Nanoparticles and dissolved organic matter: a network perspective reveals a decreasing diversity of the materials investigated – Supporting Information Appendix

Nicole Sani-Kast[a], Jérôme Labille[b,c], Patrick Ollivier[d], Danielle Slomberg[b,c], Konrad Hungerbühler[a], and Martin Scheringer[a,e,*]

[a]Institute for Chemical and Bioengineering, ETH Zürich, CH-8093 Zürich, Switzerland
[b]Aix-Marseille Université, CNRS, IRD, CEREGE UMR 7330, 13545 Aix-en-Provence, France
[c]International Consortium for the Environmental Implications of Nanotechnology, iCEINT, Aix-en-Provence, France
[d]BRGM, 3 Avenue C. Guillemin, BP 36009, 45060 Orléans, France
[e]RECETOX, Masaryk University, 625 00 Brno, Czech Republic
[*]Corresponding author: scheringer@chem.ethz.ch

ABSTRACT

Dissolved organic matter (DOM) strongly influences the properties and fate of engineered nanoparticles (ENPs) in aquatic environments. There is an extensive body of experiments on interactions between DOM and ENPs and also larger particles (we denote particles on the nano- and micrometer scale as particulate matter, PM). However, the experimental results are very heterogeneous and a general mechanistic understanding of DOM-PM interactions is still missing. In this situation, recent reviews have called to expand the range of DOM and ENPs studied. Therefore, our work focuses on the diversity of the DOM and PM types investigated. Because the experimental results reported in the literature are highly disparate and difficult to structure, a new format of organizing, visualizing and interpreting the results is needed. To this end we perform a network analysis of 951 experimental results on DOM-PM interactions. This enables us to analyze and quantify the diversity of the materials investigated. The diversity of the DOM-PM combinations studied has mostly been decreasing over the last 25 years. This is driven by an increasing focus on several frequently investigated materials such as DOM isolated from fresh water, DOM in whole-water samples, and TiO_2 and silver PM. Futhermore, there is an underrepresentation of studies into the effect of particle coating on PM-DOM interactions. Finally, it is of great importance that the properties of DOM used in experiments with PM, in particular the molecular weight and the content of aromatic and aliphatic carbon, are reported more comprehensively and systematically.

Contents of this document

- **Consistency of reported DOM parameters:** (i) a flowchart of the distribution of group-1 DOM in the publications in the database; (ii) a summary of the availability of key parameters of group-1 DOM (**pp. 2–4**).
- **Frequency and prevalence of materials in the experiments:** (i) a list of abbreviations of particulate matter (PM) and dissolved organic matter (DOM) types that appear in the database; (ii) a heatmap representation of the experimental network (Figure S2). It depicts the frequency of employment of each given DOM-PM combination in the experiments by means of a color gradient; (iii) two sets of bar charts that summarize the temporal trends in

the use of the various materials, separately for DOM and PM (Figures S3 and S4); (iv) a graphical representation of the distributions of whole-water-sample DOM vs. isolated DOM in the empirical network (Figure S5) (**pp. 4–11**).

- **Environmental sources and chemical composition of dissolved organic matter:** Detailed discussion of the results of the principal component analysis of dissolved organic matter (presented in the main text). It explains the distribution of the 80 DOM types analyzed in the lower-dimensional space of their carbon distribution. It presents three figures that summarize this analysis, Figures S6 to S8 (**pp. 12–16**).
- **Comparison to random networks:** Quantitative differences between the empirical network and an ensemble of simulated networks created by a random linking process (**p. 17**).
- **Simulated networks of high and low diversity:** Detailed discussion of the source of variability in the diversity of DOM-PM combinations obtained for the simulated networks of high and low diversity (**p. 17**).
- **Resilience of the empirical network:** Results of tests that investigate the influence of various modifications in the database on the observed diversity of DOM-PM combinations. Specifically: (i) the effect of considering only experiments that employ humic substances as their DOM component; and (ii) bootstrap analysis of the publications in the database (**pp. 18–21**).
- **Temporal dimension of the network's structure:** The empirical network with links colored by the year of study (**p. 21**).
- **Annex A:** Output of the principal component analysis as obtained from the statistics software R (**p. 22**).
- **Annex B:** The R code used to generate the and analyze the bootstrap of the publications in the database (**pp. 22–23**).
- **References:** References for this document (**p. 23**).

Consistency of reported DOM parameters

DOM purchased from the International Humic Substances Society

More than 50% of the publications in our database that employ DOM purchased from the International Humic Substances Society (IHSS) do not explicitly report the unique identifier number (Cat. No.) of the DOM. Overall there are 113 publications that employ DOM obtained from the IHSS, out of which 60 do not explicitly state the unique identifier of these materials (Figure S 1). While some publications state that they use the *reference* material, many simply state the name of the material. As explained in the section *Sources and chemical composition of dissolved organic matter from different natural environments* (below), different samples of IHSS DOM from the same source exhibit variations in their chemical composition. These could originate from natural variations and/or variations due to potential changes in the extraction process.

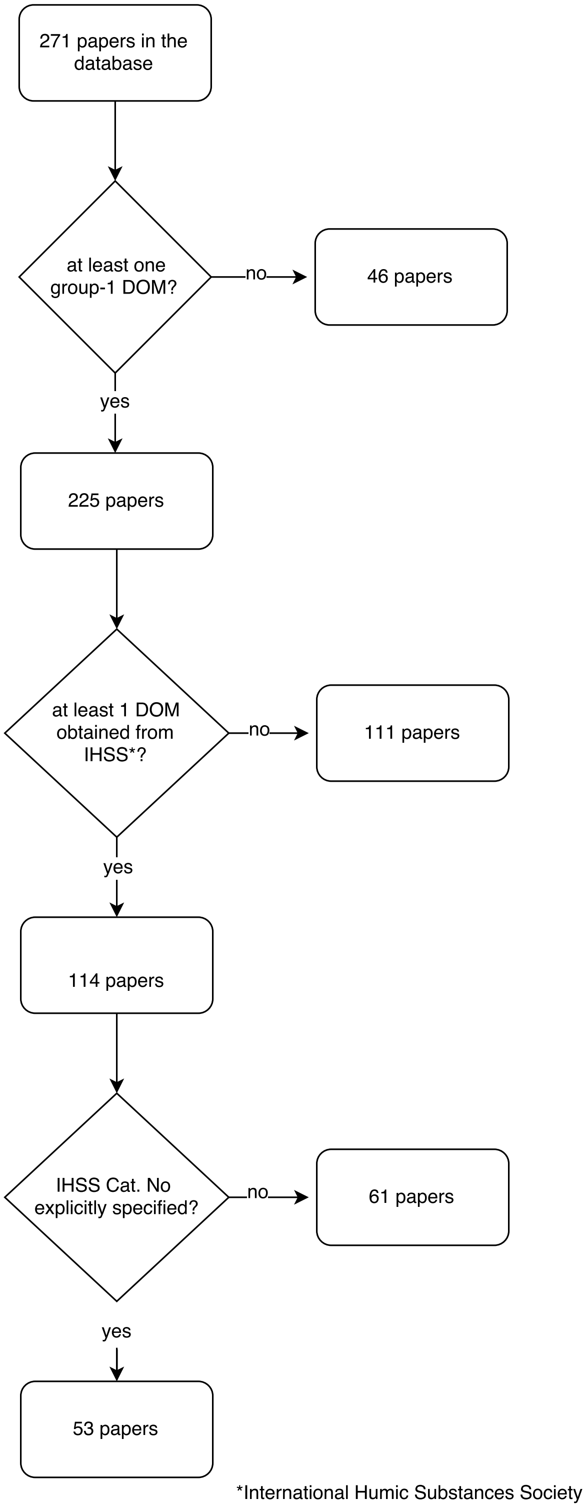
271 papers in the database
at least one group-1 DOM?
no
46 papers
yes
225 papers
at least 1 DOM obtained from IHSS*?
no
111 papers
yes
114 papers
IHSS Cat. No explicitly specified?
no
61 papers
yes
53 papers
*International Humic Substances Society

Figure S 1. Fractionation of the publications in the database according to their usage of DOM from the IHSS.

Availability of group-1 DOM properties

There are 62 publications that employ at least two group-1 DOM types. In 18 of these publications molecular weight is unknown for some of the studied group-1 DOM, and in 17 publications molecular weight is unknown for all of the studied group-1 DOM (similar proportions are found also for elemental composition, SUVA and ^{13}C NMR spectra). This observation means that in such publications where a given property – such as molecular weight, elemental composition, SUVA and ^{13}C NMR spectra – is known for only some of the studied DOM (if at all), this property cannot be used to interpret the results. There are only 21 publications that employ more than one group-1 DOM type and for which all the above parameters are known / measured for all the employed group-1 DOM.

Frequency and prevalence of materials in the experiments

Abbreviations

Tables S1 and S2, see next pages, detail the abbreviations of the different types of particulate matter (PM) and dissolved organic matter (DOM) employed in the experiments included in our database.

Table S 1. Particulate matter (PM) abbreviations. NP: nanoparticles

abbreviation	meaning
Acr-Au	acrylate-stabilized gold NP
Al-SiO_2	Al^+-functionalized silica NP
AmherstHA-Al_2O_3	Al_2O_3 coated with Amherst peat soil humic acid
AmherstHA-TiO_2	AmherstHA-coated TiO_2 NP
AmherstHA-ZnO	AmherstHA coated ZnO
amine-peg-QDCdSe ZnS	amine polyethylene glycol (PEG)-functionalized CdSeZnS
amine-QDCdSe	amine-functionalized cadmium selenium, quantum dot (eFluor)
ARSHAP	alizarin red S labeled hydroxyapatite nanoparticles
B	boron
carb-paa-QDCdTe CdS	carboxylic polyacrylic acid-functionalized CdTe CdS
carb-peg-QDCdSe ZnS	carboxylic polyethylene glycol (PEG)-functionalized CdSe ZnS
carb-QDCdSe	carboxyl functionalized cadmium selenium, zinc sulfate quantum dot (eFluor)
Dmsa-TiO_2	dmsa (dimercaptosuccinic acid) coated TiO_2 NP
Fh	ferrihydrite
Ga-Ag	gum-arabic coated Ag NP
Hf(0.37)ZrO_2(0.63)	particles containing a mixture of Hf and ZrO_2
Hmc-Ag	hydroxylammonium chloride stabilized Ag NP
latex-amidine	amidine modified latex particles
latex-sulf	latex sulfonate
Lig-TiO_2	lignin-coated TiO_2 NP
Mag-Fe_2O_3	magnetic Fe_2O_3
mix1-AlSiO	mixture of 27% montmontmorillonite, 24% illite and 38% kaolinite
mix2-AlSiO	mixture of 20% montmontmorillonite, 29% illite and 45% chloride
mix3-AlSiO	mixture of 45% montmorilonite, 18% illite and 30% kaolinite
Mua-Au	11-mercaptoundecanoic acid coated gold NP
oa-QDCdSe	oleic acid coated cadmium selenium quantum dots
Oh-SiO_2	OH-functionalized silica NP
PAA-CeO_2	polyacrylic acid CeO_2
paa-Fe	polyacrylic acid coated NZVI (nanoscale zero-valent iron)
peg-QDCdSe	polyethylen glycol coated cadmium selenium quantum dot (eFluor)
Peptone-Al_2O_3	peptone-coated Al_2O_3
Peptone-TiO_2	peptone-coated TiO_2 NP
Peptone-ZnO	peptone-coated ZnO
PMMA	poly(methylmethacrylate nanoparticles
PMMA-PHEMA	poly(methylmethacrylate-co-hydroxyethylmethacrylate) nanoparticles
PMMA-PSMA	poly(methylmethacrylate-co-stearylmethacrylate) nanoparticles
Pogto-Ag	polyoxyethylene glycerol trioleate coated Ag NP
protein-Ag	protein-capped Ag NP
Pvp-Ag	polyvinylpyrrolidone-coated Ag NP
Sds-methacrylate	SDS stabilized methylacrylate NP
srnom-Fe_2O_3	Suwannee River natural organic matter coated Fe_2O_3
starch-Ag	starch-capped AgNP
Ta-Al_2O_3	tannic-acid-coated Al_2O_3
Ta-TiO_2	tannic-acid-coated TiO_2 NP
Ta-ZnO	tannic-acid-coated ZnO NP
teos-Fe	tetraethyl orthosilicate coated Fe NP
TiO_2/AC	TiO_2 and activated carbon
TiO_2-TN	titanium dioxide titanate nanotubes (TiO2HxNa2-xTi3O7)
T-lite	TiO_2 coated with hydrated silica, dimethicone/methicone copolymer, and aluminum hydroxide (purchased from BASF)
Tma-bentonite	tetramethylammonium bromide coated benotite
Tma-kaolinite	tetramethylammonium bromide coated kaolinite
TN	titanate nanotubes (HxNa2-xTi3O7)
topo-Hf(0.37)ZrO2(0.63)	trioctylphosphine oxide coated Hf(0.37)ZrO2(0.63)
topo-HfO_2	trioctylphosphine oxide coated HfO2
topo-QDCdSe	trioctylphosphine oxide coated cadmium selenium quantum dots
topo-ZrO_2	trioctylphosphine oxide coated ZrO_2
tween20-Ag	tween20 coated Ag NP
tween20-Fe	tween20 coated NZVI (Fe)
Z-cote	ZnO coated with triethoxycaprylylsilan (purchased from BASF)

Table S 2. Dissolved organic matter (DOM) abbreviations

abbreviation	meaning
2,3-DHBA	2,3-dihydroxybenzoic acid
BSA	bovine serum albumin
CAPA	capric acid
CAPRYA	caprylic acid
CMC	carboxymethyl cellulose
CTAB	cetyltrimethylammonium bromide (amine-based cationic quaternary surfactant)
EPS	extracellular polymeric substance
HMM acid	hydrophilic macromolecular acids
JBR215	rhamnolipid (a glycolipid)
LA	lauric acid
MWAP71	bacterial polysaccharide
PAA	polyacrylic acid
PAH	polycyclic aromatic hydrocarbon
PAM	polyacrylic co maleic acid
PAP	polyaspartate
PGUA	poly(galacturonic acid), a polysaccharide (pectic acid)
PHEMA	poly 2-hydroxyethyl methacrylate
PMA	polymaleic acid
PSS	poly(styrene sulfonate)
PVP	polyvinylpyrrolidone
R95	rhamnolipid (glycolipid)
RMDP17	bacterial polysaccharide
STP effluent	sewage treatment plant effluent
STP influent	sewage treatment plant influent
YAS34	bacterial polysaccharide

Frequency and prevalence of materials in the experiments: heatmap

Figure S 2 depicts in a "heatmap" the prevalence of materials in the experiments included in our database. The sparsity of the experimental field is evident here by the large blank areas in the heatmap. These blank areas correspond to combinations of PM and DOM that have not been studied yet.

Frequency and prevalence of materials in the experiments: bar charts

The temporal distributions of the DOM types and PM types used in the experiments between 1990–2015 are shown in Figure S 3 and Figure S 4, respectively. These Figures support the interpretation of Figures 3b and 3c in the main text.

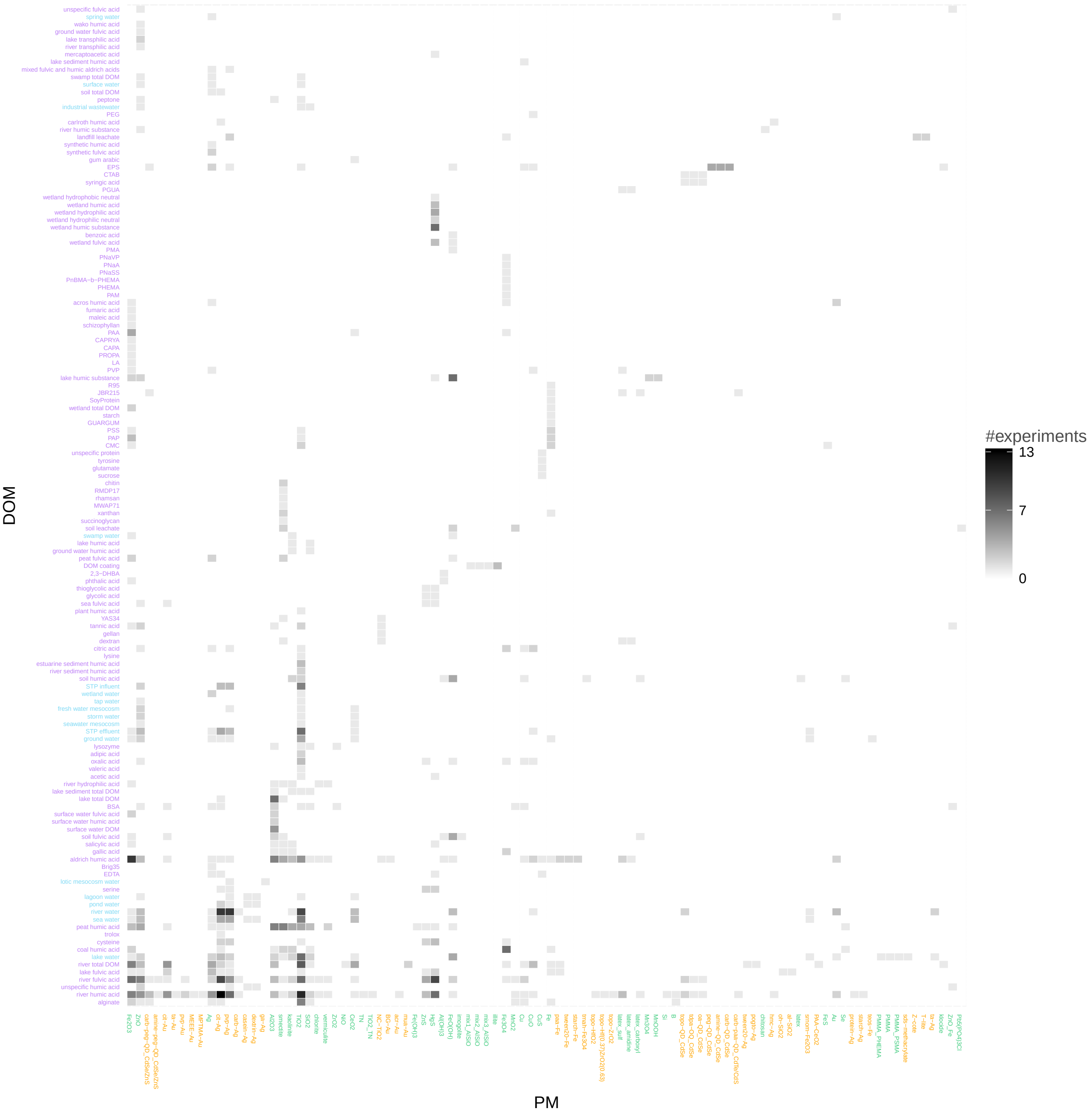

Figure S 2. Prevalence of DOM-PM combinations investigated in the experiments in our database. The horizontal and vertical axes list the various PM and DOM types employed in the experimental papers analyzed, respectively. PM with initial coating and uncoated PM are listed in orange and green colors, respectively. Isolated DOM is listed in purple, and water samples are listed in light blue. The color intensity corresponds to the number of times a given DOM-PM combination was employed in the experiments.

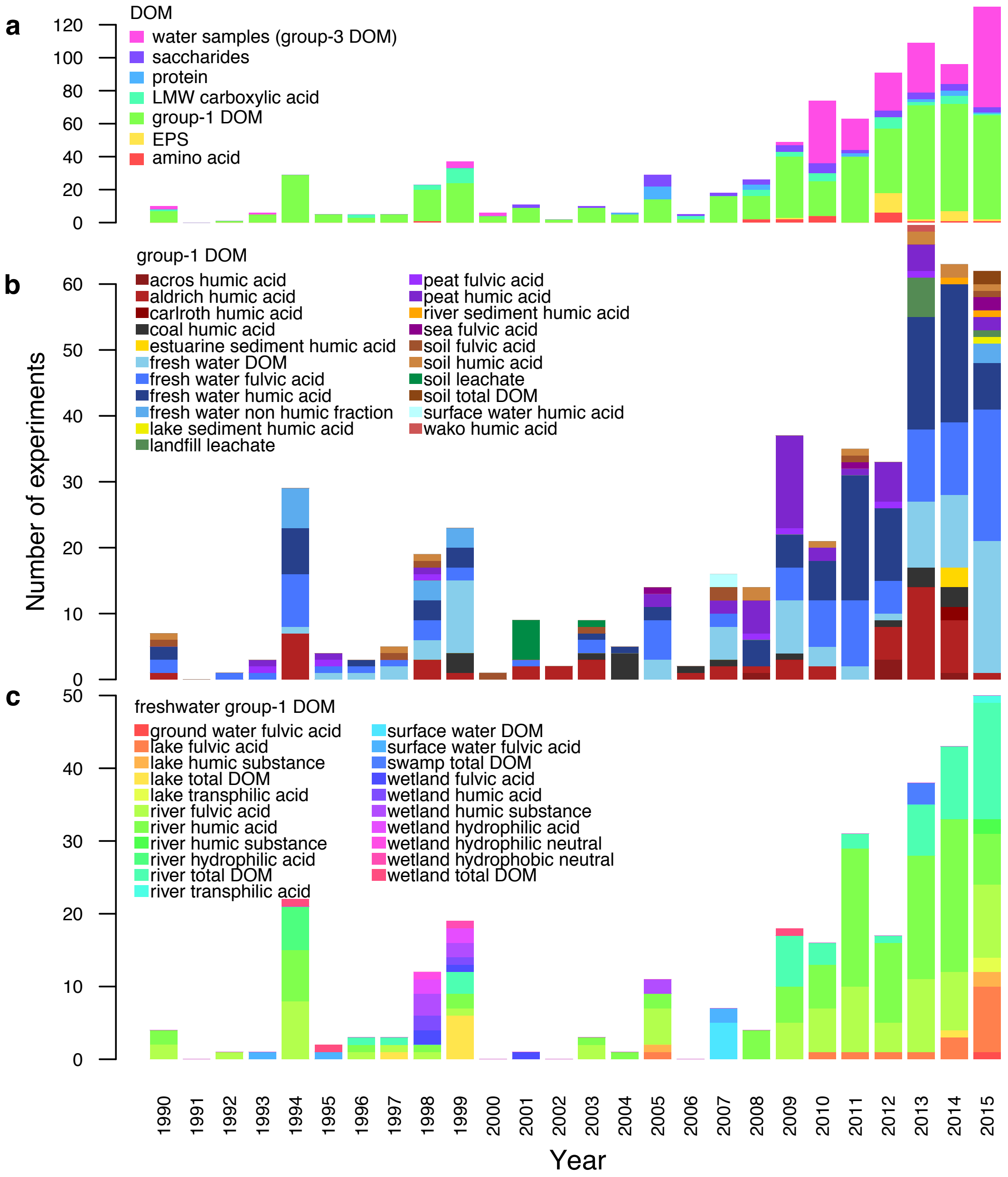

Figure S 3. Temporal trends in the use of DOM types between 1990–2015. **a,** distribution of group-1, group-2 and group-3 DOM, where group-2 DOM is separated into specific groups of material types; **b,** temporal distribution of group-1 DOM only; **c,** temporal distribution of fresh water group-1 DOM only.

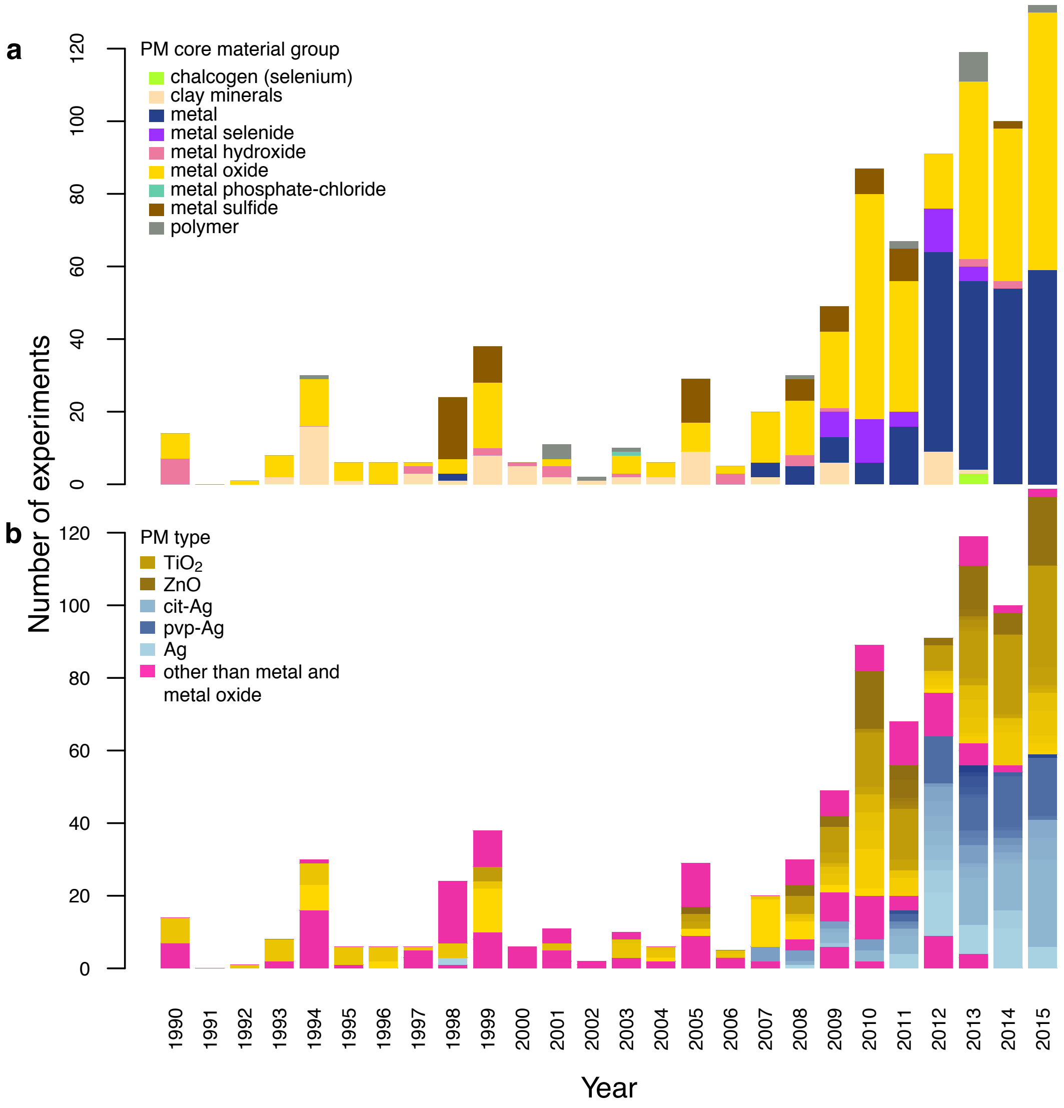

Figure S 4. Temporal trends in the use of PM types between 1990–2015. **a,** PM aggregated based on groups of core materials; **b,** temporal trends of selected PM types.

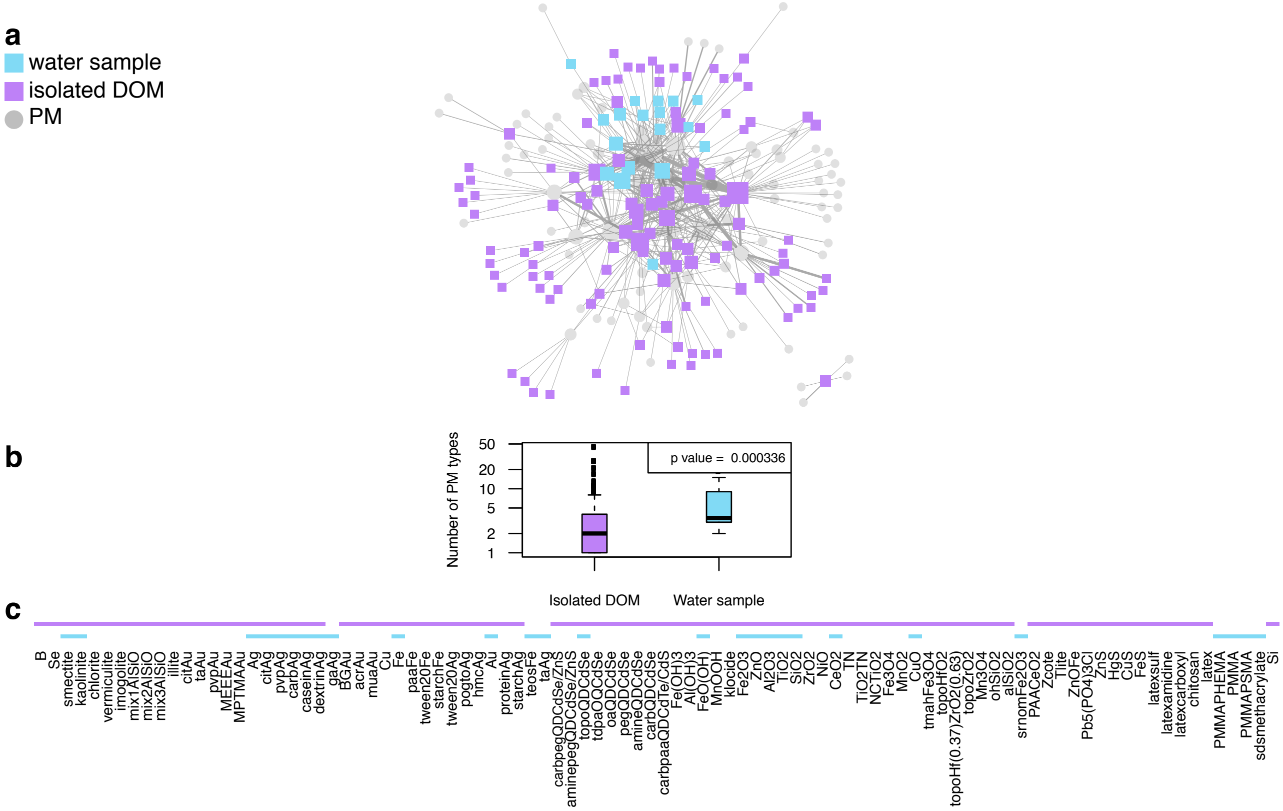

Figure S 5. Differences between water samples vs. isolated DOM in the empirical network. **a,** the empirical network; nodes representing water samples are colored in light blue; nodes for isolated DOM are colored in purple. The size of the nodes is proportional to the degree of the nodes. **b,** boxplots showing the distributions of the number of PM types studied with water-sample DOM vs. the number of PM types studied with isolated DOM, i.e., the degrees of the DOM nodes. The p value is for the two-sample, non-parametric Mann-Whitney U test (alternative hypothesis: water samples are studied with more PM types than isolated DOM). **c,** range of PM types studied with isolated DOM (purple) vs. water samples (light blue). The PM types labeled at the bottom are ordered by type (e.g. clay minerals, metals and metal oxides).

Environmental sources and chemical composition of dissolved organic matter

Figure S 6 depicts the distribution of the 80 DOM samples that were used in the principal component analysis (PCA) in the two-dimensional space spanned by the first two principal components (PC1 and PC2). Figure S 7 depicts the same distribution, but without the labels of the individual samples and with several domains of certain DOM types highlighted.

The PC1 and PC2 values of a given DOM sample are the weighted averages of its aliphatic, aromatic and carbonyl carbon contents, and are defined as:

$$\mathrm{PC1} = 0.67 \cdot \%\mathrm{C}_{\mathrm{aliphatic}} - 0.57 \cdot \%\mathrm{C}_{\mathrm{aromatic}} - 0.47 \cdot \%\mathrm{C}_{\mathrm{carbonyl}},$$
$$\mathrm{PC2} = -0.59 \cdot \%\mathrm{C}_{\mathrm{aromatic}} + 0.81 \cdot \%\mathrm{C}_{\mathrm{carbonyl}}.$$

Fulvic acids, in general, have higher values of PC2 compared to humic acids, which reflects their lower aromatic carbon content, see Figure S 6. This observation is in accordance with the literature, reporting lower aromatic carbon content in fulvic acids than in humic acids from the same environment[1].

Most marine DOM samples in this analysis have high PC1 values compared to the DOM samples from non-marine environments (i.e. marine DOM has high PC1 values and is mostly located to the right of the other DOM samples in Figure S 7). This indicates that marine DOM has a higher content of aliphatic carbon compared to DOM from fresh water and from soil. This observation is in agreement with differences between DOM from marine and terrestrial environments, as reported in the literature. Specifically, DOM originating from terrestrial sources tends to have high aromatic carbon content compared to marine DOM, whereas the latter exhibits high content of branched aliphatic carbons[2].

There are several relatively small regions in Figure S 7 that correspond to DOM samples that share similar values of the first two PCs. This reflects high similarity of chemical composition such as the one observed for river fulvic acids (region 1) and river humic acids (region 3, with Ogeechee river humic acid being an outlier and marked with C). Our results confirm previous results that show high similarity of river fulvic acids across rivers of different characteristics[1]. Another relatively small region is that of estuarine sediment humic acid (region 5), which exhibits both high aliphaticity as well as relatively high aromatic carbon content. An exception is Chesapeake bay sediment humic acid (dot marked by F), which has higher aromaticity than the other estuarine sediment humic acids. Regions of larger sizes correspond to DOM samples from similar environments that exhibit larger variability in one or two of the PCs values. For example, soil fulvic acid fractions in region 2 span a wide range of PC1 values, corresponding to a large variability in the aliphatic carbon content. However, the narrow range of PC2 values for region 2 indicates that all samples of soil fulvic acid fractions in our analysis have similar content of aromatic carbon. Overall, the majority of DOM from aquatic sources lie in the upper left and upper right regions, with the upper left region being dominated by freshwater DOM.

The proximity of aldrich humic acid, labeled with D in Figure S 7, to the soil and coal humic acids provides an indication that this material has terrestrial origin, as was suggested multiple times in the literature.

Soil humic acids are the most scattered material in the PC1-PC2 space. They occupy all but the upper right quadrant of the space (Figure S 6). This corresponds to a diverse chemical composition of soil humic acids, which can originate from either the high heterogeneity of soil composition or sensitivity of sample compositions to DOM extraction procedures, or both. The chemical composition of soil humic acids ranges from materials having both high aliphatic and aromatic carbon content (lower-right quadrant), high aromatic and low aliphatic carbon content (lower-left quadrant), and low aliphatic and low aromatic carbon content (upper-left quadrant). The effect of the DOM extraction and purification methods can not be ruled out as a contributor to the high variability in the measured chemical composition of soil humic acids. Notably, it has been suggested that extraction and purification methods of soil humic acids may affect the measured carbon content to an extent that can modify the measured aromaticity of the extracted fulvic and humic acid fractions[1]. Moreover, the International Humic Substances Society (IHSS) states that humic acids isolated from soils are not operationally equivalent to humic acids extracted from aquatic origin, since the former ones may contain, in addition, hydrophilic acids[3].

We strongly recommend that authors of experimental papers always report the reference number of DOM samples that were purchased from the IHSS. Our results show that chemical composition of different reference samples of DOM from the same source can vary substantially. We observe that in our DOM-PM experimental database the reference number of DOM from IHSS is not always reported (see section *Consistency of reported DOM parameters* above). Such missing information impedes any attempts to chemically characterize the DOM employed in the various experiments, and therefore makes it difficult or even impossible to compare quantitative results from different experiments. We observe that the variability in chemical composition of different reference materials from the same source increases in the following order: river fulvic acids (e.g. *A1* and *A2*), river humic acids (e.g. *B1* and *B2*), and soil humic acids (e.g. *E1* and *E2*).

Figure S 8a compares the distribution of euclidean distances, in the PC1-PC2 space, between materials from the same environment type to the distribution of distances among materials from different environment types (note that distances from a given material to itself were excluded in order not to bias the distribution towards lower values by including the zero distance from a material to itself). Overall, materials from the same environment types tend to be located closer to each other in the PC1-PC2 space. This proximity, which reflects similarities in the carbon distribution of the respective materials, is unlikely to be coincidental: when the DOM sample locations in the PC1-PC2 space were associated randomly with environmental sources, the difference between the two distributions disappears (see Figure S 8b).

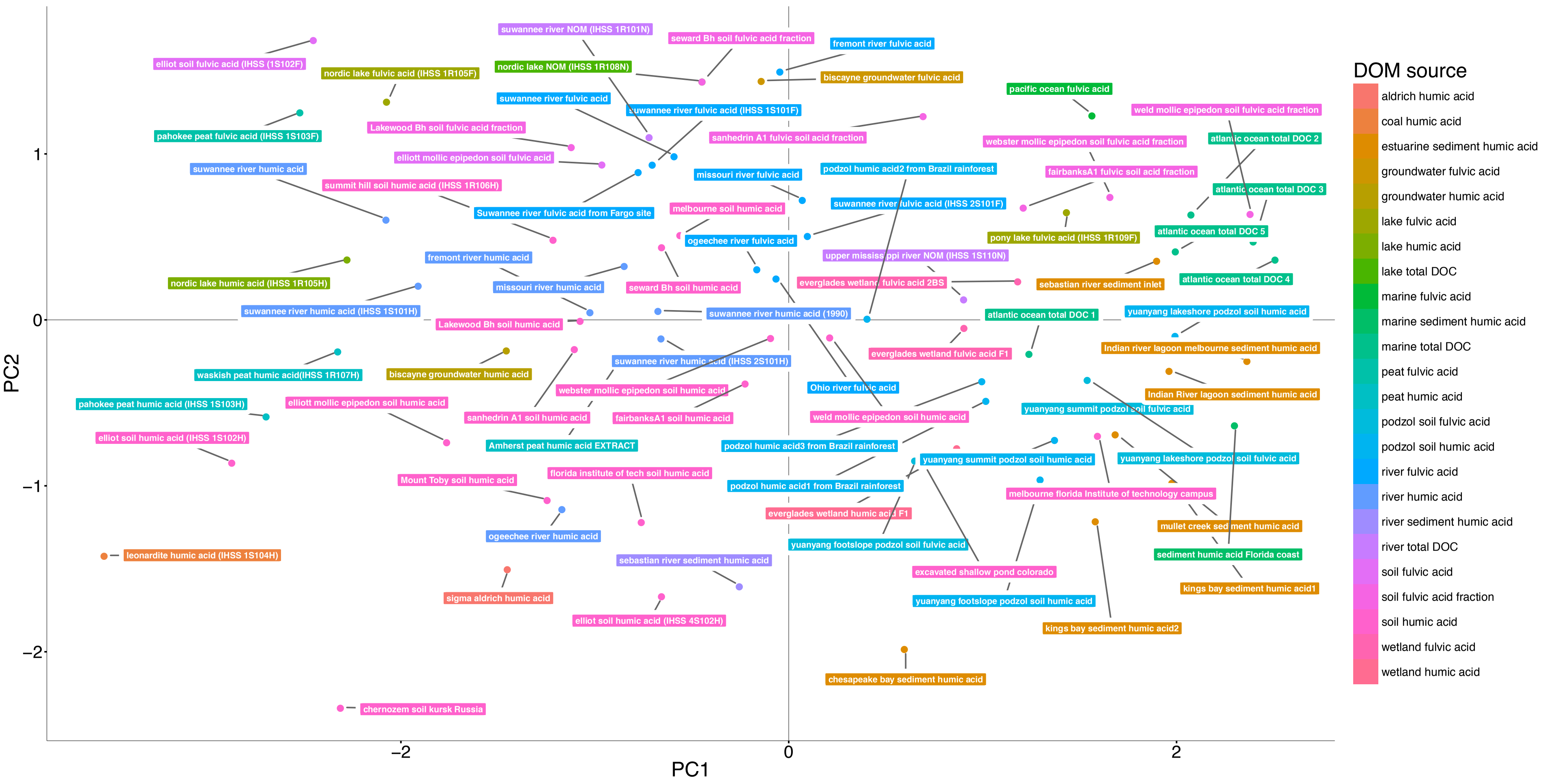

Figure S 6. Distribution of various DOM samples in a lower-dimensional representation of their carbon distribution space, given by the first two principal components. The environmental sources of the materials are listed in the legend and in the labels of the individual data points.

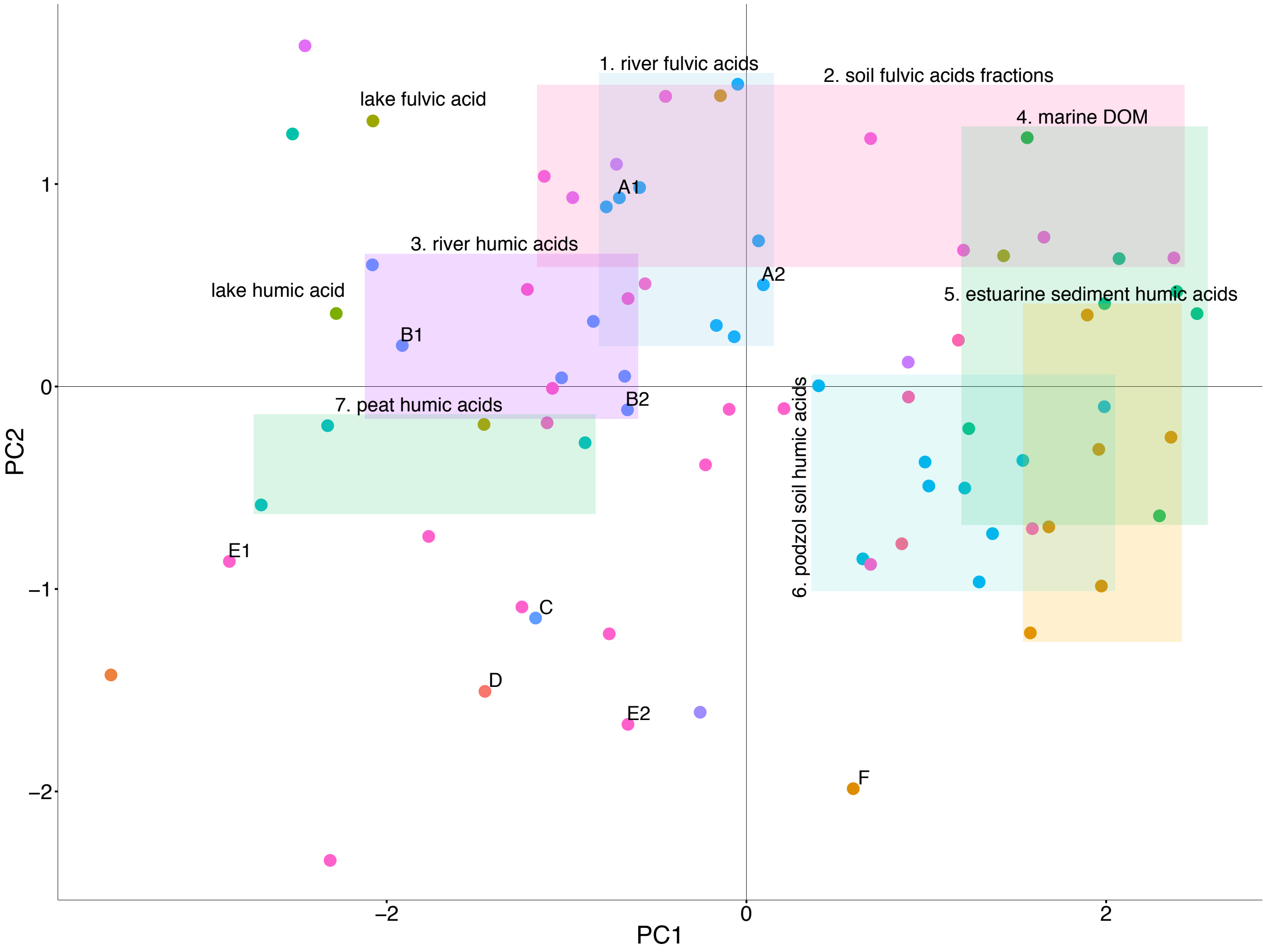

Figure S 7. Distribution of the unlabeled DOM samples in a lower-dimensional representation of their carbon distribution space, given by the first two principal components. The colored rectangles mark regions that include sets of materials from similar environments, as indicated by their labels. **A1-2**, **B1-2** and **C1-2** are different IHSS standards of Suwanee river fulvic acid, Suwanee river humic acid and Eliot soil humic acid, respectively. **D** is Aldrich humic acid, and **F** is Chesapeake bay sediment humic acid.

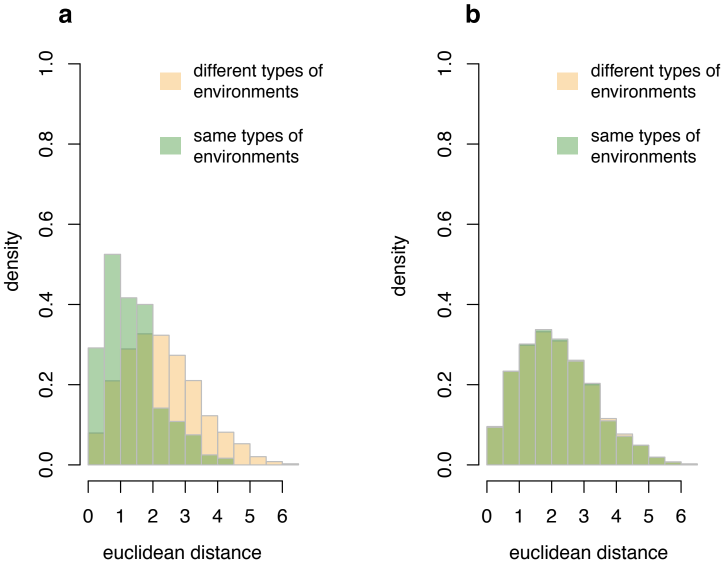

Figure S 8. Euclidean distances between the different DOM types in the two dimensional space spanned by the first two principal components. The distribution of the distances is separated into distances among DOM fractions from the same environment type (e.g. river fulvic acids) and distances among DOM fractions from different environment types (e.g. river fulvic acid vs. soil fulvic acid). **a,** distances as calculated on the original data (Figure S7). **b,** distances calculated for randomized data (i.e. randomized data resulted in disassociation between the location in the principal component dimensions in Figure S7 and the type of environmental source of the DOM).

Comparison to random networks

The following section highlights the most important topological features in the empirical network. To specify these features, we identify which features are unlikely to be found in an ensemble of random networks of comparable size and link density as our empirical network.

The empirical network has a density, p, that is the fraction of existing links out of all possible links, i.e. the number of DOM-PM combinations studied out of all possible combinations. The density is calculated as $\rho = \frac{|E|}{|U| \cdot |V|}$; E, U and V are the sets of links, of nodes of DOM type, and of nodes of PM types, respectively. For the empirical network investigated here, $\rho = 4.3\%$.

The random networks were simulated by means of the $G(n,p)$ model implementation for a bipartite network. For each pair of nodes a link is created with probability p, here $p = \rho$. This ensures that the random and empirical networks have similar densities (see Table S3), and therefore the main difference between the networks is the configuration of links, which causes different topological features.

Table S 3. Comparison between main parameter values of the empirical network and of 1000 random realizations of networks obtained by the $G(n,p)$ model adjusted for bipartite networks.

parameter	empirical	random [a]
mean degree[b]	4.7	$[4.3, 5.1]$
diameter[c]	6 *	$[7, 9]$
density ρ[d]	0.043	$[0.039, 0.046]$
degree assortativity[e]	-0.31 *	$[-0.22, -0.054]$

[a] 95% confidence interval for 1000 random network realizations defined as the range of values between the 2.5–97.5% empirical quantiles.

[*] Value lies outside the 95% confidence interval defined above.

[b] The mean degree is the average number of links per node in the network.

[c] The network diameter is the longest geodesic (shortest) path found for any pair of nodes in the network

[d] Network density (ρ), describes the fraction of links present in the network out of all possible links, $\rho = \frac{|E|}{|U| \cdot |V|}$, where $|E|$, $|U|$ and $|V|$ are the numbers of links, of DOM nodes, and of PM nodes in the network.

[e] Degree assortativity (r) is the correlation between the degree (number of links) of neighboring nodes, $r \in [-1, 1]$; for a completely random network $r \approx 0$.

Simulated networks of high and low diversity

In the high-diversity simulation each experiment almost always adds new DOM-PM combinations to the network, i.e. creates new links. Therefore, the diversity index, D_{comb}, is similar for all 1000 high-diversity networks simulated in each year (indicated by the narrow D_{comb} ranges in the upper series of boxplots, Figure 2a in the main text). On the other hand, D_{comb} values vary considerably between the 1000 low-diversity networks simulated in each year (lower series of boxplots, Figure 2a in the main text). This variability results from the way DOM-PM combinations are sampled in the construction of the low-diversity networks; when all DOM-PM combinations in the cited references of a given paper were studied with the same frequency, the DOM-PM combinations assigned to that given paper are sampled uniformly at random from the previously studied ones, which yields a large variation in the number of combinations studied (n_{comb}) and therefore a large variation in D_{comb}.

Resilience of the experimental network

Bootstrap of publications in the database

In this section we investigate the influence of potential missing publications on the values of the diversity index. The main objective is to inspect whether or not the overall trend (of a decreasing diversity) is robust in light of a possible change in the publication list used to assemble the database. We do this in order to account for the possibility that some publications were not retrieved by the publication search. Our assumption is that the DOM-PM combinations that were studied in the publications analyzed herein are representative, in terms of identity and frequency, of the overall materials studied (i.e., the search queries used to retrieve the publications have no bias towards certain DOM-PM combinations). To perform the analysis, we employed the bootstrap approach by sampling with replacement the publications in the database. Subsequently, we obtained sets of publications, herein referred to as the bootstrap samples, that contain multiple instances (i.e. sampled with replacement) of certain publications and therefore differ from the list of publications that comprises the original database. For each bootstrap sample, similar to the original publication list, the studied DOM-PM combinations were extracted and the D_{comb} was calculated.

Under the assumption of representativeness described above, the variation in the D_{comb} of the bootstrap samples around the empirical D_{comb} reflects the variation in D_{comb} expected from a random sample of publications around the true D_{comb}. To assess the uncertainty in the D_{comb} of the original database, we calculated the 95% bootstrap confidence interval, which is given by:

$$D_i - \widehat{\beta}_i \pm z_{0.975} \cdot \widehat{v}_i^{\frac{1}{2}}, \tag{1}$$

where D_i is the calculated D_{comb} for the original publications list in the year i, $\widehat{\beta}_i$ is the estimated bias of the calculated D_{comb} value in year i, $z_{0.975}$ is the 97.5% quantile of the standard normal distribution, and $\widehat{v}_i^{\frac{1}{2}}$ is the estimated standard error of the D_{comb} value in year i. Both $\widehat{\beta}_i$ and $\widehat{v}_i^{\frac{1}{2}}$ were estimated from the bootstrap samples.

To perform the bootstrap simulation, each combination of DOM-PM presented in the dataset was labeled as ”new“ or ”old“. Experiments labeled as “new” are those experiments carried out in year i that study DOM-PM combinations not studied in previous years. By using these labels, we avoid the situation where the set of PM-DOM combinations of the empirical network is the most diverse one compared to all possible bootstrap samples. In order to account for the temporal aspect, the sampling was stratified by years, which means that in the bootstrap samples each year contained only papers published within that given year. The total number of simulated bootstrap samples was 9999. The analysis was carried out using the *boot* function of the boot package in R; the code is given in Appendix B.

The resulting bootstrap 95% normal confidence interval is depicted in Figure S 9. The confidence interval still demonstrates a general trend of a decrease in D_{comb} over the analyzed period. The uncertainty in the D_{comb} estimates is reduced when the number of publications is higher, reflected by the decrease in the width of the confidence interval after the year 2010.

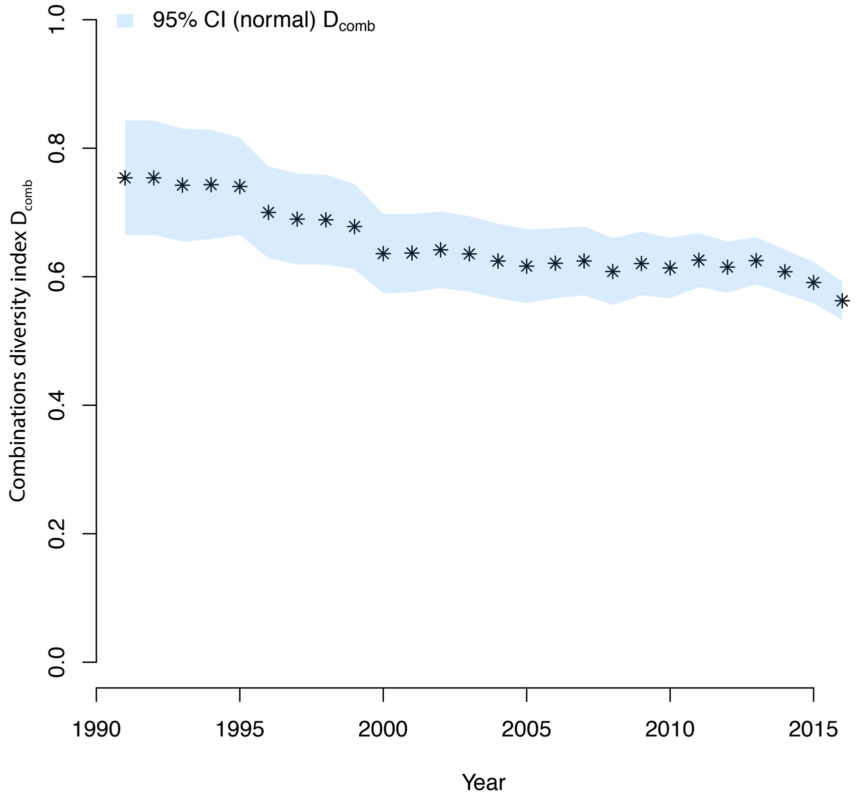

Figure S 9. The empirical D_{comb} values (asterisks) and the 95% bootstrap normal confidence interval of the D_{comb} obtained from the bootstrap samples.

The effect of considering only experiments with humic substances as their DOM constituent

In the empirical network the various DOM types include both humic substances as well as other macromolecules. We here analyze a subset of the network (a "reduced" network) that is comprised of only the experiments that employ humic substances as their DOM constituent. Overall about 49% of the experiments employ humic substances as their DOM constituent (497 of 951 experiments).

Figure S 10 compares the subset network to the original empirical network, and Table S 4 compares their basic properties. The reduced network is smaller than the original network and its density is higher. This implies that in the reduced network there are more types of DOM-PM combinations that were studied relative to the network size, when compared to the original network. However, the diversity index, D_{comb}, is *lower* for the subset network compared to that of the original network, i.e. 0.48 compared to 0.56. The D_{comb} of the subset network implies that the tendency to focus on a small set of DOM-PM combinations is even stronger in the experiments that employ humic substances as DOM. Both networks, empirical and reduced, share the central DOM nodes (e.g."river humic acid" and "aldrich humic acid"), which means that in both cases humic acids are the main focus of the experimental effort. Overall the strong focus towards humic acids creates similarly sparse networks that share similar topological features (i.e. core-periphery structure, which is reflected by the similar negative degree assortativity, Table S 4) and medium diversity of the PM-DOM combinations.

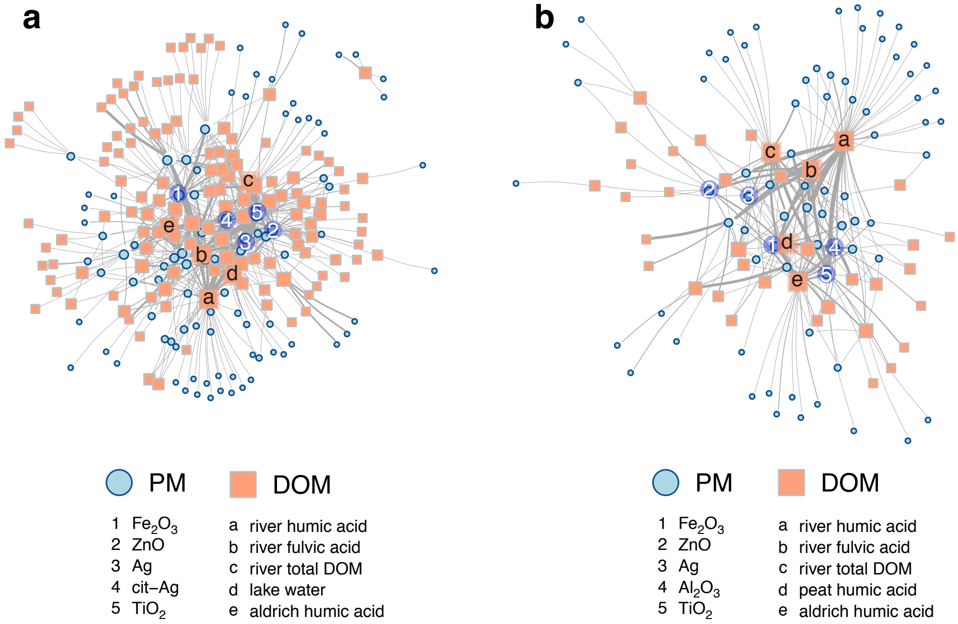

Figure S 10. Comparison between the original empirical network (**a**), and a reduced network containing only the experiments that employ humic substances as their DOM constituent (**b**). The materials listed in the respective legends are the ones that have the highest degree (i.e. were studied with the largest number of counterparts).

Table S 4. Basic properties of the full (empirical) and the reduced network in comparison.

	all DOM	humic substances only
number of nodes	227	112
number of PM nodes	94	74
number of DOM nodes	133	38
number of combinations	535	240
number of experiments	951	497
degree assortativity	−0.311	−0.412
density	0.0428	0.0853
combinations diversity	0.563	0.483

Temporal dimension of the network's structure

Here we inspect whether or not the network's complex topology can be explained by temporal segregation, i.e. if specific regions in the network are more recent than others. To this end, we add a temporal dimension by coloring the links according to the latest publication year of the respective experiments (Figure S 11). We observe that the network is roughly split in half, where the first half has both core (central part) and some outer branches, all of which correspond to rather recent experiments. The other half, which is comprised of older experiments, contains more periphery nodes than central nodes. From this we conclude that recent experiments are not confined to any specific region in the network. Therefore, any imbalance between the study of central vs. peripheral materials cannot be explained by a temporal trend.

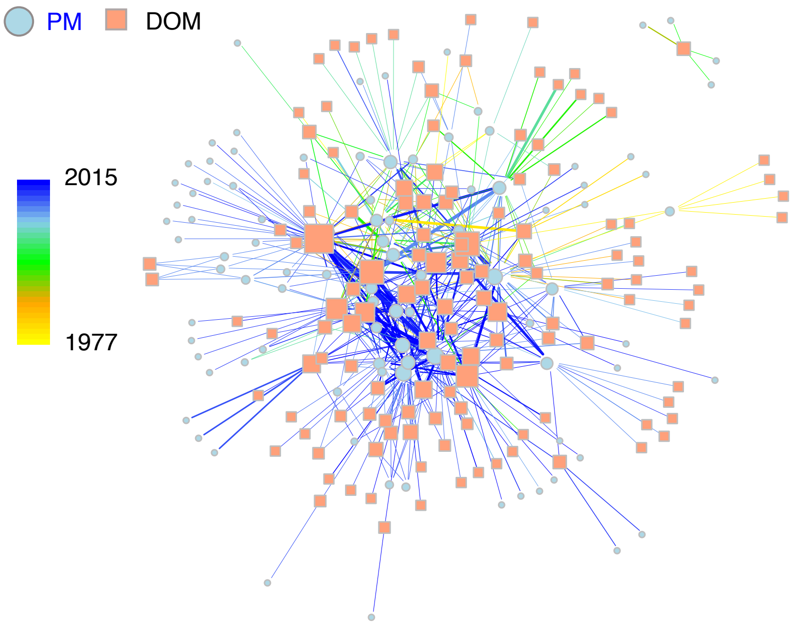

Figure S 11. The empirical network with links colored by latest year of publication. Size of nodes is proportional to their degree; width of links is proportional to the number of experiments studying the connected materials. The color of links corresponds (according to the legend) to the latest year the connected materials (i.e. PM and DOM) were studied together.

Annex A: Principal component analysis – R output

```

Importance of components:
                          Comp.1    Comp.2       Comp.3
Standard deviation     1.4918198 0.8796544 0.0261146260
Proportion of Variance 0.7418421 0.2579306 0.0002273246
Cumulative Proportion  0.7418421 0.9997727 1.0000000000

Loadings:
                     Comp.1 Comp.2 Comp.3
Aliphatic.0..110ppm.  0.669        -0.740
Aroamtic.110..165.   -0.574 -0.588 -0.570
Carbonyl.165..220.   -0.473  0.806 -0.356
```

Annex B: Bootstrap analysis – R code

```
d.data.nom.type$comb  <- sapply(1:nrow(d.data.nom.type),function(x) paste(d.data.nom.type$ENP[x],d.data.nom.type[x,nom.data.column],sep = "-"))
#Label each DOM-PM as "new" or "old"
d.data.nom.type$label.diversity <- rep("old",nrow(d.data.nom.type))
for(i in unique(d.data.nom.type$year)){
 #this loop assigns the label "new" or "old" to each DOM-ENP combination,
 #based on the first definition of diversity
  #unique(d.data.nom.type$comb[!d.data.nom.type$comb[d.data.nom.type$year == i]%in%d.data.nom.type$comb[d.data.nom.type$year < i]])
  #this year combinations:
  current.comb  <- unique(d.data.nom.type$comb[d.data.nom.type$year == i])
  n.new.comb <- sum(!current.comb%in%d.data.nom.type$comb[d.data.nom.type$year < i])#the number of unique and new comb added in the current year
  #now assign the label: "new" to n.new.comb and label: "old" to the rest of the combination in this year
  relv.rows  <- which(d.data.nom.type$year == i)#the rows the correspond to experiments that were published in year i
  #assign "new" label to the first n.new.comb rows from relv.rows and all the rest assign the label: "old"
  d.data.nom.type$label.diversity[relv.rows]  <-  c(rep("new",n.new.comb),rep("old",length(relv.rows)-n.new.comb))
}
#The following is a function to measure materials' diversity.
diversity.func.boot <- function(d.data, year.start){
  #This function calculates the diversity according to the first definition of the diversity index. It takes as arguments the dataset and the year from which to start the calculations
  diversity.trend <- c() # place holder for the trend values to be computed below
  for(year in year.start:max(d.data$year)){
    d.data.slice  <- d.data[d.data$year <= year, "label.diversity"]
    n.experiments  <- length(d.data.slice)#the number of experiments done up to the given year
    n.com <- sum(d.data.slice == "new")#the number of unique combinations studied up to the given year
     diversity.trend  <- c(diversity.trend, n.com/n.experiments)
  }
  return(diversity.trend)
}

bootstrap.publications <- function(data, i, year = 1990){
  #this function takes the data set of all DOIs and the index of the lines that are being resampled, where resampling is stratified by years and calculates the diversity index over the years. Arguments: data = dataset, i = resampled rows, year = the year from which diversity should be calculated, by default is set to 1990
  data  <- data[i,]
  diversity.trend  <- diversity.func.boot(d.data = data, year.start = year)
}
###########################################################################
R  <- 9999
set.seed(1)
publication.boot  <- boot(d.data.nom.type, bootstrap.publications, R = R, strata = factor(d.data.nom.type$year))
#95% normal bootstrap confidence interval:
publication.boot.ci.normal  <- sapply(1:length(publication.boot$t0), function(x) {boot.ci(boot.out = publication.boot, type = c("norm"), index = x, conf = 0.95)$normal[2:3]})
```

References

1. Malcolm RL (1990) The uniqueness of humic substances in each of soil, stream and marine environments. *Anal Chim Acta* 232:19–30.
2. Esteves VI, Otero M, Duarte AC (2009) Comparative characterization of humic substances from the open ocean, estuarine water and fresh water. *Org Geochem* 40(9):942–950.
3. International Humic Substances Society (2016) *Available at: http://www.humicsubstances.org [Accessed: May 18, 2016].*