Environmental Science & Technology

Contaminants in Aquatic and Terrestrial Environments

Improved understanding of dissolved organic matter processing in freshwater using complementary experimental and machine learning approaches

Peter Herzsprung, Valerie Wentzky, Norbert Kamjunke, Wolf von Tumpling, Christin Wilske, Kurt Friese, Bertram Boehrer, Thorsten Reemtsma, Karsten Rinke, and Oliver Jens Lechtenfeld

Environ. Sci. Technol., **Just Accepted Manuscript** • DOI: 10.1021/acs.est.0c02383 • Publication Date (Web): 23 Sep 2020

Improved understanding of dissolved organic matter processing in freshwater using complementary experimental and machine learning approaches

Peter Herzsprung,[*,†] *Valerie Wentzky,*[†,#] *Norbert Kamjunke,*[‡] *Wolf von Tümpling,*[‡] *Christin Wilske,*[‡] *Kurt Friese,*[†] *Bertram Boehrer,*[†] *Thorsten Reemtsma,*[§,⊥] *Karsten Rinke,*[†] *and Oliver J. Lechtenfeld* [*,§,‖]

AUTHOR ADDRESS

† UFZ – Helmholtz Centre for Environmental Research, Department Lake Research, Brückstraße 3a, D-39114 Magdeburg, Germany

State Agency for Agriculture, Environment and Rural Areas Schleswig-Holstein - LLUR, Lake Department, Hamburger Chaussee 25, D-24220 Flintbek, Germany

‡ UFZ – Helmholtz Centre for Environmental Research, Department River Ecology, Brückstraße 3a, D-39114 Magdeburg, Germany

§ UFZ – Helmholtz Centre for Environmental Research, Department Analytical Chemistry, Permoserstr. 15, D-04318 Leipzig, Germany

⊥ Institute of Analytical Chemistry, University of Leipzig, D-04103, Leipzig, Germany

18 ‖ UFZ – Helmholtz Centre for Environmental Research, ProVIS – Centre for
19 Chemical Microscopy, Permoserstr. 15, D-04318 Leipzig, Germany

20

21 KEYWORDS

22 DOM, molecular reactivity, microbial processes, photochemical transformations,

23 drinking water reservoir, ultra-high resolution mass spectrometry, machine learning,

24 predictive models

25 ABSTRACT

26 Dissolved organic matter plays an important role in aquatic ecosystems and poses a
27 major problem for drinking water production. However, our understanding of DOM
28 reactivity in natural systems is hampered by its complex molecular composition. Here,
29 we used Fourier-transform ion cyclotron resonance mass spectrometry (FT-ICR-MS)
30 and data from two independent studies to disentangle DOM reactivity based on
31 photochemical and microbial induced transformations. Robust correlations of FT-ICR-
32 MS peak intensities with chlorophyll a and solar irradiation were used to define 9
33 reactivity classes for 1277 common molecular formulas. Germany`s largest drinking
34 water reservoir was sampled for one year and DOM processing in stratified surface

waters could be attributed to photochemical transformation during summer months. Microbial DOM alterations could be distinguished based on correlation coefficients with chlorophyll a and often shared molecular features (elemental ratios, mass) with photo reactive compounds. Specifically, many photo products and some microbial products were identified as potential precursors of disinfection byproducts. Molecular DOM features were used to further predict molecular reactivity for the remaining compounds in the data set based on a random forest model. Our method offers an expandable classification approach to integrate reactivity of DOM from specific environments and link it to molecular properties and chemistry.

1 Introduction

Dissolved organic matter (DOM) sources in surface waters can be classified as allochthonous or autochthonous organic material.[1] Both sources contribute to a large molecular diversity of the DOM pool.[2] The molecular composition and properties of DOM play a critical role in a number of biogeochemical processes such as metal redox cycling,[3] microbial growth, maintenance of community structure,[4] and photochemical processes.[5-7] A multitude of processes, e.g. microbial production (exudation from

autotrophic algae) and (heterotrophic) degradation, photochemical reactions (photo production and photo degradation), leaching of organic material from plant debris, soils or sediments, can in turn alter the molecular composition of freshwater DOM.[3,6,8,9] In natural systems, these processes are often difficult to disentangle due to our limited analytical abilities to distinguish multiple concurrently acting molecular transformations of DOM.[10]

In surface waters of lakes and reservoirs for example, solar radiation enhances both, autotrophic processes and photochemical reactions.[11] Algal growth can be monitored e.g. by chlorophyll a (Chl-a) concentrations and results in a direct release of DOM as exudates and a release of DOM degradation products via heterotrophic processes.[8] In addition, photo products are also formed from pre-existing or newly released DOM molecules. All these processes may be positively correlated with Chl-a concentration (or cell density, solar irradiation dose, or others). In order to discriminate photo products from microbially derived products, additional data are required, e.g. on molecular DOM properties and reactivity, which can serve as markers for individual processes. This challenge is exemplified by the problem of increasing levels of dissolved organic carbon (DOC) in drinking water reservoirs, which poses serious

challenges for the commercial supply of drinking water.[12,13] Here, DOM sources and its quality play a crucial role for DOC removal during drinking water production.[14,15] The discrimination of photochemical and microbial DOM transformation processes in drinking water reservoirs is important as they alter the DOM quality and its removal potential.[16] Thus it is of high importance to better understand relationships between DOM molecular composition, chemical characteristics and biogeochemical reactivity in the reservoir as well as removal[17] and disinfection byproducts formation potential.[18,19]

Ultra-high resolution mass spectrometry (FT-ICR or Orbitrap MS) provides molecular chemical information about DOM molecules, and thus provides a direct link to microbial, photochemical, or geochemical molecular reactivity.[20-23] Experimental approaches with isolated transformation processes provide direction and strength of molecular changes of DOM. However, the transferability to natural environments with coupled transformation processes can be limited and need to be combined with field observations[8,9,24-28]

This study employs a novel approach to evaluate the relative importance of biogeochemical processes by linking the DOM quality change in two different experiments: First, the DOM composition in a reservoir was correlated to a biological

parameter (Chl-a concentration) across seasons and water depth (mixed microbial and photochemical processes). Second, the DOM composition was correlated to radiation dose (only photochemical transformations) as recently published in an irradiation experiment using water from a reservoir tributary[29]. We used a simple, yet robust model based on Spearman`s rank correlation which is an established tool to correlate peak intensities of FT-ICR-MS derived molecular formulas with external parameters with a semi-quantitative validity.[12,15,30-35] Combining two experimentally derived correlations yielded nine potential correlation types for DOM compounds, which were interpreted as molecular reactivity classes. Using a random forest model we predicted the classification of further compounds in the data set to the nine reactivity classes. Primarily this study aimed at establishing a versatile and transferable framework to determine and to predict DOM transformation reactions based on mass spectrometry derived molecular information. The molecular reactivity classes, which we derived from our datasets, can thus be used and validated in future studies.

2 Materials and methods

2.1 Study site and sampling

The Rappbode Reservoir is the largest drinking water reservoir in Germany (catchment size 274 km^2, volume 113 • 10^6 m^3, 400 m a.s.l.).[36,37] Samples were collected on seven dates between January 16 and December 6, 2016 at the deepest location of the lake close to the dam wall (51°44'18" N 10°53'29" E) with a 4.2 L water sampler (Limnos) at nine depths (0–70 m; Table S1) at around 10 am local time. Only 60 out of 63 (7 dates x 9 depths) potential samples were collected (in January the 30 m sample was missing; in October and December the water level was too low for a 70 m sample). Aliquots of 1.5 L were filtered within a few hours after sampling using glass fiber filters (Whatman GF/F). Algal pigments were determined in filtered samples after extraction with 5 mL ethanol and several freezing/thawing cycles with HPLC-UVD-FLD-DAD system (Ultimate 3000, Thermo fisher, USA).[38] DOC was measured via the high-temperature catalytic oxidation method (DIMATOC2000, Dimatec, Germany). Rappbode Reservoir is a dimictic water body. In 2016, summer stratification persisted from April until November while mixing extended to a depth of 10 m from early May until late September (Fig. S2.; see Wentzky et al. for details[37]). During the growing

117 season, measured Secchi depths between 3.5 and 5 m indicated photic depths
118 between 10 and 14 m.[39]
119

120 2.2 FT-ICR-MS data sets

121 The analysis of the filtered samples from the Rappbode Reservoir included solid
122 phase extraction (SPE)[40,41], analysis with ultrahigh resolution Fourier-transform ion
123 cyclotron resonance mass spectrometry[29], molecular formula (MF) assignment[42,43],
124 calculation of relative peak intensities from mono-isotopic data sets[44] , molecular
125 degradation index (I_{DEG})[44] and removal of process blanks (for details see SI2). In total
126 6181 unique MF representing on average > 99% of assigned intensity in all 60 samples
127 were used for further data evaluation ("Lake data set", Fig. S3). 1704 compounds (≈
128 82% of assigned intensity on average) were jointly present in all 60 samples of the
129 Lake data set and constitute the "Chla data set".
130 In addition, we used a data set from a photo degradation experiment ("Photo data
131 set"; comprising 7211 unique MF), originating from a study where stream water
132 samples had been exposed to natural sunlight.[29] Briefly, the samples originated from
133 a forested catchment dominated by terrestrial DOM and only little exposure to solar

irradiation for photochemical alteration. Catchment characteristics and DOM quality were similar to those of the Lake data set and representative for drinking water catchments in Germany.[3] The Photo data set from Wilske et al. (2020) consisted of water samples which were exposed to solar radiation (Rad) in quartz glass bottles and sampled 12 times within six days.[29] 3214 MF (≈ 89% of assigned intensity on average) were jointly present in all 13 Photo data set samples and constitute the “Rad data set”.

For this study, the Lake data set was combined with the Photo data set. This was done by intersecting the 1704 Chla MF with the 3214 Rad MF. This joint data set contained 1277 shared MF (≈ 73% intensity from the Chla data set and ≈ 64% from the Rad data set, Fig. S3.), which were present in 73 samples from both data sets.

Throughout this article, we also refer to a calculated MF from a DOM sample as “compound”, although a DOM MF potentially represents multiple isomers.[16,40]

2.3 Statistical approach and model

For the Spearman`s rank correlation the intensity ranking was calculated for the 1704 MF in the Chla data set and the 3214 MF in the Rad separately. The rank correlation coefficients r_s (Chla) were calculated[12] for each MF using their inter sample ranks

151 (based on 60 samples) and 60 Chl-a concentration ranks. The inter sample ranking
152 calculation is a kind of double ranking (two step rank calculation) and explained in more
153 detail in SI4. Independently, rank correlation coefficients r_s (Rad) were calculated for
154 the Rad data set (inter sample ranks of MF based on 13 samples) with 13 cumulated
155 solar radiation dose ranks.[29] The *p*-values were calculated from the r_s-values as
156 function of the sample number *n* and were adjusted for multiple testing using the
157 Benjamini-Hochberg correction (B-H-cor; 5% false discovery rate).[45,46]
158 We used Chl-a to link DOM processes in the epilimnion of the lake as planktonic
159 algae release a considerable proportion of dissolved photosynthesis products which
160 may reach > 50% of total primary production at low nutrient concentration.[47]
161 For the definition of reactivity classes, the two r_s – values from both correlation types
162 were considered each for the 1277 MF of the joint data set. Due to the different degrees
163 of freedom (58 vs 11), B-H-cor *p*-values and thus significance levels differed for the
164 same r_s-values in both data sets. For each MF in the joint data set, the two correlation
165 features (r_s and B-H-cor *p*-values from Chla and Rad data set, respectively) were used
166 as measure for the averaged reactivity of underlying DOM compounds. We propose
167 that the combination of two independent correlation types from distinct data sets (Chla

and Rad) results in a better assessment of DOM reactivity as compared to individual data sets, provided, that the underlying samples are comparable and the studied processes are sufficiently orthogonal. Table 1 shows the definitions of reactivity classes as function of r_s - and B-H-cor p-values. MF having B-H-cor p-values < 0.05 were considered reactive, MF with B-H-cor p-values > 0.05 were considered non-reactive towards one or both processes (Table 1). The mean molecular chemical properties of the derived reactivity classes (cf. section 3.4) were robust against sample sizes in both data sets (Figure S5), suggesting that sample sizes in our study did not substantially affect the classification using B-H-cor p-values.

A Random forest model (RF) was used to expand the molecular reactivity classes (defined for the 1277 joint MF comprising the training data) to all MF which were at least present once in both data sets (n = 3162, see details in SI9). Random forests are built on decision trees, which do not require linear relationships between variable or multivariate normal distributions.[48] Fitting multiple decision trees to the data combines two statistical concepts, bootstrap aggregating (bagging) and random predictor subset selection reducing the influence of predictor collinearity.[49] We used the RF in regression mode[48] using chemical properties derived from the 1277 joint MF as

predictors (after exclusion of near zero-variance and highly correlated (r^2 > 0.8) predictors) and the correlation coefficients r_s as variables. The model predicted r_s-values, referred to as modeled r_s.

For each correlation type (Chla and Rad, see above) 500 trees were grown down to a minimum node depth of 5 using 11–13 randomly selected predictors at each split. Initial model validation was done with repeated cross-validation (10-fold, 10-times) and the model performance was evaluated with out-of-bag prediction estimates. Further Chla (Rad) model validation included a set of MF, for which r_s values were only available in the Chla (Rad) data set. For more details on model setup, validation and performance refer to SI9. Modeled r_s values for 3162 MF were transformed into reactivity classes as described above using the same degree of freedom (*df*) to calculate the B-H-cor *p*-values as in the respective original data set. The random forest model was implemented in R[45] using the packages randomForest[50] for model algorithm and caret[51] for hyperparameter tuning.

2.4 Suitability of the approach

The interpretation of DOM compositional changes based on molecular formulas is based on an extremely large, yet unknown isomeric diversity of each molecular formula.[52,53] Although biogeochemical processes will affect isomers differently, it can be assumed that the average mass spectral representation of compounds did not change from sample to sample (for consequences see SI5.1). For comparison of samples within a data set we used the ranking of mass peak intensities after harmonization of carbon concentrations for MS measurements (see SI5.2, SI5.3). A change in ranks of the same MF between samples is interpreted as a corresponding increase in its concentration and used in the correlation of inter sample ranks with external parameters like the Chl-a concentration or the cumulated radiation.

Evidently, the experimental design (number of samples, sampling time and depth intervals) and the statistical approach (rank correlation and data set combination) affect the interpretation of results (SI5.4, SI5.5). Moreover, if environmental processes are derived from correlations (like the interpretation of Chl-a correlation), possible superposition of multiple reactions need to be considered (SI5.6). The presented

approach offers one possibility to deal with this via combination of multiple experiments.

3 Results

3.1 Rank correlation of DOM compound intensities with Chl-a concentration

1277 unique MF were shared among all lake monitoring ($n = 60$) and photo degradation samples ($n = 13$), divided into 1002 (78%) CHO, 266 (21%) CHNO and 9 (1%) CHOS compounds (Table S4). The 1277 MF (out of 6181 in the Lake data set) represented on average 72.9% ± 1.7% of the total intensity in each of the 60 samples of the Lake data set. Out of all shared compounds, 342 (27%) correlated positively, 345 (27%) correlated negatively, and 590 (46%) did not correlate significantly with Chl-a concentration (Table S4). If significant, CHNO compounds mostly correlated positively (39% vs 5% negative), while CHOS (although very few) correlated only negatively with Chl-a concentrations. CHO compounds were more balanced between positive (24%) and negative (33%) correlation.

3.2 Rank correlation of DOM compound intensities with cumulated radiation

The 1277 shared MF (out of 7211 in the Photo data set) represent on average 64.2% ± 2.8% of the total intensity in each of the 13 samples of the Photo data set. Among the 1277 shared compounds, 288 were photo produced (positive correlation with cumulative radiation, 22%), 508 were photo degraded (negative correlation, 40%), and 481 compounds were non-reactive (no significant correlation 38%; Table S4). Out of the CHNO compound class, 15% were photo degraded, 36% were photo produced and the remaining 49% were non-reactive. Like in the Chla data set, nearly no positive correlation of CHOS compounds was observed in the Rad data set. Again, the CHO class was dominated by negatively correlating compounds (47% vs 19% positive). All these observed transformations solely result from direct photo degradation.

3.3 Classification of molecular reactivity

The combination of both correlation types, i.e. rank correlation of DOM compounds with Chl-a and with cumulative radiation, resulted in nine distinct reactivity classes (Figure 1, Table 1). Each class is characterized by either positive, negative, or non-significant correlation in each correlation type. E.g. $C_{25}H_{28}O_9$ showed negative correlation with Chl-a concentration ('Chla') but non significant correlation with

249 cumulated irradiation ('Rad'), resulting in a classification as $Chla^{-}Rad^{0}$ (Table
250 S3,Figure 1).
251
252 **Figure 1.**
253
254 The fraction of CHO (CHNO) compounds (out of 1277) which correlated both
255 positively with Chl-a and radiation ($Chla^{+}Rad^{+}$) was 85% (71%) with respect to all
256 positive correlating CHO (CHNO) compounds in the photo degradation experiment and
257 68% (65%) with respect to the Chla data set (Table S4). Similarly, 48% (5%) of the
258 CHO (CHNO) compounds with negative correlation in the photo degradation
259 experiment and 69% (14%) of the CHO (CHNO) compounds having negative Chl-a
260 correlation formed the class $Chla^{-}Rad^{-}$. 44% (40%) of all CHO (CHNO) compounds
261 showed significant correlation only in one dataset (classes $Chla^{+}Rad^{0}$, $Chla^{-}Rad^{0}$,
262 $Chla^{0}Rad^{-}$, $Chla^{0}Rad^{+}$) while less than 0.4% (4.9%) of the CHO (CHNO) compounds
263 showed contrasting correlation behavior (classes $Chla^{+}Rad^{-}$, $Chla^{-}Rad^{+}$). Overall, 17%
264 and 42% of the CHO and CHNO compounds showed mutually no correlation in both

265 data sets (class $Chla^0Rad^0$), while 83% and 64% of the MF could be classified as

266 reactive.

267

268 **Table 1. Classification approach, definition of correlation types based on rank**
269 **correlation in two data sets**

270

correlation with Chl-a, Chla data set		correlation with radiation, Rad data set		correlation type	reaction type
r_s [a]	*p*-value (B-H-cor)[b]	r_s [a]	*p*-value (B-H-cor)[b]		
positive	$p < 0.05$	positive	$p < 0.05$	$Chla^{+}Rad^{+}$	photo product [c]
±	non sign	positive	$p < 0.05$	$Chla^{0}Rad^{+}$	photo product, no net reactivity in lake water
negative	$p < 0.05$	positive	$p < 0.05$	$Chla^{-}Rad^{+}$	microbially degraded photo product
positive	$p < 0.05$	±	non sign	$Chla^{+}Rad^{0}$	microbially derived product
±	non sign	±	non sign	$Chla^{0}Rad^{0}$	non-reactive
negative	$p < 0.05$	±	non sign	$Chla^{-}Rad^{0}$	microbially degraded compound
positive	$p < 0.05$	negative	$p < 0.05$	$Chla^{+}Rad^{-}$	photo degraded microbial product
±	non sign	negative	$p < 0.05$	$Chla^{0}Rad^{-}$	photo degraded compound, no net reactivity in lake water

negative	$p < 0.05$	negative	$p < 0.05$	Chla$^-$Rad$^-$	photo degraded compound

[a] Spearman's rank correlation coefficient

[b] Significance level (Benjamini-Hochberg corrected) derived from r_s and df in each data set

[c] photo product includes photo-enriched and photo-resistant molecules, cf. SI5.3

± Correlation coefficient could be positive or negative

3.4 Chemical properties of reactivity classes

So far, the presented results are evaluations of the number and direction of correlations of DOM compounds with environmental parameters, used to define reactivity classes. In order to link reactivity with molecular properties, we utilized the molecular information of DOM compounds obtained from the FT-ICR-MS data (Figure 2).

Figure 2.

Compounds in the class Chla$^-$Rad$^-$ (photo degraded compounds; Figure 2a, b) were on average less saturated and of higher molecular weight whereas more saturated and

289 lower molecular weight compounds were dominating the class $Chla^{+}Rad^{+}$ (photo
290 products; Figure 2c, d). CHNO compounds generally followed the distribution of the
291 CHO compounds in the van Krevelen space but with lower average masses (Figures
292 S6-7, Tables S6-7). Notably compounds in class $Chla^{-}Rad^{0}$ (microbially degraded
293 compounds, Figure 2a, b) had higher molecular weight than compounds in class
294 $Chla^{+}Rad^{0}$ (microbially derived products, Figure 2c, d; Tables S6-7). The latter class
295 also had a higher average O/C ratio (Tables S6-7) compared to $Chla^{-}Rad^{0}$.
296 Compounds of the class $Chla^{0}Rad^{+}$ and $Chla^{0}Rad^{-}$ exhibited similar chemical
297 properties as classes $Chla^{+}Rad^{+}$ and $Chla^{-}Phot^{-}$, respectively (Figure 2e, f). Only 15
298 MFs were classified into the contrasting reactivity class $Chla^{+}Rad^{-}$ and none into $Chla^{-}$
299 Rad^{+}. Expectedly, the compounds of the non-reactive class $Chla^{0}Rad^{0}$ were evenly
300 distributed in the van Krevelen and H/C vs mass space (Figure **2**e, f).
301

302 3.5 Modeled reactivity classes

303 The random forest model performed with an r^2 (post-resample) of > 0.97 for both
304 correlation types ('Chla' and 'Rad', Table S8, Figure S10). In both cases, most
305 important predictors (assessed by its influence on model accuracy) were the number

306 of N atoms in a MF, with carbon nominal oxidation state (NOSC) values, mass and
307 DBE also being important (Figure S10). Together with the MF's X_C values (as measure
308 of aromaticity) and O/C ratio, these predictors were also most frequently selected as
309 splitting variables at the top of the regression trees (Figure S11-12). The predicted r_s -
310 values bimodally (trimodally) distributed for Rad (Chla) correlation types, similar to the
311 observed r_s -values (Figure 3).
312
313 **Figure 3.**
314
315 The average coordinates in the van Krevelen space of the predicted reactivity
316 classes generally followed the initially classified MF (as based on the calculated r_s -
317 values), with however mostly expanded ranges (Figures S13-14). No MF was assigned
318 to the two classes with contrasting reactivity ($Chla^{-}Rad^{+}$ and $Chla^{+}Rad^{-}$). The modelled
319 reactivity classes covered a larger fraction of N-and S-containing MF (36% and 13%,
320 Table S8) as compared to the observations used in the model (21% and < 1%, Table
321 S4). Model predictors did account for the elemental composition of individual MF by
322 including the number of N and S, resulting in assignments of reactivity classes also for

CHNOS (in addition to CHNO and CHOS) MF which were not part of the joint data set. However, the functional groups of these MFs were not part of the model, which may result in unaccounted reactivity for heteroatom-containing MF. The fraction of MF classified as non-reactive (55%) was almost three times larger than for the 1277 shared MFs in the joint data set.

Figure 4.

4 Discussion

4.1 Improved links between chemical properties and biogeochemical reactivity

During lake stratification in spring and summer months algal blooms establish in the epilimnion of the Rappbode Reservoir[54] (Figure S2, Chl-a profiles), increasing the DOC concentration and altering DOM composition via release and transformation of photoautotrophically produced compounds. However, also photochemical processes may have altered the DOM composition and therefore photochemically produced molecules and microbial derived molecules cannot easily be distinguished via environmental time series alone. E.g. using the ranges of the molecular degradation

index I_{DEG}[44], photo degradation may account for most of the peak intensity variability in surface samples of the Chla data set (Figure S17). Our approach of using data from an independent but comparable photo degradation experiment allowed us to better differentiate biological and photochemical origin and potential fate of DOM compounds. Further linking the observed biogeochemical reactivity (via classification based on experimental data sets) with molecular chemical properties (via machine learning models) offers more insights into the chemical driving factors of DOM reactivity in the environment.

Photo products ($Chla^{+}Rad^{+}$) are MF which showed positive correlation in both data sets (cf. SI5.3). The correlation to radiation in the photo degradation experiment clearly describes the chemical reactivity of compounds. In contrast the correlation to Chl-a is ambiguous, although Chl-a concentration in surface waters is often well correlated to solar radiation. However, it cannot be excluded that some of the photo products (or a subset of the underlying isomers) can also be microbially produced or directly released as algal exudates. Samples from the lake monitoring comprise different seasons and depth. Positive correlations of the photo products with Chl-a suggest that these compounds accumulate in the surface layer or are selectively enriched (the

358 concentrations did not necessarily increase, see caveats in SI5.3). The epilimnion in
359 Rappbode Reservoir was rather shallow due to the wind-protected position of the
360 reservoir (see depth profiles in Fig. S2). However, when mixing of the water column
361 starts in autumn, relative intensities of photo products decrease towards values of the
362 hypolimnion (as shown for $C_9H_{12}O_5$ in Figure S18).

363 The compounds from $Chla^+Rad^+$ were relatively saturated (H/C > 1, AI_{mod} < 0.34),
364 had low (negative) DBE-O values (Tables S6-7, FigureS8) and showed an enrichment
365 in the epilimnion of the Rappbode Reservoir during summer months and a monotonous
366 normalized intensity increase during irradiation (Figure S18 a-d). Experimental
367 evidence for the production of aliphatic photo products in freshwater suggests this as
368 being a common process throughout environments.[9,28,55,56] In a photo degradation
369 experiment with water from ten world rivers photo produced MF in the range H/C > 1.4
370 and O/C > 0.5 are in agreement with the photo products in our study.[9] For the range
371 1.0 < H/C < 1.4 and O/C < 0.5 we found both photo products and photo degraded
372 compounds whereas Riedel et al. found mainly photo products.[9] Similarly, aliphatic
373 compounds, which were enriched in stratified marine surface waters, were suggested

to be produced via photochemical transformation, suggests this as being a common process throughout environments.[26]

Aliphatic and potentially microbial derived CHNO compounds were also part of the photo product class, consistent with observations from rivers and marine systems.[25,57,58]

Photo degraded compounds ($Chla^{-}Rad^{-}$) were on average more unsaturated (lower H/C ratio, higher AI_{mod} values), less oxygenated (lower O/C ratios, higher DBE-O values; Tables S6-7, Figure S8) and of higher molecular mass than photo products, similar to other studies.[24,26,56,59] Photo degraded compounds in the range H/C < 1 and O/C > 0.5 were previously found in other studies.[9,11] Compounds with similar chemical characteristics (higher aromaticity, low nitrogen content) are typically found in freshwater systems and are mainly derived from terrestrial sources.[9] Photochemical reactivity of aromatic compounds like $C_{19}H_{22}O_7$ (strong relative depletion in the surface of the Rappbode Reservoir during summer month, Figure S18 e-h) is attributed to the high degree of aromatic moieties and is also suggested due to a pronounced loss in specific UV absorption.[56] However, the majority of MF in the $Chla^{-}Rad^{-}$ class were found in a range 0.2 < O/C < 0.4 and 1 < H/C < 1.6. Gonsior et al. also found in a

boreal lake DOM photo degradation of compounds with comparable molecular elemental ratios.[11] In comparison to the $Chla^{+}Rad^{+}$ class the $Chla^{-}Rad^{-}$ class comprises very few CHNO compounds (Table S4, Figure 1). Hence, photo degradation of CHNO seems to play a minor role.

Concerning the differences in average oxygenation, addition of reactive oxygen species to double bonds, in contrast to photo decarboxylation, may explain the relative higher average O/C and lower DBE-O values for photo products as compared to the photo degraded compounds.[60,61]

Non-reactive ($Chla^{0}Rad^{0}$) compounds showed conservative behavior within the observation periods (days to month) and boundary conditions (oxic surface water). As reported earlier not all compounds show biogeochemical reactivity. Bittar et al. found many compounds which originated from axenic algae cultures to be both photo- and bioresistant.[8] However, these compounds may also represent intermediate products in complex biogeochemical reaction cascades and their reactivity attribution may depend on the investigated time range. Also, simultaneous degradation both in epilimnion and hypolimnion or a continuously supply via the tributaries and mixing would lead no observable correlation with biogeochemical proxies like Chl-a. Some algal exudates

408 are more photolabile and less biolabile[8] possibly resulting in no detectable correlation.
409 In any case this class shows no net reactivity of compounds and the mean molecular
410 properties of this class are close to the overall average of all samples (Tables S6-7).

411 **Microbially derived products** ($Chla^{+}Rad^{0}$) are postulated if no correlation to radiation
412 but a significant positive correlation with Chl-a was observed. They were also evidently
413 enriched in the epilimnion of the Rappbode Reservoir during summer months (as
414 shown for $C_{11}H_{11}N_1O_6$ in Figure S19 a-d). This class potentially represents the
415 autochtonous DOM in the lake: the molecules are on average small (290 Da) and
416 oxygen rich (Tables S6-7). A large number of potentially microbial derived products will
417 not be detectable by this approach due to low extraction and ionization efficiency (like
418 polysaccharides) and turnover rates in the range of minutes to days.[14] Also, using a
419 correlation to Chl-a, differentiation between autotrophic and heterotrophic processes
420 is not possible. Compounds identified via Chl-a correlation can be algal exudates or
421 molecules released after processing by bacteria and both processes are known for
422 introducing rather small and polar compounds into the DOM pool.[59,62] Although these
423 compounds share chemical characteristics with the photo product class, we assume
424 high production/release of the compounds in the surface layer due to biological activity.

425 A part of this group shows similar van Krevelen coordinates (0.6 < H/C < 1.3; 0.5 <
426 O/C) with compounds which were highly negatively correlated to the DOC
427 concentration (which was decreasing) in a biodegradation experiment with samples
428 from ten world rivers.[9] In a degradation experiment with leaf leachates MF with 0.5 <
429 H/C < 1.5 and < 450 Da were microbially produced.[62] Bacterial metabolites with
430 comparable molecular properties were stable on time scales of weeks-month.[22]
431 Compounds of the $Chla^{+}Rad^{0}$ class share molecular characteristics with the island of
432 stability (0.9 < H/C < 1.4; 0.4 < O/C < 0.7) suggesting long term stability as found for
433 marine DOM.[40,43,63]

434 **Microbially degraded compounds** ($Chla^{-}Rad^{0}$) were suggested when depletion in the
435 epilimnion of the Rappbode Reservoir (as shown for $C_{25}H_{28}O_9$ in Figure S19 e-f),
436 negative correlation with Chl-a, and no alteration in the photo degradation experiment
437 were found.

438 Negative correlation with Chl-a may also result from an enrichment of a compound
439 in the hypolimnion versus the epilimnion, resulting in an interpretation as (hypolimnetic)
440 microbial product. The decomposition of particulate organic matter after settling to the
441 hypolimnion or sediment might produce such compounds. However it was shown that

positive flux of DOC from the sediment to the hypolimnion in a predam of Rappbode Reservoir occurred only at dissolved oxygen (DO) concentrations < 6 mg L^{-1}.[64] DOC release under oxic conditions was only found at T > 10°C. During our study, the temperature was always below 10 °C at the lake bottom and DO was below the threshold of 6 mg L^{-1} only in October. We conclude that DOM release to the hypolimnion in the Rappbode main dam Reservoir was negligible.

CHO compounds in the Chla-Rad^0 class were on average larger (452 Da) than CHNO (382 Da), but both were larger than molecules in the microbial derived product class. Both results suggest preferential release of smaller and degradation of larger molecular weight compounds by microbial heterotrophic processes.[22,62] Compounds in this class which are degraded by heterotrophic processes can originate from autochthonous and/or allochthonous sources. Bittar et al.[8] found biolabile and at the same time photo-resistant compounds using an axenic algae culture (autochthonous source) in the mass range 400–600 Da and with 0.7 < H/C < 1.3 which is in agreement with compounds in the (Chla-Rad^0) class. In a degradation experiment with terrestrial source (leaf leachate), compounds in the same range were found to be degraded.[62] Riedel et al.[9] found compounds in the range 0.9 < H/C < 1.5 and O/C < 0.5 positively

459 related to DOC (which decreased) suggesting bio-degradation in agreement to the
460 ($Chla^{-}Rad^{0}$) group. Lusk et al.[65] reported that biodegraded CHNO are less oxygenated
461 than bio-produced CHNO which is in agreement to our results (Table S7). Alternatively,
462 the negative correlation of non-photo reactive compounds to Chl-a can be explained
463 by enrichment of allochthonous versus autochthonous compounds in winter.

464 The remaining classes **$Chla^{-}Rad^{+}$, $Chla^{+}Rad^{-}$**, **$Chla^{0}Rad^{+}$, $Chla^{0}Rad^{-}$** are discussed
465 in SI11. The first three comprised less than 6% of all classified MF, the latter one 21%.
466

467 4.2 Relevance for drinking water production

468 Depending on the applied treatment process DOM compounds of different
469 biogeochemical origin (and different molecular characteristics) can be removed (i.e. by
470 coagulation, oxidation or other processes) during drinking water production.[17] Mainly
471 oxygen-rich and relatively unsaturated compounds (O/C > 0.5; H/C < 1.5; m/z > 400
472 Da) can be readily removed via conventional full scale process (coagulation with
473 $Al_2(SO_4)_3$, sedimentation, rapid sand filtration, slow sand filtration, UV disinfection,
474 dosing of NH_2Cl).[14,17] Such compounds are depleted in epilimnetic waters mainly by
475 photochemical degradation (class $Chla^{-}Rad^{-}$, Figure 2). In contrast, photo products

476 (class $Chla^{+}Rad^{+}$) produced in the epilimnion can be potential precursors for
477 disinfection byproducts.[18, 19] Reported potential precursor compounds match with MF
478 from our study and are listed in Table S9, many of them classified here as photo
479 products. Some of the putative precursors like the photo product $C_9H_{12}O_6$ had two
480 times higher intensity in the epilimnion compared to the hypolimnion in the Rappbode
481 Reservoir (SI database.xlsx). We suggest that a tracing of photo products as potential
482 DBP precursors need be taken into account for drinking water processing in the future.
483

484 ## 4.3 Final statements

485 A classification approach to discriminate biogeochemical reaction types founded on
486 independent experimental information and data sets was developed. Our approach
487 does not rely on an a priori definition of biogeochemical reactivity classes based or
488 questionable assumptions on macromolecular precursors.

489 Nevertheless, it appears to manifest that based on molecular formulas, DOM
490 compounds can have comparable chemical and biological properties (photo reactivity,
491 microbial lability, etc.) in comparable biogeochemical contexts[8,9,40,43,66] and the results
492 of this study clearly support the reactivity continuum models of DOM. [40,44,63,67]

493 However, if two or even more reactivity types are taken into consideration,
494 multidimensional reactivity continua largely overlap and molecular reactivity classes
495 cannot be easily projected in a two dimensional (e.g. van Krevelen) molecular space
496 any more. Thus, for the interpretation of DOM changes in biogeochemical and
497 ecological studies, a classification of molecular formulas into (environment-dependent)
498 reactivity types may prove much more valuable and of practical usefulness than
499 putative structural assignments in a low-dimensional chemical space (like H/C versus
500 O/C or H/C versus mass).

501 Here, we have demonstrated the potential to extend DOM biogeochemical properties
502 and reactivity information based on molecular information via predictive models to less
503 frequently observed compounds, further taking advantage of the reactivity continua
504 within DOM.

505 Expanding our approach to an array of experiments/monitoring data sets from alike
506 and contrasting environments and interlinked via databases and machine learning
507 methods will provide more accurate and more detailed biogeochemical reactivity
508 classes assigned to individual MF as ever before.

509

5 ASSOCIATED CONTENT

The Supporting Information is available free of charge at https://pubs.acs.org/doi/

Study site and water chemistry details; quality control of FT-ICR-MS; examples for rank correlation; data set combination description; additional results from rank correlation in tables and diagrams; random forest model results in tables and diagrams; compounds intensity versus depth or cumulated radiation diagrams; additional references (pdf)

A data base with all molecular formulas used in this study (database.xlsx)

6 AUTHOR INFORMATION

Corresponding Authors

*E-mail: peter.herzsprung@ufz.de, Phone: +49 391 810 9330

*E-mail: oliver.lechtenfeld@ufz.de, Phone: +49 341 235 1020

Present Addresses

State Agency for Agriculture, Environment and Rural Areas Schleswig-Holstein - LLUR, Lake Department, Hamburger Chaussee 25, 24220 Flintbek, Germany Author Contributions

525 The manuscript was written through contributions of all authors. All authors have
526 given approval to the final version of the manuscript.

527 Notes

528 The authors declare no competing financial interest.

529 7 ACKNOWLEDGMENTS

530 Jan M. Kaesler is gratefully acknowledged for the FT-ICR-MS analysis at the Centre
531 for Chemical Microscopy (ProVIS) at the Helmholtz Centre for Environmental
532 Research which is supported by European Regional Development Funds (EFRE-
533 Europe funds Saxony) and the Helmholtz Association. We thank Kai Franze for
534 software development, Mick Wu for support with the random forest model, Ina Siebert
535 for DOC determination, Ines Locker for SPE and Kerstin Lerche for Chl-a analysis. We further
536 thank Karsten Rahn, Martin Wieprecht and Michael Herzog for collecting water samples at
537 Rappbode Reservoir. The comments of three anonymous reviewers improved the manuscript
538 considerably.

539

8 REFERENCES

1. McKnight, D.; Thurman, E. M.; Wershaw, R. L.; Hemond, H. Biogeochemistry of aquatic humic substances in Thoreau bog, Concord, Massachusetts. *Ecology* **1985,** *66* (4), 1339-1352.
2. Morling, K.; Raeke, J.; Kamjunke, N.; Reemtsma, T.; Tittel, J. Tracing aquatic priming effect during microbial decomposition of terrestrial dissolved organic carbon in chemostat experiments. *Microbial Ecology* **2017,** *74* (3), 534-549.
3. Dadi, T.; Harir, M.; Hertkorn, N.; Koschorreck, M.; Schmitt-Kopplin, P.; Herzsprung, P. Redox conditions affect dissolved organic carbon quality in stratified freshwaters. *Environmental Science & Technology* **2017,** *51* (23), 13705-13713.
4. Osterholz, H.; Kirchman, D. L.; Niggemann, J.; Dittmar, T. Environmental drivers of dissolved organic matter molecular composition in the Delaware Estuary. *Frontiers in Earth Science* **2016,** *4*, 95
5. Cottrell, B. A.; Gonsior, M.; Timko, S. A.; Simpson, A. J.; Cooper, W. J.; van der Veer, W. Photochemistry of marine and fresh waters: A role for copper-dissolved organic matter ligands. *Mar. Chem.* **2014,** *162*, 77-88.
6. Maizel, A. C.; Li, J.; Remucal, C. K. Relationships between dissolved organic matter composition and photochemistry in lakes of diverse trophic status. *Environmental Science & Technology* **2017,** *51* (17), 9624-9632.
7. Berg, S. M.; Whiting, Q. T.; Herrli, J. A.; Winkels, R.; Wammer, K. H.; Remucal, C. K. The role of dissolved organic matter composition in determining photochemical reactivity at the molecular level. *Environmental Science & Technology* **2019,** *53* (20), 11725-11734.
8. Bittar, T. B.; Vieira, A. A. H.; Stubbins, A.; Mopper, K. Competition between photochemical and biological degradation of dissolved organic matter from the cyanobacteria Microcystis aeruginosa. *Limnology and Oceanography* **2015,** *60* (4), 1172-1194.
9. Riedel, T.; Zark, M.; Vahatalo, A. V.; Niggemann, J.; Spencer, R. G. M.; Hernes, P. J.; Dittmar, T. Molecular signatures of biogeochemical transformations in dissolved organic matter from ten world rivers. *Frontiers in Earth Science* **2016,** *4*, 85.

570 10. Kellerman, A. M.; Kothawala, D. N.; Dittmar, T.; Tranvik, L. J. Persistence of
571 dissolved organic matter in lakes related to its molecular characteristics. *Nature*
572 *Geoscience* **2015,** *8* (6), 454-U52.
573 11. Gonsior, M.; Schmitt-Kopplin, P.; Bastviken, D. Depth-dependent molecular
574 composition and photo-reactivity of dissolved organic matter in a boreal lake under
575 winter and summer conditions. *Biogeosciences* **2013,** *10* (11), 6945-6956.
576 12. Herzsprung, P.; von Tumpling, W.; Hertkorn, N.; Harir, M.; Buttner, O.;
577 Bravidor, J.; Friese, K.; Schmitt-Kopplin, P. Variations of dom quality in inflows of a
578 drinking water reservoir: Linking of van Krevelen diagrams with EEMF spectra by
579 rank correlation. *Environmental Science & Technology* **2012,** *46* (10), 5511-5518.
580 13. Evans, C. D.; Monteith, D. T.; Cooper, D. M. Long-term increases in surface
581 water dissolved organic carbon: Observations, possible causes and environmental
582 impacts. *Environ. Pollut.* **2005,** *137* (1), 55-71.
583 14. Raeke, J.; Lechtenfeld, O. J.; Tittel, J.; Oosterwoud, M. R.; Bornmann, K.;
584 Reemtsma, T. Linking the mobilization of dissolved organic matter in catchments and
585 its removal in drinking water treatment to its molecular characteristics. *Water Res.*
586 **2017,** *113*, 149-159.
587 15. Lavonen, E. E.; Kothawala, D. N.; Tranvik, L. J.; Gonsior, M.; Schmitt-
588 Kopplin, P.; Kohler, S. J. Tracking changes in the optical properties and molecular
589 composition of dissolved organic matter during drinking water production. *Water Res.*
590 **2015,** *85*, 286-294.
591 16. Lavonen, E. E.; Gonsior, M.; Tranvik, L. J.; Schmitt-Kopplin, P.; Kohler, S. J.
592 Selective chlorination of natural organic matter: identification of previously unknown
593 disinfection byproducts. *Environmental Science & Technology* **2013,** *47* (5), 2264-
594 2271.
595 17. Andersson, A.; Lavonen, E.; Harir, M.; Gonsior, M.; Hertkorn, N.; Schmitt-
596 Kopplin, P.; Kylin, H.; Bastviken, D. Selective removal of natural organic matter
597 during drinking water production changes the composition of disinfection by-products.
598 *Environmental Science-Water Research & Technology* **2020,** *6* (3), 779-794.
599 18. Phungsai, P.; Kurisu, F.; Kasuga, I.; Furumai, H. Molecular characterization
600 of low molecular weight dissolved organic matter in water reclamation processes
601 using Orbitrap mass spectrometry. *Water Res.* **2016,** *100*, 526-536.

19. Phungsai, P.; Kurisu, F.; Kasuga, I.; Furumai, H. Changes in dissolved organic matter composition and disinfection byproduct precursors in advanced drinking water treatment processes. *Environmental Science & Technology* **2018,** *52* (6), 3392-3401.

20. Wunsch, U. J.; Acar, E.; Koch, B. P.; Murphy, K. R.; Schmitt-Kopplin, P.; Stedmon, C. A. The molecular fingerprint of fluorescent natural organic matter offers insight into biogeochemical sources and diagenetic state. *Anal. Chem.* **2018,** *90* (24), 14188-14197.

21. Raeke, J.; Lechtenfeld, O. J.; Seiwert, B.; Meier, T.; Riemenschneider, C.; Reemtsma, T. Photochemically induced bound residue formation of carbamazepine with dissolved organic matter. *Environmental Science & Technology* **2017,** *51* (10), 5523-5530.

22. Lechtenfeld, O. J.; Hertkorn, N.; Shen, Y.; Witt, M.; Benner, R. Marine sequestration of carbon in bacterial metabolites. *Nature Communications* **2015,** *6*, 6711.

23. Kamjunke, N.; Nimptsch, J.; Harir, M.; Herzsprung, P.; Schmitt-Kopplin, P.; Neu, T. R.; Graeber, D.; Osorio, S.; Valenzuela, J.; Reyes, J. C.; Woelfl, S.; Hertkorn, N. Land-based salmon aquacultures change the quality and bacterial degradation of riverine dissolved organic matter. *Scientific Reports* **2017,** *7*, 43739

24. Rossel, P. E.; Vahatalo, A. V.; Witt, M.; Dittmar, T. Molecular composition of dissolved organic matter from a wetland plant (Juncus effusus) after photochemical and microbial decomposition (1.25 yr): Common features with deep sea dissolved organic matter. *Org. Geochem.* **2013,** *60*, 62-71.

25. Mesfioui, R.; Abdulla, H. A. N.; Hatcher, P. G. Photochemical alterations of natural and anthropogenic dissolved organic nitrogen in the York River. *Environmental Science & Technology* **2015,** *49* (1), 159-167.

26. Medeiros, P. M.; Seidel, M.; Powers, L. C.; Dittmar, T.; Hansell, D. A.; Miller, W. L. Dissolved organic matter composition and photochemical transformations in the northern North Pacific Ocean. *Geophysical Research Letters* **2015,** *42* (3), 863-870.

27. Mostovaya, A.; Hawkes, J. A.; Dittmar, T.; Tranvik, L. J. Molecular determinants of dissolved organic matter reactivity in lake water. *Frontiers in Earth Science* **2017,** *5*, 106.

634 28. Stubbins, A.; Spencer, R. G. M.; Chen, H. M.; Hatcher, P. G.; Mopper, K.;
635 Hernes, P. J.; Mwamba, V. L.; Mangangu, A. M.; Wabakanghanzi, J. N.; Six, J.
636 Illuminated darkness: Molecular signatures of Congo River dissolved organic matter
637 and its photochemical alteration as revealed by ultrahigh precision mass
638 spectrometry. *Limnology and Oceanography* **2010,** *55* (4), 1467-1477.
639 29. Wilske, C.; Herzsprung, P.; Lechtenfeld, O. J.; Kamjunke, N.; von Tumpling,
640 W. Photochemically induced changes of dissolved organic matter in a humic-rich and
641 forested stream. *Water* **2020,** *12* (2), 331.
642 30. Stubbins, A.; Lapierre, J. F.; Berggren, M.; Prairie, Y. T.; Dittmar, T.; del
643 Giorgio, P. A. What's in an EEM? Molecular signatures associated with dissolved
644 organic fluorescence in Boreal Canada. *Environmental Science & Technology* **2014,**
645 *48* (18), 10598-10606.
646 31. Kellerman, A. M.; Dittmar, T.; Kothawala, D. N.; Tranvik, L. J. Chemodiversity
647 of dissolved organic matter in lakes driven by climate and hydrology. *Nature*
648 *Communications* **2014,** *5*, 3804.
649 32. Ohno, T.; Parr, T. B.; Gruselle, M. C. I.; Fernandez, I. J.; Sleighter, R. L.;
650 Hatcher, P. G. Molecular composition and biodegradability of soil organic matter: A
651 case study comparing two New England forest types. *Environmental Science &*
652 *Technology* **2014,** *48* (13), 7229-7236.
653 33. Martinez-Perez, A. M.; Nieto-Cid, M.; Osterholz, H.; Catala, T. S.; Reche, I.;
654 Dittmar, T.; Alvarez-Salgado, X. A. Linking optical and molecular signatures of
655 dissolved organic matter in the Mediterranean Sea. *Scientific Reports* **2017,** *7*.
656 34. Roullier-Gall, C.; Hemmler, D.; Gonsior, M.; Li, Y.; Nikolantonaki, M.; Aron,
657 A.; Coelho, C.; Gougeon, R. D.; Schmitt-Kopplin, P. Sulfites and the wine
658 metabolome. *Food Chem.* **2017,** *237*, 106-113.
659 35. Kamjunke, N.; Hertkorn, N.; Harir, M.; Schmitt-Kopplin, P.; Griebler, C.;
660 Brauns, M.; von Tumpling, W.; Weitere, M.; Herzsprung, P. Molecular change of
661 dissolved organic matter and patterns of bacterial activity in a stream along a land-
662 use gradient. *Water Res.* **2019,** *164*., 114919
663 36. Rinke, K.; Kuehn, B.; Bocaniov, S.; Wendt-Potthoff, K.; Buttner, O.; Tittel,
664 J.; Schultze, M.; Herzsprung, P.; Ronicke, H.; Rink, K.; Rinke, K.; Dietze, M.;
665 Matthes, M.; Paul, L.; Friese, K. Reservoirs as sentinels of catchments: the

Rappbode Reservoir Observatory (Harz Mountains, Germany). *Environmental Earth Sciences* **2013,** *69* (2), 523-536.

37. Wentzky, V. C.; Tittel, J.; Jager, C. G.; Rinke, K. Mechanisms preventing a decrease in phytoplankton biomass after phosphorus reductions in a German drinking water reservoirresults from more than 50years of observation. *Freshwater Biology* **2018,** *63* (9), 1063-1076.

38. van Pinxteren, M.; Barthel, S.; Fomba, K. W.; Muller, K.; von Tumpling, W.; Herrmann, H. The influence of environmental drivers on the enrichment of organic carbon in the sea surface microlayer and in submicron aerosol particles - measurements from the Atlantic Ocean. *Elementa-Science of the Anthropocene* **2017,** *5*, 35.

39. French, R. H., Cooper, J. J., Vigg, S. Secchi Disc Relationships1 *JAWRA Journal of the American Water Resources Association* **1982,** *18* (1), 121-123.

40. Lechtenfeld, O. J.; Kattner, G.; Flerus, R.; McCallister, S. L.; Schmitt-Kopplin, P.; Koch, B. P. Molecular transformation and degradation of refractory dissolved organic matter in the Atlantic and Southern Ocean. *Geochim. Cosmochim. Acta* **2014,** *126*, 321-337.

41. Raeke, J.; Lechtenfeld, O. J.; Wagner, M.; Herzsprung, P.; Reemtsma, T. Selectivity of solid phase extraction of freshwater dissolved organic matter and its effect on ultrahigh resolution mass spectra. *Environmental Science-Processes & Impacts* **2016,** *18* (7), 918-927.

42. Herzsprung, P.; Hertkorn, N.; von Tumpling, W.; Harir, M.; Friese, K.; Schmitt-Kopplin, P. Understanding molecular formula assignment of Fourier transform ion cyclotron resonance mass spectrometry data of natural organic matter from a chemical point of view. *Analytical and Bioanalytical Chemistry* **2014,** *406* (30), 7977-7987.

43. Koch, B. P.; Kattner, G.; Witt, M.; Passow, U. Molecular insights into the microbial formation of marine dissolved organic matter: recalcitrant or labile? *Biogeosciences* **2014,** *11* (15), 4173-4190.

44. Flerus, R.; Lechtenfeld, O. J.; Koch, B. P.; McCallister, S. L.; Schmitt-Kopplin, P.; Benner, R.; Kaiser, K.; Kattner, G. A molecular perspective on the ageing of marine dissolved organic matter. *Biogeosciences* **2012,** *9* (6), 1935-1955.

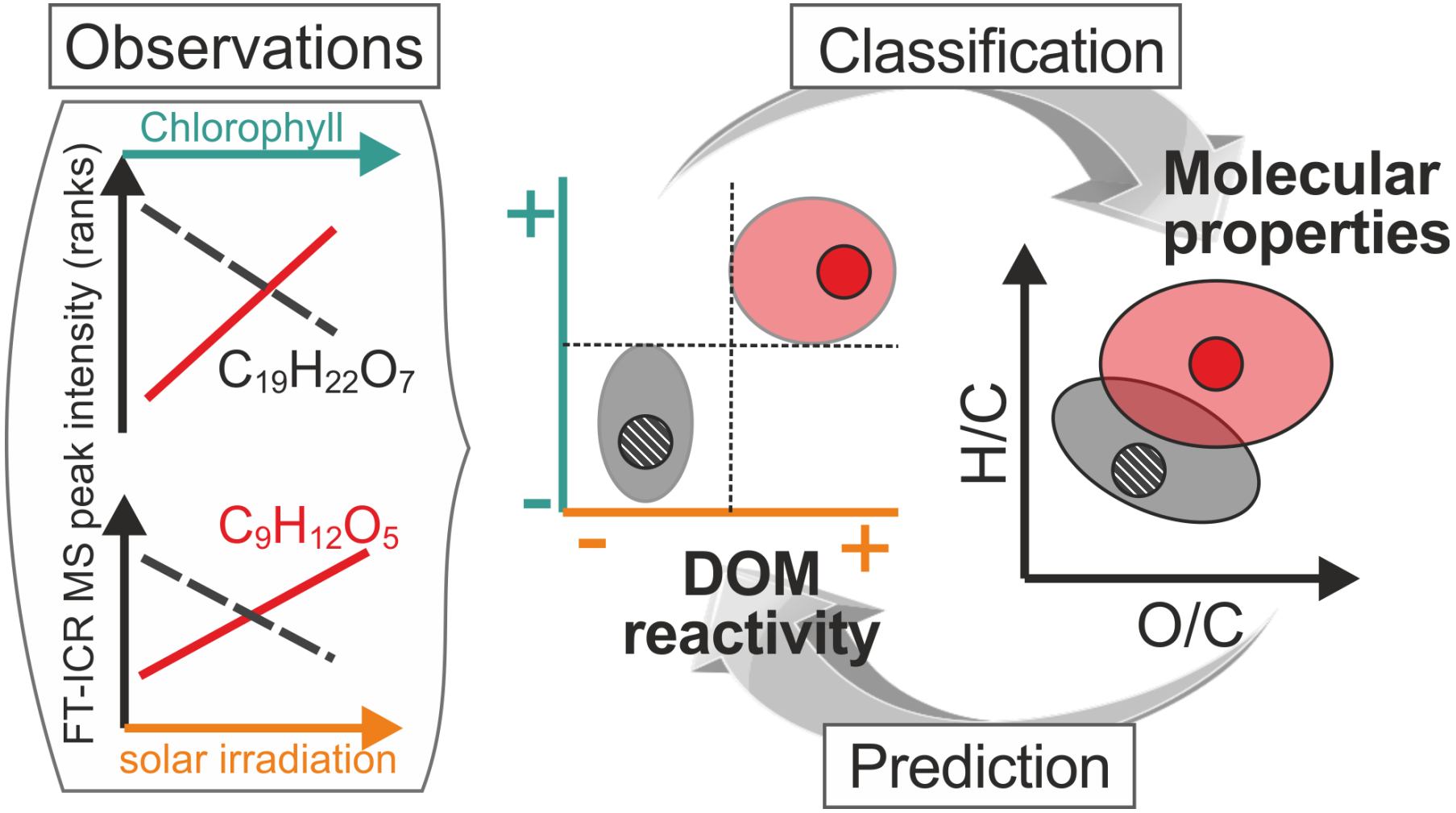
45. *R: A Language and Environment for Statistical Computing*, 3.6.1; R Foundation for Statistical Computing: Vienna, Austria, **2019**.
46. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **1995**, *57* (1), 289-300.
47. Baines, S. B.; Pace, M. L. The production of dissolved organic-matter by phytoplankton and its importance to bacteria - patterns across marine and fresh-water systems. *Limnology and Oceanography* **1991,** *36* (6), 1078-1090.
48. Breiman, L. Random forests. *Machine Learning* **2001,** *45* (1), 5-32.
49. Tomaschek, F.; Hendrix, P.; Baayen, R. H. Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics* **2018,** *71*, 249-267.
50. Liaw, A.; Wiener, M. *Classification and Regression by randomForest*, 4.6-14; **2002**.
51. Kuhn, M. *caret: Classification and Regression Training*, 6.0-85; **2020**.
52. Hertkorn, N.; Ruecker, C.; Meringer, M.; Gugisch, R.; Frommberger, M.; Perdue, E. M.; Witt, M.; Schmitt-Kopplin, P. High-precision frequency measurements: indispensable tools at the core of the molecular-level analysis of complex systems. *Analytical and Bioanalytical Chemistry* **2007,** *389* (5), 1311-1327.
53. Hawkes, E. A.; Patrlarca, A.; Sjoberg, P. J. R.; Tranvik, L. J.; Bergquist, J. Extreme isomeric complexity of dissolved organic matter found across aquatic environments. *Limnology and Oceanography Letters* **2018,** *3* (2), 21-30.
54. Wentzky, V. C.; Frassl, M. A.; Rinke, K.; Boehrer, B. Metalimnetic oxygen minimum and the presence of Planktothrix rubescens in a low-nutrient drinking water reservoir. *Water Res.* **2019,** *148*, 208-218.
55. Seidel, M.; Yager, P. L.; Ward, N. D.; Carpenter, E. J.; Gomes, H. R.; Krusche, A. V.; Richey, J. E.; Dittmar, T.; Medeiros, P. M. Molecular-level changes of dissolved organic matter along the Amazon River-to-ocean continuum. *Mar. Chem.* **2015,** *177*, 218-231.
56. Gonsior, M.; Peake, B. M.; Cooper, W. T.; Podgorski, D.; D'Andrilli, J.; Cooper, W. J. Photochemically induced changes in dissolved organic matter identified by ultrahigh resolution Fourier transform ion cyclotron resonance mass spectrometry. *Environmental Science & Technology* **2009,** *43* (3), 698-703.

730 57. Timko, S. A.; Maydanov, A.; Pittelli, S. L.; Conte, M. H.; Cooper, W. J.;
731 Koch, B. P.; Schmitt-Kopplin, P.; Gonsior, M. Depth-dependent photodegradation of
732 marine dissolved organic matter. *Frontiers in Marine Science* **2015,** *2*, 66.
733 58. Bittar, T. B.; Stubbins, A.; Vieira, A. A. H.; Mopper, K. Characterization and
734 photodegradation of dissolved organic matter (DOM) from a tropical lake and its
735 dominant primary producer, the cyanobacteria Microcystis aeruginosa. *Mar. Chem.*
736 **2015,** *177*, 205-217.
737 59. Kujawinski, E. B.; Del Vecchio, R.; Blough, N. V.; Klein, G. C.; Marshall, A.
738 G. Probing molecular-level transformations of dissolved organic matter: insights on
739 photochemical degradation and protozoan modification of DOM from electrospray
740 ionization Fourier transform ion cyclotron resonance mass spectrometry. *Mar. Chem.*
741 **2004,** *92* (1-4), 23-37.
742 60. Waggoner, D. C.; Wozniak, A. S.; Cory, R. M.; Hatcher, P. G. The role of
743 reactive oxygen species in the degradation of lignin derived dissolved organic matter.
744 *Geochim. Cosmochim. Acta* **2017,** *208*, 171-184.
745 61. Ward, C. P.; Cory, R. M. Complete and partial photo-oxidation of dissolved
746 organic matter draining permafrost soils. *Environmental Science & Technology* **2016,**
747 *50* (7), 3545-3553.
748 62. Kamjunke, N.; von Tumpling, W.; Hertkorn, N.; Harir, M.; Schmitt-Kopplin,
749 P.; Norf, H.; Weitere, M.; Herzsprung, P. A new approach for evaluating
750 transformations of dissolved organic matter (DOM) via high-resolution mass
751 spectrometry and relating it to bacterial activity. *Water Res.* **2017,** *123*, 513-523.
752 63. Kellerman, A. M.; Guillemette, F.; Podgorski, D. C.; Aiken, G. R.; Butler, K.
753 D.; Spencer, R. G. M. Unifying concepts linking dissolved organic matter composition
754 to persistence in aquatic ecosystems. *Environmental Science & Technology* **2018,** *52*
755 (5), 2538-2548.
756 64. Dadi, T.; Friese, K.; Wendt-Potthoff, K.; Koschorreck, M. Benthic dissolved
757 organic carbon fluxes in a drinking water reservoir. *Limnology and Oceanography*
758 **2016,** *61* (2), 445-459.
759 65. Lusk, M. G.; Toor, G. S. Dissolved organic nitrogen in urban streams:
760 Biodegradability and molecular composition studies. *Water Res.* **2016,** *96*, 225-235.

761 66. Pisani, O.; Boyer, J. N.; Podgorski, D. C.; Thomas, C. R.; Coley, T.; Jaffe,
762 R. Molecular composition and bioavailability of dissolved organic nitrogen in a lake
763 flow-influenced river in south Florida, USA. *Aquatic Sciences* **2017,** *79* (4), 891-908.
764 67. Mostovaya, A.; Hawkes, J. A.; Koehler, B.; Dittmar, T.; Tranvik, L. J.
765 Emergence of the reactivity continuum of organic matter from kinetics of a multitude
766 of individual molecular constituents. *Environmental Science & Technology* **2017,** *51*
767 (20), 11571-11579.

768

769


Observations
Chlorophyll
FT-ICR MS peak intensity (ranks)
$C_{19}H_{22}O_7$
$C_9H_{12}O_5$
solar irradiation
Classification
Molecular properties
H/C
O/C
+
-
DOM reactivity
+
-
Prediction

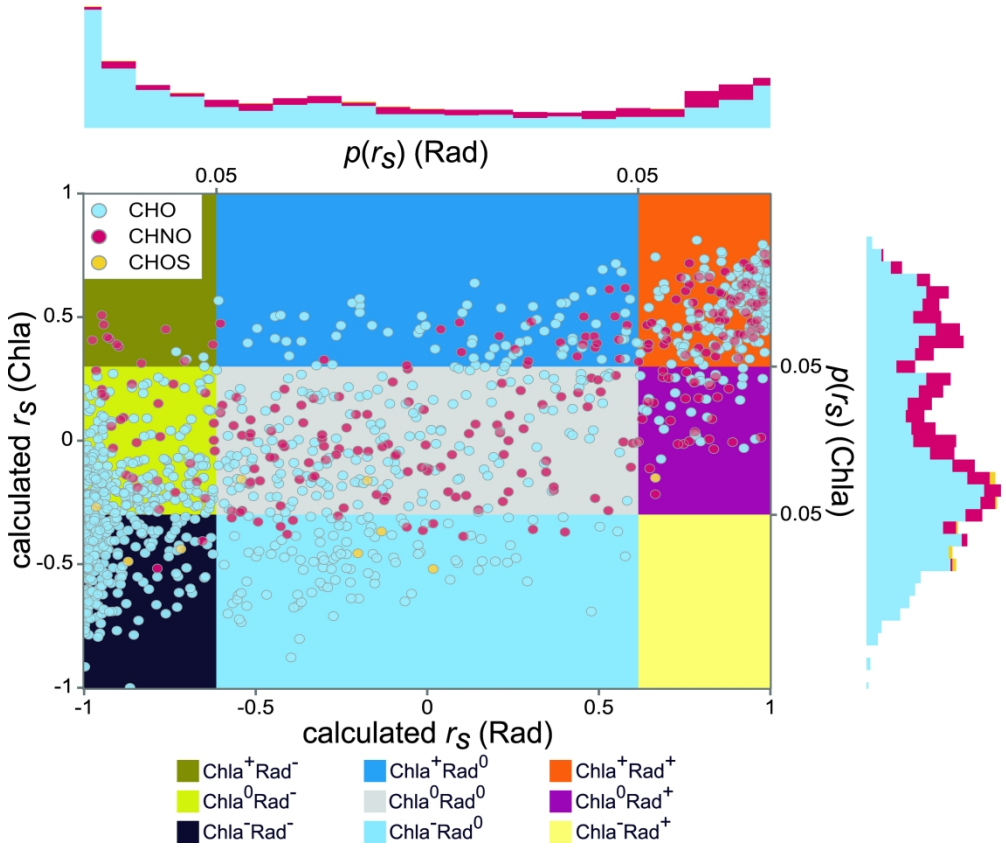


Figure 1. Rank correlation coefficients rs and p(rs) derived from MF intensity rank and chlorophyll a concentration (Chla; 60 samples) or cumulated sunlight irradiation (Rad; 13 samples). The reactivity classes as defined in Table 1 are highlighted with colors. MF classes are denoted as CHO (blue), CHNO (red), and CHOS (orange) and their rs frequency distribution is shown as marginal histograms.

141x119mm (600 x 600 DPI)

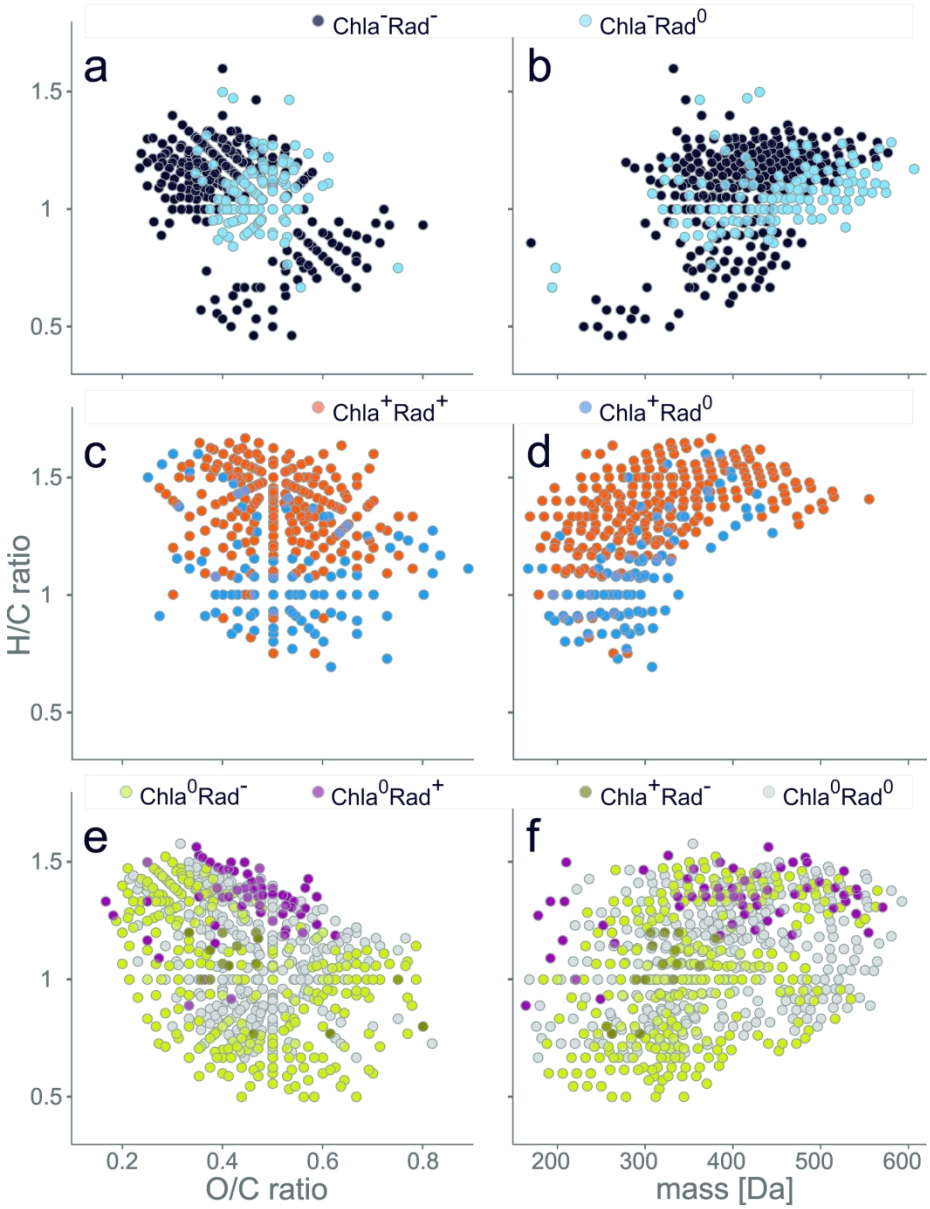


Figure 2. Van Krevelen (a, c, e) and H/C vs mass (b, d, f) diagrams for 1277 MF with colors according to their assigned reactivity class from calculated rs values. The distribution into CHO and CHNO MF into the reactivity classes is shown in Figures S6-7.For the definition of the classes, refer to Table 1.

84x110mm (600 x 600 DPI)

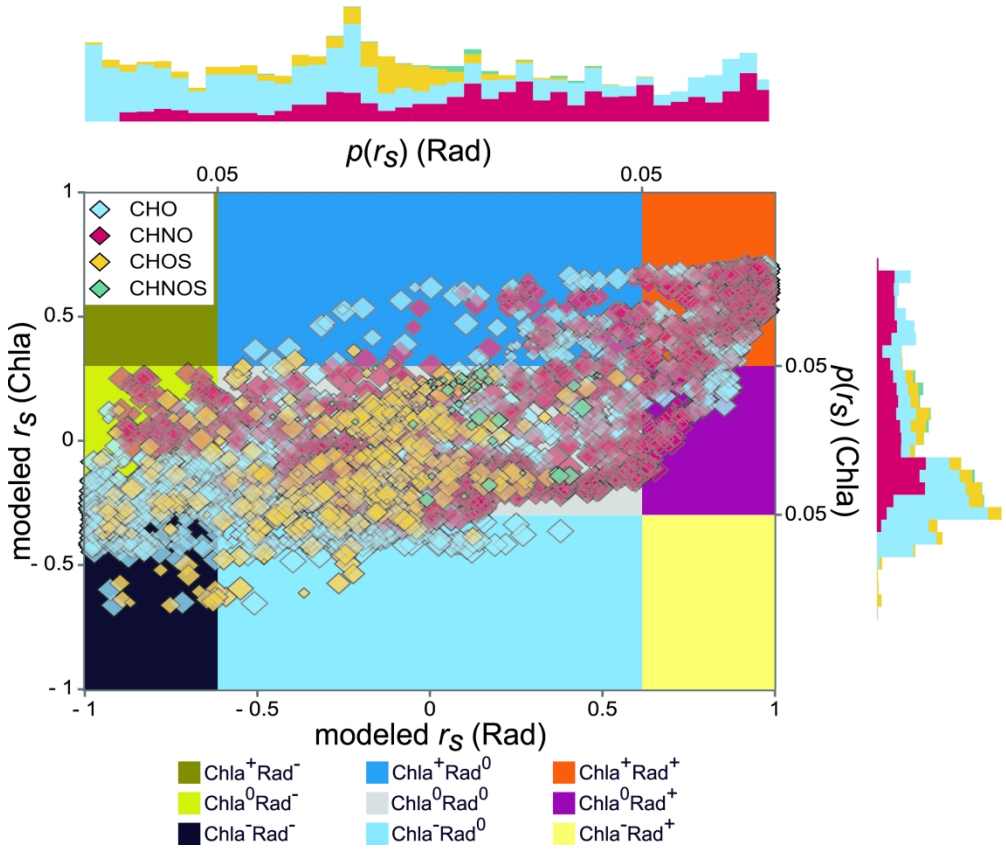


Figure 3. Modelled rank correlation coefficients rs and significance p(rs) for 3162 MF, detected at least once in both data sets. The reactivity classes as defined in Table 1 are highlighted with colors (same colors as in Figure 1). MF classes are denoted as CHO (blue) and CHNO (red), CHOS (orange), and CHNOS (cyan) and their modeled rs frequency distribution is shown as marginal histograms. The size of the diamonds indicates the frequency of detection (2 – 72) within the both data sets.

141x118mm (600 x 600 DPI)

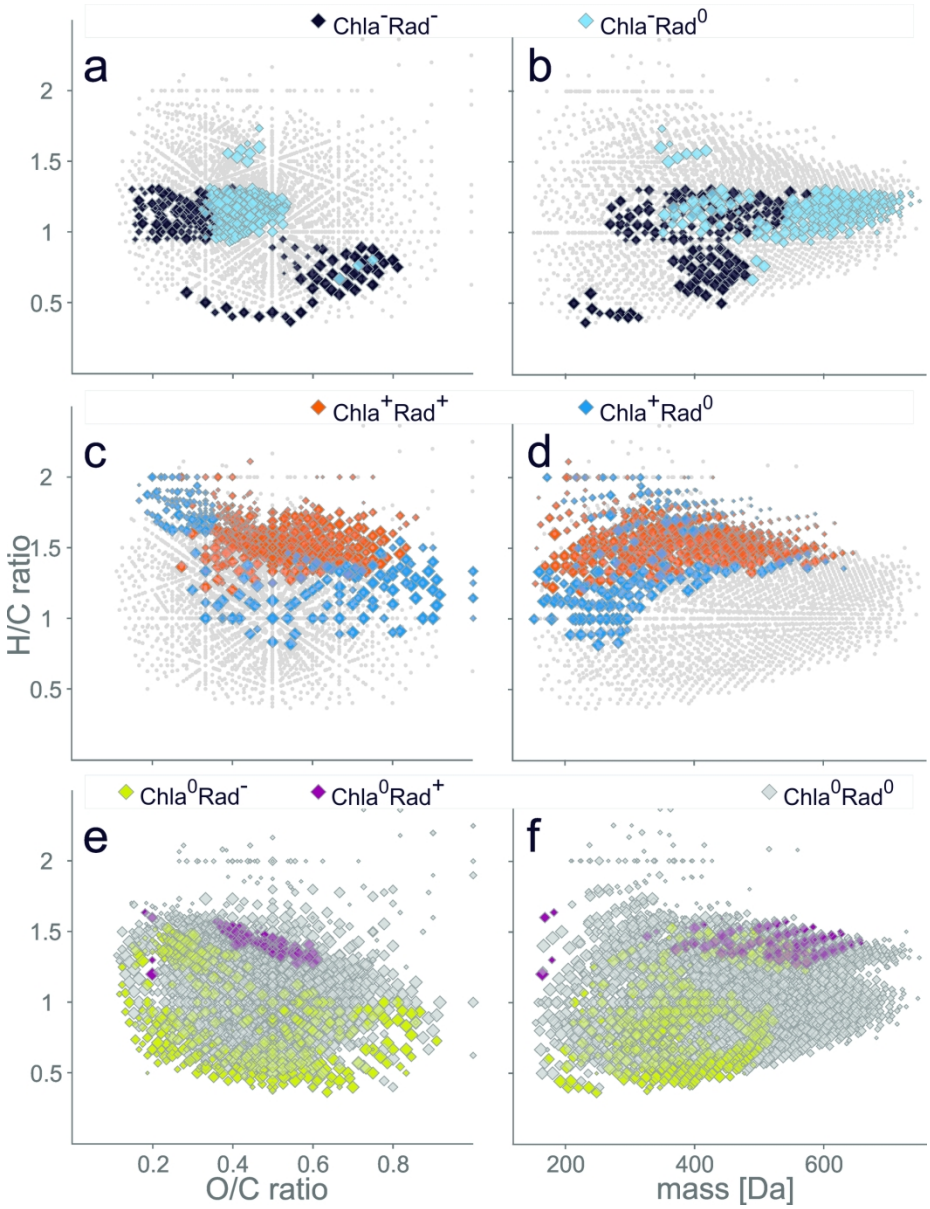


Figure 4. Van Krevelen (a, c, e) and H/C vs mass (b, d, f) diagrams for 3162 MF detected at least once in both data sets. The size of the diamonds indicates the frequency of detection (2 – 72) within the both data sets. The small grey circles in a-d represent the full chemical diversity of all 3162 MF used for the model. . For the definition of the classes, refer Table 1.

84x110mm (600 x 600 DPI)