its central role in aquatic ecosystems and is intrinsically linked to its chemical 38 composition.³ The wide variety of environmental processes involved results in an extremely complex pool of 39 40 organic compounds that spans nearly all possibilities defined by the laws of chemical bonding.⁴⁻⁶ In order to understand the biogeochemical role of DOM in natural waters, it is essential to reduce the complexity of 41 analytical data and to trace and characterize various underlying fractions. Owing to its molecular complexity, 42 43 the simultaneous quantification and characterization of DOM presents a formidable analytical challenge. Users often have to choose between techniques with different strengths and limitations.^{7,8} The insight gained from 44 independent analytical techniques strongly depends on the experimental design and data analysis approach, 45 and is ultimately limited by the intrinsic constraints of the individual approach. 46

47 The determination of ultraviolet-visible (UV-Vis) spectroscopic properties (targeting chromophoric and 48 fluorescent DOM, CDOM and FDOM, respectively) represents a rapid method to follow DOM dynamics.^{9,10} An ever-increasing number of studies focus on FDOM, since its measurement is cost-efficient, highly sensitive 49 50 and suitable for field deployment.¹¹ Fluorescence excitation emission matrices (EEMs) are frequently decomposed into the underlying independently fluorescing components using multiway techniques such as 51 Parallel Factor Analysis (PARAFAC).¹² However, since fluorescence and absorbance require optically-active 52 compounds, they can only target a fraction of the DOM pool. The chemical structures responsible for the UV-53 Vis spectroscopic properties of DOM have yet to be uncovered.¹⁰ Moreover, the chemical interpretation of 54 PARAFAC spectra is inherently difficult and often results in the ambiguous labeling of components that 55 suggests representation of molecular species (e.g. proteins or humic substances).¹³ 56

Another analytical approach to DOM characterization is ultrahigh-resolution mass spectrometry,¹⁴ which 57 determines the exact masses, and thus molecular formulas, of organic substances present.⁵ Importantly, while 58 59 DOM can be mass-resolved, ultrahigh-resolution mass spectrometry cannot routinely distinguish structural 60 isomers for a given mass peak and thus still produces convoluted analytical signals that can be challenging to interpret in complex mixtures. To date, this has mainly been addressed by multivariate analysis with Principal 61 Component Analysis (PCA) or Hierarchical Cluster Analysis.¹⁵ PCA in particular is a powerful approach to 62 63 reduce complexity and isolate factor loadings that correspond to the chemical imprint of environmental processes. However, the properties of the decomposition might hinder the discovery of true chemical signals. 64 65 For example, in PCA components are orthogonal, i.e. all factors have a loading similarity of zero. However, 66 many properties of DOM can be expected to be correlated (non-orthogonal). A more flexible multivariate approach to distinguish underlying factors of molecular formula matrices is therefore needed. 67

The comparison of fluorescence spectroscopy and ultrahigh-resolution mass spectrometry makes apparent that 68 the strengths and limitations of both approaches are diametrically opposite: The former approach, when 69 70 coupled with PARAFAC allows meaningful statistical description, but offers only limited insight into DOM 71 chemical composition, while the latter offers a wealth of qualitative chemical information with limited means 72 to systematically elucidate the primary factors responsible for observed dynamics. Experience in other 73 disciplines, such as metabolomics, shows that considerable analytical advances are achieved when two or more complementary datasets are jointly analyzed.¹⁶ For DOM, recent studies have employed post-hoc rank-74 correlation analysis to establish links between optical and chemical properties of DOM.^{17,18} However, 75 76 considering that thousands of signals are compared, the risk of false positive correlations is significant. 77 Moreover, correlations may sometimes be hard to interpret, e.g. when negative correlations are reported while 78 both signals in principle correspond to analyte concentrations. A promising approach is to jointly decompose 79 these datasets into multiple underlying factors using advanced data fusion that can account for their convoluted 80 character. However, no such models have been tested for DOM.

81 The heterogeneous nature of DOM datasets requires an approach that can handle the different types of datasets

82 while accounting for the partial overlap of detector signals. A data fusion model based on simultaneous

83 factorization of multiple datasets, called Advanced Coupled Matrix and Tensor Factorization (ACMTF) model

- 84 has been developed specifically for such scenarios.¹⁹ In recent years, ACMTF has been used in metabolomics²⁰
- 85 and medical applications.²¹ Here, we applied ACMTF to simultaneously analyze and decompose data from
- 86 two popular DOM characterization techniques, fluorescence EEMs and Fourier transform ultrahigh-resolution
- 87 mass spectrometry (FT-ICR-MS) molecular formulas (N = 527) for samples (N = 174) originating from three
- 88 Arctic Fjords. The associations reported by the statistical components identified by ACMTF are subsequently
- 89 compared to those suggested by post-hoc correlation analysis. We propose that data fusion represents a vital
- 90 step towards a more holistic data analysis that can help to better elucidate the complex dynamics of DOM.

91 Materials and Methods

92 Sample Collection

93 In July 2016, 174 water samples were collected onboard R/V Maria S Merian (cruise MSM56, Ecological 94 Chemistry in Arctic Fjords) over a three-week period (see Fig. S1 and Table S1 for an overview of sampling regions, salinity-, temperature-, DOC-, and depth ranges). The transect encompassed three fjords 95 (Kongsfjorden [Longyearbyen], Scoresby Sound [East Greenland], and Arnarfjörður [West Iceland]) spanning 96 from 79 °N to 65 °N and 28 °W to 12 °E. Kongsfjorden and Scoresby Sound both have marine terminating 97 98 glaciers, while Arnarfjörður further south in Iceland does not. All fjords receive limited DOM input from rivers 99 in their catchments and the flux of marine DOM from the shelf and production of DOM associated with plankton productivity dominate. Scoresby Sound, similar to many East Greenland fjords, receives terrestrial 100 organic matter from the Arctic transported in the East Greenland Current (EGC) and has low productivity. 101 102 Kongsfjorden and Arnarfjörður are not influenced by the EGC and are primarily supplied with Atlantic water 103 with little or no terrestrial DOM. Water samples were collected at depths ranging from 1.7 to 1397 m with a 24-bottle CTD Rosette, equipped with Niskin bottles and immediately filtered using pre-combusted GF/F 104 105 filters (0.7 µm, Whatman) by applying a vacuum of < 200 mbar. After filtration, DOM was immediately solidphase extracted using 200 mg PPL-resins as described previously.²² Cartridges were desalted and dried 106 onboard and stored dark and frozen at -20°C. In the home laboratory, DOM was eluted with 1000 µL methanol; 107 the final extract volume was determined by weight and samples were stored at -18° C until analysis. 108

109 Spectroscopic measurements

Fifty µL DOM extract (in methanol) were dried using a gentle stream of N2 at room temperature, reconstituted 110 in 4 mL of 150 mM ammonium acetate (pH 7) in pre-combusted, amber glass vials, and equilibrated at room 111 112 temperature for 30 minutes. Fluorescence and absorbance measurements were obtained using a HORIBA AquaLog fluorometer using a 10 mm quartz cuvette (Helma Analytics). Fluorescence emission was detected 113 114 in the range of 240 - 600 nm (increment ~3.3 nm) at excitation wavelengths between 240 nm and 450 nm (increment 3 nm between 240 and 360 nm, 9 nm between 360 and 450 nm). A separate absorbance 115 measurement was carried out to measure absorbance between 240 and 600 nm at increments matching the 116 fluorescence excitation. The fluorescence data was processed using the dr EEM toolbox.¹² Data were corrected 117 for inner filter effects using the absorbance-based method (absorbance between 0.014 and 0.07 cm⁻¹ at 118 260 nm).²³ First and 2nd order physical scatter was removed and not interpolated. A fluorescence EEM of 119 120 150 mM ammonium acetate was subtracted as the spectroscopic blank. Average EEMs for the three fjords are

121 depicted in Fig. S2.

122 Spectrometric measurements

Fourier transform ion cyclotron resonance mass spectra were collected with a 12 T Bruker Solarix mass 123 124 spectrometer (Bruker Daltonics, Bremen, Germany) using an Apollo II electrospray ionization (ESI) source in negative ionization mode. Samples were diluted in 50/50 (v/v) methanol / water to a concentration of 1.5 125 nmol DOC mL⁻¹, and injected into electrospray source at a flow rate of 0.12 mL h⁻¹ with a nebulizer gas 126 127 pressure of 2.2 bar and a drying gas flow rate of 4 L min⁻¹. Spectra were externally calibrated using arginine clusters, and then internally calibrated using marine DOM molecular formulas.²⁴ The spectra were acquired at 128 129 4 Mega words over a mass range of m/z 100 - 1000, and 300 scans were accumulated for each spectrum. The 130 average mass resolution of all signals at 400 m/z was 375,000. The formula assignment was carried out as 131 detailed in the SI. Briefly, molecular formulas were calculated from m/z values allowing for elemental combinations ${}^{12}C_{0-\infty}{}^{13}C_{0-1}{}^{1}H_{0-\infty}{}^{14}N_{0-2}{}^{16}O_{0-\infty}{}^{32}S_{0-1}$, a mass accuracy threshold of $|\Delta m| \le 0.2$ ppm, and a relative 132 133 abundance of > 1%. Formulas which were either detected in process blanks (PPL extraction of ultrapure water) or contained in the list of potential surfactants were removed from the entire data set.²⁵ It should be considered 134 135 that every assigned molecular formula most likely implies an immense structural diversity of isomers.⁶ Since FT-ICR-MS alone is unable to distinguish between such structural isomers, we reserve the term component 136

137 for statistical components, and use the term *molecular formula* in the context of spectrometric signals.

138 Data processing

139 In order to describe optical and chemical properties of DOM using statistical models, data were pre-processed as follows: (1) EEM data were normalized to reduce concentration effects and match the character of relative 140 formula abundances (using "normeem" function in drEEM); (2) signals in EEMs and the molecular formula 141 142 matrix were scaled with their Euclidean norm to equalize their numerical leverage; (3) FT-ICR-MS formula 143 abundances were scaled by division with the square root of the standard deviation to reduce modeling leverage 144 of highly abundant formulas; (4) Undetected FT-ICR-MS formulas in a given sample were assigned as 145 "missing" and formulas with more than 3.6 % missing detections (i.e. more than 10 samples without formula observation) were excluded from further analysis (81 % of the 2776 formulas in the original data set). While 146 147 this represents a significant reduction in molecular formula data, a linkage between fluorescence signals present in all samples with molecular formulas only present in a small fraction of samples is unlikely. 148 149 Preliminary ACMTF models indicated high residuals below m/z 300 and above m/z 500, therefore these 150 formulas were removed to mitigate disturbances. The resulting datasets had the following structure and 151 dimensions: sample x emission x excitation (174 x 91 x 44), and sample x formula abundance (174 x 527) for fluorescence EEMs and FT-ICR-MS formula abundances, respectively. Based on the original relative 152 153 molecular formula abundances, the subset of 527 modeled molecular formulas represented 38 ± 5 % of the 154 ESI-MS molecular formula abundance (Fig. S3).

155 Advanced Coupled Matrix and Tensor Factorization

A detailed description of the fundamental principle of the ACMTF model is presented in the SI (section S1). 156 157 Briefly, ACMTF jointly decomposes fluorescence EEMs and molecular formula matrices into a set of trilinear fluorescence components and bilinear molecular formula components by fitting a PARAFAC model to the 158 159 fluorescence EEMs and factorizing the molecular formula matrix in a way that the component scores are identical. Component weights (λ for EEMs and σ for formula matrices) are used to evaluate whether a 160 particular component is shared between both analytical datasets. ACMTF modeling was carried out using the 161 Matlab CMTF toolbox^{19,26} in conjunction with the Tensor toolbox²⁷ and the SNOPT toolbox.²⁸ ACMTF 162 factorization is computationally intense and calculations were therefore carried out using a set of IBM 163 164 NeXtScale nx360 M4 nodes, with 100 models being fit simultaneously (reducing the analysis time for 100 165 models from 6 days on a single-core computer to 3 h with parallel computing).

Nonnegativity constraints in all modes of both datasets were applied during the modeling. Furthermore, 166 angular constraints in the excitation mode were applied to prevent the algorithm from converging on solutions 167 with highly similar factors (violating model assumptions): Model components were constrained to have Tucker 168 169 Congruence Coefficient (TCC) values between all excitation spectra of less than 0.93 (limit set by maximum similarity between PARAFAC excitation spectra). ACMTF models were evaluated by (1) an assessment of 170 the fluorescence spectra (chemical coherence); (2) a variability-assessment of the component weights λ and σ 171 (model uniqueness); (3) split-half validation. ACMTF component scores were converted to F_{max} -values by 172 multiplying component scores with the spectral maximum of fluorescence excitation and emission, which 173 returns scores in the unit of the modeled data. Here, F_{max}-values represent unitless, relative values, since both 174 175 datasets were scaled and normalized prior to analysis. Contrary to similar procedures during PARAFAC 176 analysis, these pre-processing steps are currently irreversible.

177 The chemodiversity of ACMTF components was estimated as the richness estimator Chao 1 using the R 178 software package vegan (R v3.5.1) with molecular formulas as species and component loadings as species 179 counts.^{29,30} To mimic species counts, the loadings of all components were normalized by the maximum loading 180 across all components, multiplied by 100 and rounded in order to represent integer species counts.

181 Parallel Factor Analysis, Principal Component Analysis, and Pearson Rank Correlation

182 The underlying components of fluorescent DOM in 191 samples (data set contained a small number of samples 183 for which no mass spectra were collected) were isolated using Parallel Factor Analysis using the *drEEM* toolbox.¹² To do so, models with different numbers of components with nonnegative loadings and scores were explored (four to seven components). Ultimately, a six-component PARAFAC model with a core consistency of 1.5% and an explained variance of 99.9% was found to best represent the dataset.³¹ This model was validated using a split-half validation, for which the whole dataset (N = 191) was split into six separate randomly split halves (94 > N < 97).

189 PCA was performed on the molecular formula matrix exclusively to determine the explanatory power of this

190 technique in comparison to ACMTF and to compare the extent of autocorrelation between component loadings.

191 A six component PCA model was calculated for auto-scaled and mean-centered molecular formula abundances

192 (in addition to the preprocessing detailed above). PCA models were calculated using PLS_toolbox in Matlab

193 (Eigenvector Research Inc. v.8.52). Factor similarities between ACMTF and PCA components was quantified

194 using Tucker congruence coefficients for all unique combinations of components (N = 15).³²

195 Pearson rank correlation was performed on the processed data set described above. Molecular formula relative

abundances were correlated to the F_{max} -values of a split-half validated six component PARAFAC model using only pair-wise complete comparisons. A Holm-Bonferroni correction for multiple comparisons was applied to

address the possibility of type I errors,³³ eliminating 19.5 % of correlations that would have otherwise been

reported as significant. The matrix of correlation coefficients (r) was subsequently restricted to comparisons

satisfying the significance threshold of $\alpha = 0.01$; correlation coefficients with $p > \alpha$ (based on Holm-

201 Bonferroni corrected p-values) were ignored in subsequent analyses.

202 Results & Discussion

203 Model validation

204 Similar to other multivariate models such as PARAFAC, the validity of ACMTF models primarily depends 205 upon the applicability of the underlying model to explain the data set variability, as well as choosing the right 206 number of components. The application of data fusion furthermore depends on a stable, reproducible 207 relationship between signals obtained on different instruments. ACMTF was applied under the assumption that 208 the statistically identifiable signals in fluorescence EEMs and molecular formulas respond linearly to the 209 presence of the corresponding (unknown) analytes. The adherence to this assumption was investigated by 210 judging the robustness and representativeness of the model.. This validation of the selected ACMTF model 211 was carried out by analyzing the overall degree of explained variance, the randomness of residuals, the chemical coherence of component loadings, and the ability of reproducing the overall model from fully 212 213 independent subsets of the overall data set.

After the initial data exploration, a six-component ACMTF model was found to best explain the variability in 214 fluorescence EEMs and the molecular formula matrix. With six components, ACMTF explained 99.84 % of 215 variance in both datasets and featured mostly random, low model residuals (example shown in Fig. 1, Fig. S4). 216 217 Fluorescence spectra (Fig. 2, top row) were generally consistent with those expected from pure fluorophores 218 (single emission peak, Stokes' shift between 0.55 and 1.13 eV, Table 1). Further investigations, described in the supplementary information (SI, Figs. S4-8, Table S2), indicated the suitability of ACMTF for the 219 220 simultaneous decomposition of fluorescence EEMs and molecular formulas. The split-half validation indicated 221 that a relatively low number of the modeled formulas did not produce the same component loadings in both 222 independent data set halves (Fig. S6). This indicates that the dynamics of this small subset of modeled formulas 223 deviated from the ideal, linear behavior to some degree. It is possible that these molecular formulas were either not detected reproducibly by ESI-MS, or that they represented independent molecular structures that were not 224 225 represented by a statistical component (molecular fingerprint) present across a range of samples. Overall, the 226 majority of analytical signals were represented by the ACMTF model in a robust fashion and our results thus 227 indicate that the analytical signals identified by ACMTF scaled linearly with analyte abundance.

228 Chemical properties of fluorescence spectra

229 The six-component ACMTF model featured fluorescence components with emission maxima at 310, 350, 410, 230 420, 460, and 510 nm (Fig. 2, panel A, henceforth referred to by these emission maxima) and molecular 231 formula components with distinctly different molecular weight distributions, elemental composition and 232 overall varying degree of chemodiversity (Fig. 2, panel B-C, Table 1). A comparison with the OpenFluor database³⁴ revealed similarities of all ACMTF fluorescence components with previously identified PARAFAC 233 components (TCC_{ex,em} > 0.98). Specifically, C_{310} , C_{350} , and C_{510} were similar to components identified in the 234 coastal Canadian Arctic.³⁵ However, matches were also observed across other aquatic environments, such as 235 236 the Baltic Sea (C₃₁₀ with C5 in Stedmon et al. [2007]), small streams (C₃₅₀ with C5 in Yamashita et al. [2011],

237 or C_{410} with C1 in Graeber et al. [2012]), or drinking water (C_{420} with C3 in Shutova et al. [2014]).^{36–39}

238 The component weights λ and σ (see SI section S1 for further details) indicated that components were generally shared between both data sets. For C_{410} , C_{420} , C_{460} , and $C_{510} \lambda$ (weights for fluorescence components) and σ 239 240 (weights for molecular formula components) deviated less than 15 % between each other, indicating shared components (Table 1). In contrast, weights differed by almost 80 % for C₃₁₀ and C₃₅₀. While this represents a 241 significant difference, factors other than unshared signals may have contributed to this observation. In addition 242 243 to "sharedness" of components in both data sets, weights in the ACMTF model also reflect contributions of a 244 given component to the overall variability in each dataset. The weights λ and σ are thus influenced by detector 245 response. Furthermore, differences in component complexity may lead to different weights of shared 246 components in separate data sets. In our application, the ionization efficiencies of molecular formulas 247 associated with C_{310} and C_{350} compared to their fluorescence quantum yields may have been partly responsible for different component weights. However, further investigations are necessary to investigate this hypothesis 248

and the response of component weights to analytical factors. Overall, despite C_{310} and C_{330} 's different component weights, these findings indicate that all modeled components were shared between both datasets and therefore represented interpretable, chemically meaningful components that allow the investigation of molecular formulas associated with fluorescence spectra.

ACMTF addresses the multivariate character of molecular formulas by dissecting the abundance of one 253 254 formula into multiple components (Fig. S9), and the identification of the most prominent links between formula 255 and fluorescence described in a specific component is therefore complicated. Here, we propose an approach 256 to simplify the interpretation of ACMTF mass spectra for the purpose of initial investigations: The molecular formula loadings (Fig. 2B-C) can be simplified by identifying the component with the highest relative loading 257 258 for every molecular formula while disregarding the remaining components. In the resulting modified component mass spectra, every molecular formula is only represented once and the interpretation is simplified 259 260 (Fig. S9 depicts several examples).

When plotting modified mass spectra in the van Krevelen space (Fig. 3), the chemical properties associated 261 with fluorescence spectra clustered in specific regions. The modified component loadings tracked a continuum 262 263 of chemical properties along a diagonal line in the van Krevelen space from the most saturated formulas (high 264 H/C, low O/C) to the most oxygenated, unsaturated formulas (high O/C, low H/C) where oxygenation and unsaturation increased in the order $C_{460} < C_{310} < C_{350} < C_{410} < C_{420} < C_{510}$. This shift of elemental composition 265 between components observed in the modified component spectra accurately reflected the properties of the 266 unmodified components revealing identical shifts in the weighted averages of O/C and H/C (Table 1). This 267 268 indicates that the shift is an inherent pattern and not an artefact of the simplification of the molecular fingerprint of each component to modified loadings. However, as would be expected, there was a large discrepancy 269 between the chemodiversity of molecular formula components (Table 1) and the number of formulas 270 271 represented in modified mass spectra (148 > N > 42, Fig. 3). Despite considerable data reduction, modified 272 mass spectra appear to offer adequate insight into the compositional differences between components, but further in-depth investigations of chemical properties require the consideration of the complete component 273 274 molecular fingerprint and thus all further model interpretation refer to the loadings depicted in Fig. 2C.

275 Because ACMTF indicated that all identified fluorescence components were linked (i.e. shared) with distinctly 276 different mass spectra, our study allowed the first multivariate estimate of the molecular fingerprint of 277 fluorescent organic matter. As can be seen in Fig. 2C, the chemodiversity and elemental composition of 278 components differed significantly, but fluorescence properties such as emission maximum or Stokes Shift did not correlate with molecular properties such as chemodiversity, weighted average elemental composition, or 279 280 weighted average molecular weight. This suggests that while EEM fluorescence was possibly caused by the excitation of chemical moieties summarized in the molecular fingerprint of ACMTF components, these 281 282 fluorescing moieties do not dominate the molecular fingerprint to an extent that would allow the usage of 283 fluorescence properties as indicator for, or predictor of, associated molecular fractions (and vice versa). 284 Moreover, it should not be assumed that the molecular formulas summarized in ACMTF components 285 predominantly consist moieties that fluoresce, but rather that their dynamics are indistinguishable from the moieties that do. However, if further studies should reveal consistent trends in the molecular fingerprint of 286 287 fluorescent DOM, the stable, albeit non-causal relationship could be utilized routinely to expand the analytical 288 window of UV-Vis spectroscopic analyses.

289 Comparison of data fusion and traditional approaches

290 Since the present study is the first application of ACMTF to analytical DOM data sets, it is important to

291 compare the results obtained with ACMTF to existing approaches employed in previous studies. First, the

292 factorization of molecular formulas with ACMTF was compared to PCA, a widely used method to decompose

293 molecular formula tables.¹⁵ The first significant difference between PCA and ACMTF is the ability to impose

294 non-negativity constraints in ACMTF. Our analysis indicated that loadings and scores of PC1-PC6 in PCA

- 295 were both positive and negative (Fig. S10), while ACMTF component loadings were exclusively positive (Fig.
- 296 2, panel C). In ACMTF, component loadings and scores therefore directly correspond to analytical signals,

297 while in the case of PCA, the interpretation of components is less intuitive since combination of negative and

298 positive scores and loadings must be considered. The option to constrain models to nonnegative loadings and

299 scores in ACMTF represents an improvement in the characterization of complex mixtures such as DOM with

300 mass spectrometry.

Compared to the strict orthogonality of components in PCA, ACMTF allows some degree of similarity 301 302 between molecular formula loadings. A congruence analysis between all unique combination of the six 303 ACMTF components showed that ACMTF loadings of molecular formula components were autocorrelated to 304 some degree (between 0.21 and 0.85, on a scale from zero to one, Fig. S11). The similarity between ACMTF 305 components highlights that environmental processes or independent chemical fractions may overlap in their 306 spectral properties. Models with strictly orthogonal components (such as PCA) would ultimately be unable to 307 recover these spectra. Together, our comparison indicated that ACMTF provides more chemically intuitive 308 results making it a more appropriate model for the decomposition of molecular formula matrices of DOM. 309 However, it is important to stress that ACMTF models are primarily driven by the variability of the tensor (fluorescence EEMs). The description of variability beyond EEMs depends on fitting unshared components 310 311 that can describe variability independent of fluorescence. Future efforts should also include the comparison of 312 components derived from data fusion using ACMTF and those derived from factorizations based solely on 313 molecular formulas to investigate how data fusion models relate to models describing only molecular formulas. 314 However, this is pending the validation of non-negative matrix factorizations applicable to molecular formula

315 data sets of DOM, and a comparison is thus not yet possible.

316 Since ACMTF is based on PARAFAC, there should be a basic agreement between loadings and scores of a 317 PARAFAC model fit to the EEM data exclusively, and the ACMTF model describing both EEMs and 318 molecular formulas. A detailed comparison between ACMTF and PARAFAC is given in the SI (SI S2, Fig. 319 S8). In short, ACMTF loadings were highly congruent with corresponding PARAFAC components, while the 320 scores of some components (C420, C460) diverged from the PARAFAC solution despite clearly showing a positive correlation. These discrepancies are most likely attributable to the occurrence of mass spectrometry-321 322 specific disturbances. Amongst other possibilities, matrix effects in ESI-MS affecting certain groups of 323 samples could have caused deviations in component scores of shared components. Despite this, the 324 simultaneous factorization of fluorescence EEMs and the molecular formula table resulted in a model that 325 generally agreed with PARAFAC.

326 Finally, we compared associations between fluorescence EEMs and molecular formulas identified by ACMTF 327 with the correlations identified by the post-hoc Pearson rank correlation of PARAFAC F_{max} -values and 328 molecular formula abundances (Fig. 2, panel D). Although, there was a degree of overlap between associations 329 identified by each approach (Fig. 2, panel C-D), there was also substantial disagreement between the findings. 330 The correlation-weighted mass spectra (showing correlation coefficient in place of abundances) were 331 compared to ACMTF component mass spectra, using the Tucker congruence coefficient. For all comparisons, 332 TCCs were smaller than 0.45, indicating poor agreement. In extreme cases, such as C₃₁₀, C₃₅₀, and C₄₆₀, correlations were found to be inverse in areas of the van Krevelen space that showed relatively low, but positive 333 loadings in corresponding ACMTF components. In the case of C₅₁₀ was inversely correlated with highly 334 335 oxygenated, highly unsaturated molecular formulas in the rank analysis, while ACMTF indicated a strong 336 positive correlation in that compositional space.

337 Comparisons between data fusion and rank correlation are hindered by some key differences between the 338 correlation and data fusion approaches. Rank correlations in their simplest interpretation provide a binary 339 indication of direct or inverse association between fluorescence and molecular formulas. This approach does 340 not decompose a multivariate signal and often returns multiple possible correlations between variables. This 341 leaves users to decide which of the significant correlations are valid, and which can (or should) be ignored. On 342 the other hand, data fusion readily addresses this issue by dissecting a particular molecular formula abundance 343 into multiple components. This multivariate decomposition represents a more objective approach that is far 344 less vulnerable to false positive errors, and thus provides more robust estimates of the molecular fingerprint of 345 fluorescent DOM.

346 Biogeochemical sources, chemical fractions, or diagenetic state?

347 DOM is a highly complex mixture of organic compounds with contrasting chemical properties, varying 348 biogeochemical sources and sinks, and different degradation potential. It is striking that more than 99 % of the 349 modeled mass spectral variability (~38% of the mass spectra) was described by only six statistical components. 350 This represents a significant reduction in complexity that facilitates a far easier interpretation of DOM data 351 sets. However, the partitioning of DOM fluorescence and molecular formulas into statistical components poses 352 the question: Do ACMTF components indicate biogeochemical source materials (e.g. terrestrial substances), 353 reflect its diagenetic state (e.g. recalcitrant material), or represent distinct chemical fractions (e.g. high 354 molecular weight DOM)?

355 To assign an interpretation to the six ACMTF components, across-fjord patterns (Fig. 4) and depth-dependent trends in Scoresby Sound (Fig. 5) were investigated whilst bearing in mind the molecular fingerprint of 356 357 respective components (Fig. 2C). The most distinct oceanographic pattern was observed for C₄₂₀, which was 358 most prevalent in Scoresby Sound (Fig. 4). In Scoresby Sound, Fmax-values of C420 also exhibited a distinct surface maximum, followed by a decrease in its relative abundance with depth and a subsurface peak at all 359 360 stations in the polar waters (S < 34, T < 0) was observed (Fig. 5). Since Scoresby Sound receives terrestrial material from the EGC, terrestrial material is most likely a major source of this component. However, because 361 a recent study identified a component highly similar to C420 as ubiquitous across a wide range of 362 environments,⁴⁰ C₄₂₀ is likely not a highly selective proxy for terrestrial material, but rather represents a 363 fluorescence fraction that is particularly abundant in terrestrially influenced waters. Interestingly, C₄₂₀ had the 364 365 lowest average molecular weight, was strongly associated with the most oxygenated, unsaturated molecular 366 formulas, and showed the highest chemodiversity, suggesting that terrestrially derived substances represent a 367 distinguishable chemical fraction. The identification of this terrestrially dominated component provides targets 368 for further experiments, for example via MS/MS to explore molecular structures.

369 Components fluorescing in the spectral range of C₃₅₀ are commonly termed "protein-like"; a term which is 370 derived from the apparent spectral similarity with amino acid fluorescence that has been shown to correlate with amino acid concentration.⁴¹ The presence of an UV-A fluorescence signature has often been used as a 371 proxy for surface water biological activity and recent studies have demonstrated that such processes impact 372 DOM mass spectra with high selectivity.⁴² In agreement with these findings, ACMTF indicated a low 373 chemodiversity of the molecular fingerprint associated with C₃₅₀, which can be interpreted as representing 374 375 fresh organic material closely linked to planktonic productivity in the surface layer (Fig. 5). Similar to C₃₅₀, C_{310} also exhibited a subsurface decrease in Scoresby Sound (10 – 30 m), but slightly increased again at 45 m. 376 377 The higher chemodiversity of C₃₁₀ compared to C₃₅₀ suggests that C₃₁₀ may encompass degradation products 378 related to lateral terrestrial inputs, plankton productivity, or (photo-)degradation in the surface layer. A more 379 constrained assignment is not possible at present and would require further experiments.

For the remaining components, across-fjord differences were subtle, while depth profiles differed between components. The composition in Arnarfjörður and Kongsfjorden was relatively stable with depth (Figs. S12-13), whereas distinct changes were observed in Scoresby Sound (Fig. 5). C_{410} and C_{510} increased with depth, suggesting that this moderately complex fingerprint was generated by the processing of sinking organic matter. In contrast, C_{460} was invariant with water depth in all three systems, which suggests it may represent a recalcitrant component of high molecular complexity (Fig. 5, also see Figure S12-13). These systematic trends provide confidence in the components identified by this data fusion approach.

387 Challenges and future directions

388 Multivariate data fusion simultaneously decomposes multiple analytical datasets and provides a tool to link 389 heterogeneous datasets, such as fluorescence spectroscopy and mass spectrometry. The methodological 390 similarities with PARAFAC provide the opportunity to integrate data fusion algorithms into popular existing 391 software toolboxes (such as drEEM), as well as related public databases (such as OpenFluor). With the 392 continuous development of multi-core processors and improvements in modeling algorithms, the 393 computational expense of data fusion will greatly decrease in the future, making it widely applicable for the 394 scientific community. While multivariate data fusion may not be able to identify the chemical compounds 395 responsible for the optical properties of complex environmental DOM datasets, it greatly improves the 396 chemical interpretability of fluorescence data sets and offers potential for future developments. With data 397 fusion as central interface, future studies will be able to leverage the superior analytical depth of mass 398 spectrometry while utilizing the spatio-temporal resolution of ultraviolet-visible spectroscopy.

399 Assigning an interpretation to ACMTF components is confounded by the increased information content of 400 components, particularly the high complexity of molecular fingerprints. Whereas fluorescence components are 401 continuous, and their chemical interpretation may generally be assessed by comparison to pure fluorophore 402 spectra, the complex and discontinuous nature of component mass spectra constitutes a significant challenge in this regard. In light of their generally high diversity, it appears reasonable to assume that ACMTF component 403 404 mass spectra represent multiple, currently unresolvable chemical fractions. The ability to further separate 405 chemical fractions may be improved by integrating additional DOM analyses, such as absorbance, ¹³C, or ¹H 406 nuclear magnetic resonance (NMR) spectroscopy. ACMTF is theoretically able to link one trilinear data set (fluorescence EEMs) with multiple matrices (such as ¹³C and ¹H NMR). However, ACMTF requires that one 407 408 of the data structures is trilinear. The fusion of bilinear data sets, such as FT-ICR-MS and ¹³C NMR requires 409 other data fusion strategies.43

410 There are a number of methodological challenges that remain to be solved in regard to the applicability of ACMTF to DOM data sets. For example, it is very likely that FDOM EEMs and FT-ICR-MS formula matrices 411 do not share the same number of underlying components. While ACMTF is able to address this disparity 412 between two datasets with dataset- and component-specific weights, it is currently unknown how well this 413 approach can cope with the potentially large discrepancy between the number of components in fluorescence 414 415 vs. molecular formula datasets of DOM. An expanded analysis of this is warranted and this topic needs 416 consideration in future efforts. Furthermore, our study put focus on the 527 most conserved molecular formulas since the primary goal was to identify the molecular fingerprint of quasi-ubiquitous fluorescence signals. 417 418 Relaxing the data pre-processing criteria by including more unique (uncommon) molecular formulas would be 419 desirable from a biogeochemical point-of-view. However, this presents a challenge for model validation and 420 considerably increases computation time. In order to include sparsely observed molecular formulas, further 421 developments to the analysis approach are necessary.

422 The scientific community utilizing FDOM EEMs has converged towards a standard methodology for measuring samples and analyzing resulting data sets with PARAFAC, achieved through a substantial number 423 of efforts including inter-laboratory comparisons.^{12,23,44} This was driven mostly by the fact that multivariate 424 425 analysis demands a stringent and standardized sample and data processing routine, to provide globally 426 consistent data. The measurement and data analysis of FT-ICR-MS mass spectra is in the process of being standardized and this is essential if advanced data analysis techniques are to be employed.⁴⁵ For the time being, 427 428 however, different guidelines for formula assignment, peak identification, and signal normalization exist across the community and no central, open database has been developed to our knowledge. If data fusion 429 430 should become a viable tool for community-wide DOM characterization, these discrepancies must be 431 addressed to provide reproducible results that enable replication.

432 Key differences between fluorescence spectroscopy and FT-ICR-MS currently complicate joint description of fluorescence and ESI-MS data sets. Fluorescence quantum yields and ESI efficiencies of analytes are 433 434 unknown; therefore, peak intensities and proportions between peaks reflect a combination of actual 435 concentrations and differences in fluorescence and ionization efficiencies, respectively. However, ESI-MS 436 peak intensities differ from fluorescence, because carbon concentrations of samples are typically adjusted 437 before injection, and signals are normalized to the sum of peaks (or the highest peak) during post-processing. 438 These steps are designed to improve data robustness, but peak abundances subsequently depend on: (1) 439 ionization efficiencies; and (2) the abundance of the remaining peaks in relation to the total carbon 440 concentration. To ensure the compatibility of the two data sets during data fusion, fluorescence signals must be normalized to the total fluorescence per sample. As a result, ACMTF scores represent proportions. 441

442 Obtaining *more* quantitative ACMTF component scores hinges on developing approaches that yield robust 443 mass spectra without a carbon- and peak-normalization step.

Beyond issues related to sample and data treatment, practical issues related to the ion source may further 444 compromise the ability to quantify analytes with ESI-MS. The ionization of DOM constituents with ESI is 445 inherently selective and matrix effects introduce artefacts.^{3,46} However, the extent to which matrix effects 446 impact DOM mass spectra remains poorly quantified as the molecular structures of DOM remain largely 447 uncharacterized. Nonetheless, matrix effects have been documented for isolated marine metabolites.⁴⁷ The 448 449 application of models assuming linear relationships between analyte concentration and detector signal thus requires careful investigation. Here, we observed that the majority of the 527 modeled formulas could be 450 451 described using a model that assumes a linear relationship between fluorescence signals and molecular formula 452 abundances. The reproducibility of our efforts should be investigated in future studies carried out across a wide 453 variety of aquatic environments.

454 Acknowledgements

455 This study was in part funded by the Danish Council for Independent Research-Natural Sciences Grant DFF-1323-00336 and Nordic5Tech collaborative funding (Technical University of Denmark). We greatly 456 457 appreciate the constructive comments by four anonymous reviewers and the editor Professor Dovichi, which helped to improve this manuscript. We thank the captain and crew of R/V Maria S. Merian for their support 458 during cruise MSM56. Mourad Harir (Helmholtz Zentrum München), Oliver J. Lechtenfeld (Helmholtz-459 Zentrum für Umweltforschung UFZ), Claudia Burau, and Jana K. Geuer (both Alfred Wegener Institute for 460 461 Polar and Marine Research) are acknowledged for their help during sampling, extract preparations, and FT-ICR-MS measurements. UJW thanks Rasmus Bro (Copenhagen University) for helpful discussions on the 462 463 ACMTF model. EA and UJW thank Michael Saunders (Stanford University) for kindly providing access to 464 the SNOPT toolbox.

465 Supporting information

466 Supplementary results containing information on model validation (S1), twelve supplementary figures (S2),

467 and two supplementary tables (S3).

468 **References**

- 469 (1) Buchan, A.; LeCleir, G. R.; Gulvik, C. A.; González, J. M. Master Recyclers: Features and
 470 Functions of Bacteria Associated with Phytoplankton Blooms. *Nat. Rev. Microbiol.* 2014, *12*471 (10), 686–698.
- 472 (2) Mopper, K.; Kieber, D. J.; Stubbins, A. Marine Photochemistry of Organic Matter: Processes
 473 and Impacts. Processes and Impacts., Second Edi.; Elsevier Inc., 2014.
- 474 (3) Repeta, D. J. Chemical Characterization and Cycling of Dissolved Organic Matter. In
 475 *Biogeochemistry of Marine Dissolved Organic Matter*; Elsevier Inc., 2015; pp 21–63.
- 476 (4) Carlson, C. A.; Hansell, D. A. DOM Sources, Sinks, Reactivity, and Budgets. In
 477 *Biogeochemistry of Marine Dissolved Organic Matter: Second Edition*; Elsevier Inc., 2014;
 478 pp 65–126.
- Koch, B. P.; Witt, M.; Engbrodt, R.; Dittmar, T.; Kattner, G. Molecular Formulae of Marine and Terrigenous Dissolved Organic Matter Detected by Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Geochim. Cosmochim. Acta* 2005, 69 (13), 3299–3308.
- 483 (6) Hertkorn, N.; Frommberger, M.; Witt, M.; Koch, B. P.; Schmitt-Kopplin, P.; Perdue, E. M.
 484 Natural Organic Matter and the Event Horizon of Mass Spectrometry. *Anal. Chem.* 2008, *80*485 (23), 8908–8919.
- 486 (7) Mopper, K.; Stubbins, A.; Ritchie, J. D.; Bialk, H. M.; Hatcher, P. G. Advanced Instrumental
 487 Approaches for Characterization of Marine Dissolved Organic Matter: Extraction Techniques,
 488 Mass Spectrometry, and Nuclear Magnetic Resonance Spectroscopy. *Chem. Rev.* 2007, *107*489 (2), 419–442.
- 490 (8) Fellman, J. B.; Hood, E.; Spencer, R. G. M. Fluorescence Spectroscopy Opens New Windows
 491 into Dissolved Organic Matter Dynamics in Freshwater Ecosystems: A Review. *Limnol.*492 *Oceanogr.* 2010, 55 (6), 2452–2462.
- 493 (9) Coble, P. G. Characterization of Marine and Terrestrial DOM in Seawater Using Excitation
 494 Emission Matrix Spectroscopy. *Mar Chem* 1996, *51* (4), 325–346.
- 495 (10) Stedmon, C. A.; Nelson, N. B. The Optical Properties of DOM in the Ocean. In
 496 *Biogeochemistry of Marine Dissolved Organic Matter*; Hansell, D. A., Carlson, C. A., Eds.;
 497 Elsevier Inc.: San Diego, 2015; pp 481–508.
- 498 (11) Ferdinand, O. D.; Friedrichs, A.; Miranda, M. L.; Voß, D.; Zielinski, O. Next Generation
 499 Fluorescence Sensor with Multiple Excitation and Emission Wavelengths NeXOS. In
 500 OCEANS 2017 Aberdeen; IEEE: Aberdeen, 2017; pp 1–6.
- Murphy, K. R.; Stedmon, C. A.; Graeber, D.; Bro, R. Fluorescence Spectroscopy and Multi Way Techniques. PARAFAC. *Anal. Methods* 2013, 5 (23), 6557–6566.
- Rosario-Ortiz, F. L.; Korak, J. A. Oversimplification of Dissolved Organic Matter
 Fluorescence Analysis: Potential Pitfalls of Current Methods. *Environ. Sci. Technol.* 2017, 51

- 505 (2), 759–761.
- 506 (14) Stenson, A. C.; Marshall, A. G.; Cooper, W. T. Exact Masses and Chemical Formulas of
 507 Individual Suwannee River Fulvic Acids from Ultrahigh Resolution Electrospray Ionization
 508 Fourier Transform Ion Cyclotron Resonance Mass Spectra. *Anal. Chem.* 2003, 75 (6), 1275–
 509 1284.
- (15) Sleighter, R. L.; Liu, Z.; Xue, J.; Hatcher, P. G. Multivariate Statistical Approaches for the
 Characterization of Dissolved Organic Matter Analyzed by Ultrahigh Resolution Mass
 Spectrometry. *Environ. Sci. Technol.* 2010, 44 (19), 7576–7582.
- 513 (16) Acar, E.; Gürdeniz, G.; Savorani, F.; Hansen, L.; Olsen, A.; Tjønneland, A.; Dragsted, L. O.;
 514 Bro, R. Forecasting Chronic Diseases Using Data Fusion. J. Proteome Res. 2017, 16 (7), 2435–
 515 2444.
- 516 (17) Stubbins, A.; Lapierre, J.-F.; Berggren, M.; Prairie, Y. T.; Dittmar, T.; del Giorgio, P. A.
 517 What's in an EEM? Molecular Signatures Associated with Dissolved Organic Fluorescence in
 518 Boreal Canada. *Environ. Sci. Technol.* 2014, 48, 10598–10606.
- Herzsprung, P.; Von Tümpling, W.; Hertkorn, N.; Harir, M.; Büttner, O.; Bravidor, J.; Friese,
 K.; Schmitt-Kopplin, P. Variations of DOM Quality in Inflows of a Drinking Water Reservoir:
 Linking of van Krevelen Diagrams with EEMF Spectra by Rank Correlation. *Environ. Sci. Technol.* 2012, 46 (10), 5511–5518.
- Acar, E.; Papalexakis, E. E.; Gürdeniz, G.; Rasmussen, M. A.; Lawaetz, A. J.; Nilsson, M.;
 Bro, R. Structure-Revealing Data Fusion. *BMC Bioinformatics* 2014, *15* (239), 1–17.
- Acar, E.; Bro, R.; Smilde, A. K. Data Fusion in Metabolomics Using Coupled Matrix and
 Tensor Factorizations. *Proc. IEEE* 2015, *103* (9), 1602–1620.
- Acar, E.; Levin-Schwartz, Y.; Calhoun, V. D.; Adalı, T. Tensor-Based Fusion of EEG and
 FMRI to Understand Neurological Changes in Schizophrenia. *Proc. IEEE Int. Symp. Circuits Syst.* 2017, No. 1612.02189.
- 530 (22) Dittmar, T.; Koch, B. P.; Hertkorn, N.; Kattner, G. A Simple and Efficient Method for the
 531 Solid-Phase Extraction of Dissolved Organic Matter (SPE-DOM) from Seawater. *Limnol.*532 Ocean. Methods 2008, 6, 230–235.
- Kothawala, D. N.; Murphy, K. R.; Stedmon, C. A.; Weyhenmeyer, G. A.; Tranvik, L. J. Inner
 Filter Correction of Dissolved Organic Matter Fluorescence. *Limnol. Oceanogr. Methods* **2013**, *11* (DEC), 616–630.
- 536 (24) Flerus, R.; Lechtenfeld, O. J.; Koch, B. P.; McCallister, S. L.; Schmitt-Kopplin, P.; Benner,
 537 R.; Kaiser, K.; Kattner, G. A Molecular Perspective on the Ageing of Marine Dissolved
 538 Organic Matter. *Biogeosciences* 2012, 9 (6), 1935–1955.

Lechtenfeld, O. J.; Koch, B. P.; Gašparović, B.; Frka, S.; Witt, M.; Kattner, G. The Influence
of Salinity on the Molecular and Optical Properties of Surface Microlayers in a Karstic Estuary. *Mar. Chem.* 2013, *150*, 25–38.

- 542 (26) Acar, E.; Kolda, T. G.; Dunlavy, D. M. All-at-Once Optimization for Coupled Matrix and
 543 Tensor Factorizations. In *KDD Workshop on Mining and Learning with Graphs*;
 544 https://arxiv.org/abs/1105.3422, 2011.
- 545 (27) Bader, B. W.; Kolda, T. G. MATLAB Tensor Toolbox. Available online at 546 https://www.tensortoolbox.org. 2012.
- 547 (28) Gill, P. E.; Murray, W.; Saunders, M. A. SNOPT: An SQP Algorithm for Large-Scale 548 Constrained Optimization. *SIAM Rev.* **2002**, *47* (1), 99–131.
- Kellerman, A. M.; Dittmar, T.; Kothawala, D. N.; Tranvik, L. J. Chemodiversity of Dissolved
 Organic Matter in Lakes Driven by Climate and Hydrology. *Nat. Commun.* 2014, 5 (May), 1–
 8.
- (30) Chao, A.; Shen, T.-J. Nonparametric Estimation of Shannon's Index of Diversity When There
 Are Unseen Species in Sample. *Environ. Ecol. Stat.* 2003, *10*, 429–443.
- Wünsch, U. J.; Geuer, J. K.; Lechtenfeld, O. J.; Koch, B. P.; Murphy, K. R.; Stedmon, C. A.
 Quantifying the Impact of Solid-Phase Extraction on Chromophoric Dissolved Organic Matter
 Composition. *Mar. Chem.* 2018, doi: 10.1016/j.marchem.2018.08.010.
- Lorenzo-Seva, U.; ten Berge, J. M. F. Tucker's Congruence Coefficient as a Meaningful Index
 of Factor Similarity. *Methodology* 2006, 2 (2), 57–64.
- Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. Scand. J. Stat. 1979, 6 (2),
 65–70.
- Murphy, K. R.; Stedmon, C. A.; Wenig, P.; Bro, R. OpenFluor- an Online Spectral Library of
 Auto-Fluorescence by Organic Compounds in the Environment. *Anal. Methods* 2014, 6 (3),
 658–661.
- (35) Walker, S. A.; Amon, R. M. W.; Stedmon, C. A.; Duan, S.; Louchouarn, P. The Use of
 PARAFAC Modeling to Trace Terrestrial Dissolved Organic Matter and Fingerprint Water
 Masses in Coastal Canadian Arctic Surface Waters. J. Geophys. Res. Biogeosciences 2009,
 114 (4), 1–12.
- 568 (36) Shutova, Y.; Baker, A.; Bridgeman, J.; Henderson, R. K. Spectroscopic Characterisation of
 569 Dissolved Organic Matter Changes in Drinking Water Treatment: From PARAFAC Analysis
 570 to Online Monitoring Wavelengths. *Water Res.* 2014, 54, 159–169.
- 571 (37) Graeber, D.; Gelbrecht, J.; Pusch, M. T.; Anlanger, C.; von Schiller, D. Agriculture Has
 572 Changed the Amount and Composition of Dissolved Organic Matter in Central European
 573 Headwater Streams. *Sci. Total Environ.* 2012, *438*, 435–446.
- 574 (38) Yamashita, Y.; Kloeppel, B. D.; Knoepp, J.; Zausen, G. L.; Jaffé, R. Effects of Watershed
 575 History on Dissolved Organic Matter Characteristics in Headwater Streams. *Ecosystems* 2011,
 576 14 (7), 1110–1122.
- 577 (39) Stedmon, C. A.; Markager, S.; Tranvik, L. J.; Kronberg, L.; Slätis, T.; Martinsen, W.
 578 Photochemical Production of Ammonium and Transformation of Dissolved Organic Matter in

- 579 the Baltic Sea. *Mar. Chem.* **2007**, *104* (3–4), 227–240.
- Murphy, K. R.; Timko, S. A.; Gonsior, M.; Powers, L. C.; Wünsch, U. J.; Stedmon, C. A.
 Photochemistry Illuminates Ubiquitous Organic Matter Fluorescence Spectra. *Environ. Sci. Technol.* 2018, 52 (19), 11243–11250.
- 583 (41) Yamashita, Y.; Tanoue, E. Chemical Characterization of Protein-like Fluorophores in DOM in
 584 Relation to Aromatic Amino Acids. *Mar. Chem.* 2003, 82 (3–4), 255–271.
- (42) Reader, H. E.; Stedmon, C. A.; Nielsen, N. J.; Kritzberg, E. S. Mass and UV-Visible Spectral
 Fingerprints of Dissolved Organic Matter: Sources and Reactivity. *Front. Mar. Sci.* 2015, 2
 (October), 1–10.
- 588 (43) Smilde, A. K.; Måge, I.; Næs, T.; Hankemeier, T.; Lips, M. A.; Kiers, H. A. L.; Acar, E.; Bro,
 589 R. Common and Distinct Components in Data Fusion. J. Chemom. 2017, 31 (7), 1–20.
- Murphy, K. R.; Butler, K. D.; Spencer, R. G. M.; Stedmon, C. A.; Boehme, J. R.; Aiken, G. R.
 Measurement of Dissolved Organic Matter Fluorescence in Aquatic Environments: An Interlaboratory Comparison. *Environ. Sci. Technol.* 2010, 44 (24), 9405–9412.
- Leefmann, T.; Frickenhaus, S.; Koch, B. P. UltraMassExplorer a Browser-Based Application
 for the Evaluation of High-Resolution Mass Spectrometric Data. *Rapid Commun. Mass Spectrom.* 2018, doi: 10.1002/rcm.8315.
- 596 (46) Taylor, P. J. Matrix Effects: The Achilles Heel of Quantitative High-Performance Liquid
 597 Chromatography-Electrospray-Tandem Mass Spectrometry. *Clin. Biochem.* 2005, *38* (4), 328–
 598 334.
- Johnson, W. M.; Kido Soule, M. C.; Kujawinski, E. B. Extraction Efficiency and
 Quantification of Dissolved Metabolites in Targeted Marine Metabolomics. *Limnol. Oceanogr. Methods* 2017, *15* (4), 417–428.

602

603 Tables

Table 1: Properties of ACMTF components. Component weights (λ and σ) indicate the contribution of a component to the overall variability and indicate their sharedness across EEMs and FT-ICR-MS. Weighted averages (subscript wa) refer to the weighted average chemical composition of specific components. O/C: Oxygen-to-carbon ratio, H/C: Hydrogen-to-carbon ratio. m/z: mass-to-charge ratio.
 DBE: Double bond equivalent. N/C: Nitrogen-to-carbon-ratio. Ci: Chemodiversity index.

Comp.	λ (EEMs)	σ (FT-ICR-MS)	Stoke's shift (eV)	O/Cwa	H/Cwa	m/zwa	DBEwa	N/Cwa	Ci
C310	0.22 ± 0.005	0.12 ± 0.002	0.56	0.5	1.23	399.7	8.44	0.04	422
C350	0.24 ± 0.003	0.13 ± 0.007	0.8	0.52	1.24	382.6	7.8	0.02	314
C410	0.46 ± 0.007	0.40 ± 0.005	0.94	0.55	1.21	400	8.36	0.03	400
C420	0.15 ± 0.002	0.15 ± 0.002	1.13	0.56	1.16	382	8.27	0.01	459
C460	0.27 ± 0.003	0.32 ± 0.004	0.72	0.45	1.27	388.2	8.01	0.03	422
C510	0.19 ± 0.002	0.17 ± 0.003	0.61	0.6	1.15	390.2	8.35	0.02	355

608

609

610 Figure legends

611



612 Figure 1: Example of measured vs. modeled data. (A-C): Fluorescence EEMs (scaled and therefore unitless). (D-F): FT-ICR-MS

613 mass spectra scaled by the maximum peak intensity. First column: Examples of raw data. Second column: Examples of corresponding 614 modeled data. Third column: Model residuals, scaled by the maximum peak intensity in the sample. The depicted sample was taken in

615 Kongsfjorden, Norwegian monitoring station Kb5 (78.9 °N, 12.4 °E) at a depth of 30 m.



Figure 2: The molecular fingerprint of fluorescent DOM as identified by ACMTF and post-hoc correlation. Panel (A) depicts
 fluorescence loadings as a function of excitation (dashed line) and emission (solid line), panel (B) and (C) show molecular formula
 loadings as function of mass-to-charge ratio and molecular composition in the van Krevelen space. For comparison, the correlation

between PARAFAC component scores (spectrally congruent with components in top row, also provided in Figure S7) are shown in panel (**D**). Linear correlation coefficients were restricted to p<0.01 prior to visualization and are corrected for false-positives using a

622 Holm-Bonferroni correction.



Figure 3: Modified component mass spectra of ACMTF components. For the purpose of simplification, every formula is only represented when it dominates the loading of a particular component (i.e. has the highest loading). Panel (A): Modified component molecular weight distributions. Panel (B): Modified component van Krevelen plots, where the size of dots represents the component formula is.







632

Figure 5: Depth variation of physical and environmental parameters in Scoresby Sound. Panel (A): Salinity, temperature, and chlorophyll a, along with a station map. Panel (B): Depth profiles of the six ACMTF component F_{max} -values. Dots represent F_{max} values, black lines the binned average along with the standard deviation of the mean (grey).

values, black lines the billied average along with the standard deviation of the mean (gi



