Title

Unraveling longitudinal pollution patterns of organic micropollutants in a river by non-target screening and cluster analysis

Authors:

Liza-Marie Beckers[a,b*], Werner Brack[a,b], Janek Paul Dann[a,b], Martin Krauss[a], Erik Müller[a,b], Tobias Schulze[a]

[a] Helmholtz Centre for Environmental Research - UFZ, Department of Effect-Directed Analysis, Permoserstr.15, 04318 Leipzig, Germany

[b] RWTH Aachen University, Institute for Environmental Research (Biology V), Department of Ecosystem Analysis (ESA), Worringer Weg 1, 52074 Aachen, Germany

*Corresponding author:

Liza-Marie Beckers (email: liza-marie.beckers@ufz.de)

Abstract

The pollution of aquatic ecosystems with complex and largely unknown mixtures of organic micropollutants is not sufficiently addressed with current monitoring strategies based on target screening methods. In this study, we implemented an open-source workflow based on non-target screening to unravel longitudinal pollution patterns of organic micropollutants along a river course. The 47 km long Holtemme River, a tributary of the Bode River (both Saxony-Anhalt, Germany), was used as a case study. Sixteen grab samples were taken along the river and analyzed by liquid chromatography coupled to high-resolution mass spectrometry. We applied a cluster analysis specifically designed for longitudinal data sets to identify spatial pollutant patterns and prioritize peaks for compound identification. Three main pollution patterns were identified representing pollutants entering a) from wastewater treatment plants, b) at the confluence with the Bode River and c) from diffuse and random inputs via small point sources and groundwater input. By further sub-clustering of the main patterns, source-related fingerprints were revealed. The main patterns were characterized by specific isotopologue signatures and the abundance of peaks in homologue series representing the major (pollution) sources. Furthermore, we identified 25 out of 38 representative compounds for the patterns by structure elucidation. The workflow represents an important contribution to the ongoing attempts to understand, monitor, prioritize and manage complex environmental mixtures and may be applied to other settings.

Abbreviations

BR – Bode River

dd – data-dependent

DOM – dissolved organic matter

DRI – diffuse and random input

HCD - higher energy collision-induced dissociation

HDX - hydrogen-deuterium exchange

HESI - heated electrospray ionization

LC-HRMS – liquid chromatography coupled to high resolution mass spectrometry

LC-MS/MS - liquid chromatography coupled to tandem resolution mass spectrometry

m/z – mass-to-charge ratio

NTS – non-target screening

RT – retention time

WW – wastewater

WWTP – wastewater treatment plant

1. Introduction

Aquatic ecosystems are contaminated with a complex and largely unknown mixture of organic micropollutants emitted from a number of pollution sources (Richardson and Kimura, 2017). Although hundreds of compounds became analyzable in freshwaters by target screening, the large number of unknown components detected in complex and variable environmental mixtures pose a major challenge for monitoring, risk assessment and water management (Altenburger et al., 2015; Brack et al., 2018). Thus, novel approaches are needed to characterize these mixtures, link them to sources and prioritize yet unknown organic micropollutants for identification in order to allow for efficient mitigation (Altenburger et al., 2015).

Non-target screening (NTS) by liquid chromatography coupled to high-resolution mass spectrometry (LC-HRMS) provides an unbiased approach for capturing this complexity. It has been recommended as a monitoring tool (Brack et al., 2019) to identify newly emerging compounds and accidental spills of previously undetected compounds (Hollender et al., 2017) and to understand processes in drinking water (Brunner et al., 2020; Müller et al., 2011) and wastewater treatment (Nürenberg et al., 2015) such as formation of transformation products (Schollée et al., 2015) and degradation of dissolved organic matter (DOM) (Verkh et al., 2018). Furthermore, NTS may complement target screening (Hug et al., 2014; Ruff et al., 2015; Schymanski et al., 2014) and is used in effect-directed analysis to identify unknown toxicants (Muschket et al., 2018; Muz et al., 2017).

NTS generates a huge amount of data, e.g., up to millions of peaks in a set of 360 samples before data treatment (Carpenter et al., 2019) and already about 20,000 peaks in a data set of 10 WWTP effluents (Schymanski et al., 2014). Thus, the application of multivariate statistics becomes inevitable. Using exploratory data analysis tools, the complexity of the data set can be reduced and data structures may be unraveled (Carpenter et al., 2019; Hollender et al., 2017; Schollée et al., 2015). For example, time-trend analysis was recently used to detect temporal changes of individual peaks at the influent of a wastewater treatment plant (WWTP)

(Alygizakis et al., 2019). This is a valid approach for extracting individual compounds with potentially interesting trends. However, in order to draw more general conclusion on mixture dynamics, cluster analysis has been demonstrated as a valuable and time-efficient tool to understand mixture dynamics (Carpenter et al., 2019; Chiaia-Hernández et al., 2017). By means of clustering techniques, e.g. hierarchical clustering, similarities among complex mixtures were identified and sorted into distinct spatial and temporal chemical or ecotoxicological patterns (Carpenter and Helbling, 2018; Carpenter et al., 2019; Chiaia-Hernández et al., 2017; Peter et al., 2018; Zheng et al., 2012). These patterns may reflect source-related or effect-related fingerprints (Brack et al., 2018; Carpenter and Helbling, 2018; Peter et al., 2018; Zheng et al., 2012) and can be used as a prioritization tool for the identification of individual peaks as pattern representatives (Carpenter et al., 2019; Chiaia-Hernández et al., 2017). In a longitudinal setting, the advantages of time-series analysis and the reduction potential of cluster analysis can be combined to identify groups of variables with similar longitudinal behaviour. Genolini et al. (2015) developed a partitioning cluster analysis for longitudinal data ('kml') originally designed for epidemiological data. Here, each variable's course is seen as a trajectory and similar trajectories are clustered together. This approach is potentially faster than a two-step procedure as applied by Chiaia-Hernández et al. (2017) or a stepwise comparison of spatial samples (Ruff et al., 2015). With the application of a novel workflow combining NTS with partitioning clustering, we hypothesized that continuous longitudinal pollution patterns resulting from diffuse and point sources can be distinguished at least in small streams.

The objective of this study was to test this hypothesis using the Holtemme River (Saxony-Anhalt, Germany) as a case study and demonstrate this open-source workflow on a set of water samples taken according to the flow velocity along a river course. Using a multi pollution source catchment as a case study we were interested in I) whether the new approach allows for the separation of point source pollution from diffuse pollution and natural background and for the identification of source-related fingerprints, II) whether the patterns can be generally characterized based on isotopologue signatures and homologue series, and III) what are representative compounds for these patterns.

2. Methods

2.1 Study site

The Holtemme River (Saxony-Anhalt, Germany) was chosen as a case study (SI 1.1, Figure A.1). From its source in the national park of the Harz Mountains to its confluence with the Bode River, it spans over 47 km passing through an area of intensive agriculture and two medium-sized towns with wastewater treatment plants (WWTP), which discharge into the river. The catchment of the first WWTP (WWTP I) covers an urban and rural area of 300 km² with about 50,000 inhabitants and an industrial input of about 15,000 population equivalents. The second WWTP (WWTP II) covers a mostly urban area of 143 km² with about 36,800 inhabitants connected to the WWTP. The input from industry contributes approximately 5,400 population equivalents. The WWTP effluents can be considered as the largest tributaries of the Holtemme River contributing about 34% and 23% to the river's discharge on the sampling day, respectively. Further technical details provided by the operators of the WWTPs are presented in the Supporting Information (SI 1.2, Table A.1).

2.2 Sampling

Grab samples of 500 mL each were collected along the river at 16 sites (SI 1.1, Figure A.1). Glass beakers used for sampling were cleaned with LC-grade acetone, methanol and distilled water and rinsed thrice with the water from the sampling site before the actual samples was collected. The name of each sampling site consists of the abbreviation "Holt" for Holtemme River and a number representing the river kilometer where the respective sample was collected. Aliquots of 1 mL of each sample were taken for chemical analysis. The time of sampling was adjusted to the river's flow velocity to sample the same water package at each sampling site. The flow velocity was modelled by hydrologists from UFZ based on a regression analysis considering actual discharge data from official gages and distances between sampling sites. Details on the sampling sites including information on physico-chemical properties of the samples are shown in SI 1.3, Table A.2.

2.3 Chemical analysis of samples

Samples were prepared for direct large volume injection (100 µL). For sample preparation, 10 µL of a 2 M ammonium formate buffer, 25 µL of methanol and 25 µL of an internal standard mixture containing 40 isotope-labelled compounds (40 ng mL^{-1}) were added to 1 mL of sample. Details on chemicals, reagents and isotope-labelled standards are provided in SI 2.1, Tables B.1 and B.2. Chemical analysis was performed on an UltiMate 3000 LC system (Thermo Scientific) coupled to a quadrupole-Orbitrap MS (Q Exactive Plus, Thermo Scientific) with a heated electrospray ionization (HESI) source. Chromatographic separation was performed on a Kinetex 2.6 µm EVO C18 (50x2.1 mm) column equipped with a pre-column (C18 EVO 5x2.1 mm) and an inline filter. The column temperature was 40°C. The LC solvent gradient is presented in SI 2.2, Table B.3. The nominal resolving power in the fullscan experiments was 140,000 (referenced to 200 m/z). For data-dependent (dd)-MS^2 experiments, an inclusion list of the selected ions of interest was provided for ionization modes. The nominal resolving power in dd-MS^2 experiments was 70,000 (referenced to m/z 200) in fullscan mode and 35,000 (referenced to m/z 200) in dd-MS^2 scans. Two collision energies (i.e., higher energy collision-induced dissociation (HCD)) were used for dd-MS^2 experiments, i.e. HCD 55 and HCD 35, in order to obtain diagnostic fragmentation patterns for small and large molecules. Further details on settings and parameters of the Q Exactive^{TM} Plus for fullscan experiments are presented in SI 2.2, Table B.4. At the beginning and at the end of each batch, calibration standards were run at four levels (1, 10, 100 and 1000 ng L^{-1}) to check for mass accuracy, intensity changes during the run and as a quality control during peak picking. Solvent blanks (95% H_2O/ 5% methanol) were analyzed at least after every sixth sample accounting for background contamination.

2.4 Data processing

Raw data from the LC-HRMS analysis were converted into .mzML format (centroid mode) by ProteoWizard v3.0.18265 (Chambers et al., 2012). Peak lists were generated using the software MZmine v2.32 (Pluskal et al., 2010). MZmine settings are given in SI 2.3, Table B.5. Repeatability of the chemical analysis and peak picking was checked by injecting replicates of selected samples. The peak lists were exported to Microsoft Excel® for blank correction according to Eq. 1. Signals below that threshold in the samples were removed. Furthermore, an intensity cut-off at peak heights below 5,000 in negative mode and 50,000 in positive mode was included to remove noise added by gap filling. For annotated target compounds, calibration standards were checked for logical increase in peak heights. If this was not observed, the annotation was removed. For manually added “marker” compounds, the intensity cutoff limit was not an exclusion criterion as they were manually integrated and were analyzed with a full calibration curve ranging from 1 to 1000 ng/L (Beckers et al., 2018).

Eq. 1: Calculation of intensity threshold (I_{thres})

$$I_{thres} = \mu(I_{Blk}) + 2*\sigma(I_{Blk})$$

$\mu(I_{Blk})$ = mean of peak intensities in blanks; $\sigma(I_{Blk})$ = standard deviation of peak intensities in blanks

Prior to cluster analysis, isotope peaks identified by the R package ‘nontarget’ v1.9 (Loos and Singer, 2017; R Core Team, 2017) were removed and the two cleaned peak lists obtained from positive and negative ionization mode were merged. As we observed several false positive adduct peaks identified by R ‘nontarget’ already for target compounds, adduct peaks were not filtered out in the peak list. Settings used in the R ‘nontarget’ package are described in SI 2.4, Table B.6. If a target compound was detected in both ionization modes, the one showing lower peak intensities was removed from the merged peak list. Some typically detected target compounds in the Holtemme River (Beckers et al., 2018) were missed during peak detection by MZmine due to poor peak shapes. All samples of this study were re-analyzed on a LC-

MS/MS system (QTrap 6500 MS/MS, ABSciex). The data was manually evaluated with the MultiQuant Software (Sciex). Details on the LC-MS/MS method are described elsewhere (Beckers et al., 2018). In total, seven compounds were added by target analysis. These compounds included the wastewater marker compounds acesulfame and saccharin (Buerge et al., 2009) as well as the pharmaceuticals pipamperone, diphenhydramine, ofloxacin, ciprofloxacin and metoprolol acid, which were detected as important wastewater compounds in a previous study (Beckers et al., 2018).

2.5 Cluster analysis

Cluster analysis was performed on componentized peak lists of the 16 water samples along the river. Prior to cluster analysis, the peak heights were normalized by intensity of the internal standard peaks matched by retention times to account for matrix effects. The normalized peak heights were scaled to unit variance according to Eq. 2 (i.e., z-score scaling). Scaling ensures that all variables spread over the same range, i.e. all variables have equal variances.

Eq. 2: Scaling to unit variance

$$z = \frac{x - \mu}{\sigma}$$

z = standard score, µ = mean, σ = standard deviation

Non detects (i.e., zeros) were not removed from the data set. Cluster analysis was performed in R using the R package ‘kml’ to unravel longitudinal clusters of peaks along a river course (Genolini et al., 2015; R Core Team, 2017). The cluster analysis in ‘kml’ was customized by using the distance function ‘diss.CORT’ from the R package ‘TSclust’ (Montero and Vilar, 2014). The ‘diss.CORT’ function compares trajectories based on the change in direction and rate at each spot (Montero and Vilar, 2014). Thus, this distance function fitted better to our spatial data set and helped to mitigate the assumption of spherical data by Euclidean distance used in the *k*-means algorithm. The R script for kml cluster analysis can be found in SI 2.5. The final number of clusters was chosen according to a consensus score of the incorporated quality criteria. The analysis was performed on the entire data set as well as on the resulting clusters to identify potential sub-patterns masked by main patterns. The ‘kml’ package provided probabilities of individuals belonging to the different clusters. However, these probabilities should be seen as indications rather than absolute values as they depend on normal distribution of each peak’s data which does not apply for single detects.

2.6 Characterization of pattern members

The R ‘nontarget’ package was used for the characterization of the peaks in the different patterns by identifying isotopologue signatures, adducts and homologue series (Loos and Singer, 2017). The analysis was based on the most representative samples of each pattern (section 3.2). The most representative sample of each pattern was the sample in which maximum intensities of peaks in the respective pattern were observed. In case maximum peak intensities were observed in more than one sample for a pattern, more samples were selected as representatives for the respective pattern. Information on isotopologues and homologues series was merged with information on cluster assignment and displayed in scatter plots (R packages ‘ggplot2’ (Wickham, 2016) and ‘ggpubr’ (Kassambara, 2018)).

2.7 Structure elucidation

Peaks were selected for structure elucidation by intensity. The top 5 to 10 high-intensity peaks were selected in representative samples of the different patterns and sub-patterns for identification. Chemical formulas were generated with the QualBrowser in XCalibur (Thermo Scientific). Calculated formulas were tested for plausibility regarding the isotopic pattern in the QualBrower and submitted for a probable formula query in ChemSpider (Royal Society of Chemistry, 2015) and CompTox (US EPA, 2019) database. Further information for structure elucidation was obtained by re-analyzing samples again in dd-MS^2, hydrogen-deuterium exchange (HDX) and pH-dependent chromatography experiments according to Muz et al. (2017). Fragment lists from respective MS^2 spectra were submitted to MetFrag v2.3 (web tool) (Ruttkies et al., 2016) to obtain candidate lists. HDX experiments provided information on exchangeable hydrogens in a molecule (Ruttkies et al., 2019), while pH-dependent chromatography supported the identification of probable pK_a values (Dann et al., 2016). Experimentally determined pK_a value ranges were compared to calculated acidic and basic

pK_a values by JChem for Office (Excel). Spectral similarity was checked for candidates in MassBank (Horai et al., 2010) and CFM-ID (Allen et al., 2014). Details on the complete workflow for structure elucidation are provided in SI 2.6. Finally, the level of identification for each structure was reported according to confidence levels introduced by Schymanski et al. (2015).

3. Results and Discussion

In the data set, 14,235 peaks were extracted in negative and 50,446 peaks in positive mode. After blank correction and removal of isotope peaks, the final list contained 23,485 peaks including 141 annotated target compounds. Since adducts were not removed, this list still included replicate peaks of the same compound exemplified for surfactants (section 3.3). Moreover, non-target compounds might be detected in both ionization modes. The stability in mass accuracy and peak intensity of calibration standards and the performance of replicate analyses is presented in SI 3.1 (Figures C.1-4) and 3.2 (Figure C.5). The effect of normalization of peak heights by internal standards was assessed in SI 3.3, Figure C.6.

3.1 Longitudinal peak patterns

Cluster analysis is an exploratory data analysis tool which reduced the data set to three main patterns. The applicability of the cluster analysis and the validity of the identified patterns were checked by running the analysis on a subset of quantified target compounds (SI 3.4.1, Figure C.7) and a manual cross-check of spatial courses of individual compounds with the spatial course of their associated main pattern as well as knowledge on potential sources at the Holtemme River. Furthermore, the effect of single detects on the cluster stability was tested underlining the robustness of the method (SI 3.4.2, Figure C.8). Due to the nature of partitioning cluster analysis, every variable (i.e., every peak) needs to be assigned to one of the clusters. This might be problematic for variables in the overlapping region of clusters. Thus, the main pattern did not reflect each peak's intensity course. In order to “clean up” the main pattern and identify finer structures and source-related fingerprints in the data set, a second sub-clustering of the main patterns was performed (section 3.1.2). The probabilities of peaks belonging to the assigned cluster and peak intensities in the samples are presented for target and prioritized unknown compounds in SI 3.6.1, Table C.1A-C.

3.1.1 Main peak patterns along the river course

According to the score of the quality criteria (SI 3.5.1, Figure C.9), three main patterns were unraveled in the river data set by cluster analysis (Figure 1). This distinction into three patterns would be missed by target screening alone (SI 3.4.1, Figure C.7).

The first pattern exhibited maximum intensity downstream of the two WWTPs with low or no signals in the headwater and will be referred as wastewater (WW) pattern below. This pattern included 9,811 peaks representing about 42% of the data set and most of the target compounds (n = 100, SI 3.6.1, Table C.1A). The target compounds belonged mostly to the group of pharmaceuticals, industrial compounds and pesticides. A second pattern showed a distinct and sudden increase in peak intensity at the last sampling site in the river, which represents the mixing zone with the Bode River. This pattern was called Bode River (BR) pattern and contained 7,776 peaks, i.e., 33% of all peaks. As there are no major tributaries in the Holtemme River between sampling sites 40 and 42, those peaks likely originated from the Bode River. Target compounds of BR pattern included mostly industrial compounds and industrially used biocides (i.e., isothiazolinones, SI 3.6.1, Table C.1B). A third cluster with 5,910 peaks included about 25% of all peaks. It showed higher intensities in the headwater regions with a decrease downstream of the WWTP effluent sites potentially due to dilution and was termed diffuse and random input (DRI) pattern (section 2.1). Thus, the peaks of this pattern were not associated with WWTP effluents. The few target compounds that were assigned to this pattern were mainly pesticide metabolites as well as the legacy pesticide atrazine and artificial sweeteners (SI 3.6 1, Table C.1C). The presence of the artificial sweeteners cyclamate and saccharin suggested the input of untreated wastewater as they are largely degraded during the wastewater treatment process (Buerge et al., 2009). A previous study identified rain sewers as a small point source for untreated wastewater and random spills in this headwater region (Beckers et al., 2018). The input was observed even under dry weather conditions due to faulty or illicit connections in the sewer network. The occurrence of pesticides and their metabolites might also be explained by the input via rain sewers and other drainages as well as from infiltrating groundwater (Kolpin et al., 2000; Reemtsma et al., 2013). During this sampling campaign, the total discharge was solely produced by base flow generated by groundwater as well as by contributions from tributaries (including WWTP effluents). This led to a river discharge rate of 0.34 m^3 s^{-1} well below the mean annual discharge rate of 1.55 m^3 s^{-1} and consequently comparably lower dilution along the river course (LHW, 2019; Müller et al., 2018). The DRI pattern, moreover, contained many unidentified peaks which showed consistently high intensities over the whole river course. They likely represented natural background compounds. Thus, this pattern summarized both diffuse and random input of organic compounds.

3.1.2 Sub-patterns and source-related fingerprints

Based on the score of the quality criteria (SI 3.5.2, Figure C.10), cluster analysis of the WW pattern revealed four sub-patterns (Figure 2A). The majority of peaks were assigned to sub-pattern WW1, which represented peaks associated with both WWTPs. Sub-patterns WW2 and WW3 represented peaks which were more associated with either one of the WWTPs. This included peaks which solely or mainly originated from one of the WWTPs. Specific input from WWTP I included fungicides, the antibiotics roxithromycin and azithromycin, as well as coumarin derivatives (SI 3.6.1, Table C.1A). The latter were previously identified as the main drivers for anti-androgenic activity at this sampling site (Muschket et al., 2018). Several pharmaceuticals (e.g. acetaminophen and ketoprofen) were associated to a larger extent with WWTP II even though they were emitted from both WWTPs. The relatively higher input from WWTP II might be explained by shorter hydraulic residence times and thus less efficient treatment of WWTP II (SI 1.2, Table A.1). The sub-patterns WW1, WW2 and WW3 clearly assigned peaks to their sources. Thus, they may be seen as source-related fingerprints, whereas the WW1 sub-pattern is a fingerprint for common wastewater compounds with lower variability and the WW2 and WW3 sub-patterns are fingerprints for wastewater-related compounds with more variable discharges or specific sources in the WWTPs' catchments. Many of the compounds in these patterns were among frequently detected compounds at European WWTPs including the sweetener acesulfame, pharmaceuticals (e.g. carbamazepine, citalopram, diclofenac and sulfamethoxazole), pesticides (e.g. MCPA) and corrosion inhibitors such as benzotriazoles (Loos et al., 2013; Munz et al., 2017). Sub-pattern WW4 contained compounds which were predominant at the first sampling site (Figure 2A), and showed only small intensity increases downstream of both WWTPs. Already in the headwater region, there is some anthropogenic influence due to a small battery factory and a hotel upstream of sampling site Holt3. Both treat their wastewater in septic tanks and discharge rain water to the Holtemme River.

Likewise, sub-clustering of the BR pattern (Figure 2B and SI 3.5.2, Figure C.11) revealed sub-patterns of peaks that also occurred at the sites downstream of the WWTPs (i.e., BR2, BR4). However, the sampling site with highest peak intensities was still the river mouth for all sub-patterns (i.e., BR1-4).

Sub-clustering of the DRI pattern indicated a few sampling sites with elevated intensities in the urban regions (i.e., site Holt9, Holt11, Holt15 and Holt26) (Figure 2C and SI 3.5.2, Figure C.12). The sites are believed to reflect inputs from small point sources such as rain sewers. The high variation of some peaks among sampling sites is likely due to very random and inconsistent inputs from these sources directly reflecting activities in their catchment (Beckers et al., 2018). Thus, the sub-patterns of the DRI pattern may greatly vary with time. Still, the cluster analysis, especially with detailed sub-clustering, has the potential to detect even smaller point sources and is also robust enough, so that the patterns are not disturbed by single detects (SI 3.4.2, Figure C.8).

The applicability of the cluster analysis was demonstrated using data of a one-time sampling campaign. However, the stability of these patterns, sub-patterns and source-related fingerprints should be tested for temporal variations due to changing flow conditions (i.e., effects of dilution) and seasonal influences (Beckers et al., 2018) (e.g., pesticide applications in spring or changes in industrial production) in future studies. Especially, the origin of peaks in DRI pattern may become more defined and background may be better separated from input of small point sources by repeated sampling.

3.2 Characterization of pattern components

The main patterns were investigated for characteristic mass-to-charge ratio (m/z) and retention time (RT) distributions as well as for the abundance of peaks with specific isotopologue signatures and homologue series. Halogenated compounds are typically of anthropogenic origin and are often toxic and persistent. Sulfur-containing compounds especially in combination with homologue series indicate the presence of surfactants. The characterization was based on representative samples of each of the patterns. For the WW pattern, this included samples Holt17 and Holt31 corresponding to the sampling sites downstream of each of the WWTPs. Samples Holt9 and Holt26 were analyzed as representatives for the DRI pattern and sample Holt42 for the BR pattern.

By plotting m/z values against RT of the pattern components, distinct differences between the DRI pattern and the two other patterns (WW and BR) were identified (Figure 3). The DRI pattern contained a lot of peaks eluting at or close to the column dead time with high intensities (i.e., RT < 1 min). A lot of potentially halogenated and sulfur-containing compounds were among these peaks (Figure 3C). For a better identification of these compounds, an improved chromatographic separation of highly hydrophilic compounds on a more polar stationary phase would be required. This exemplifies the limit of each data set’s explanatory power based on the analytical methods used.

Also the WW and BR patterns included such early eluting peaks with this isotopologue signature. However, in these patterns more halogenated and sulfur-containing compounds were detected with higher retention times (Figures 3A, B).

The number of peaks assigned as part of a homologue series was evaluated per pattern. The number of homologue peaks increased with the effluent from the two WWTPs (n = 2282) and almost doubled with the confluence with the Bode River. In combination with the potentially high number of sulfur-containing compounds, these peaks might indicate the presence of surfactants as identified in wastewater by previous studies (e.g. Alygizakis et al., 2019; Gago-Ferrero et al., 2015; Peter et al., 2018; Schymanski et al., 2014). Dissolved organic matter (DOM) originating from wastewater has a distinctly high content of sulfur-containing species in comparison to DOM from pristine waters (Greenwood et al., 2012). The investigation of changes in DOM homologue series during wastewater treatment showed that especially compounds with CH_2-series are not readily degradable during treatment (Verkh et al., 2018). Follow-up studies in the Bode River should reveal where this high contribution of compounds in homologue series (potentially surfactants) originate from. The presence of these characteristic peaks in the WW and BR pattern supported the urban and industrial contributions indicated by target compounds (section 3.2). Some of these ions of interests were identified (section 3.3).

A consistently low number (n = 464) of peaks in a homologue series were related to the DRI pattern. Most of these homologue series (>90%) showed a mass increment of 14 m/z representing a CH_2 group. This group is commonly seen in anthropogenic homologue series but was also discovered in homologue series of natural compounds such as humic and fulvic acids (Stenson et al., 2002). Thus, the homologues series in this pattern might reflect natural background. Our results suggested that natural compounds make up a considerable part in the chemical mixtures detected along the river. Further analytical efforts are necessary to study these compounds, especially because they may play a role in the overall ecosystem health (Pignatello and Xing, 1996) and in water treatment (Neale et al., 2012).

3.3 Identification of ions of interest

In addition to target compounds, ions of interest were identified to different levels of confidence (Schymanski et al., 2015). The identified compounds supported pattern and source interpretation as well as are previously unknown representatives for these patterns. Spectra of confirmed substances were uploaded to MassBank database (SI 3.6.1, Table C.2).

The identification focused on high intensity peaks in the common wastewater WW pattern (WW1) as well as the two WWTP-specific patterns (WW2 and WW3) and the DRI and BR pattern. The results are summarized in Tables 1 and C.1A-C (SI 3.6.1). Based on determined molecular formulas, plausible candidate structures were selected using MS^2 spectra, pK_a values (indicated by pH-dependent retention times) and the number of exchangeable hydrogens. Finally, commercial relevance was considered as an indication to occur in a wastewater-impacted river. The MS^2 spectra of the compounds in the original sample and the respective reference standards are presented the SI, section 3.6.2.

In the WW sub-patterns, several pharmaceuticals (i.e., lamotrigine, methocarbamol, irbesartan and olmesartan) and some pharmaceutical transformation products (i.e., gabapentin-lactam and valsartan acid) were confirmed by reference standards. The peak of lamotrigine was also correctly identified by the R ‘nontarget’ package as ion with chlorine isotopes further supporting the confirmation based on the mass spectra of the reference standard. Lamotrigine was assigned to the WW3 sub-pattern and showed a distinct peak at WWTP I (SI 3.6.1, Table C.1A). The intensity was reduced to 30% of the original peak over the course of the river. WWTP I had a specific input of other pharmaceuticals such as the antidepressant pipamperone (SI 3.6.1, Table C.1A). This might be explained by the presence of a pharmaceutical manufacturer connected to the WWTP as there is no difference in hospital size or specialization. Lamotrigine is a ubiquitous pharmaceutical previously detected, e.g., in the Rhine River, in Swiss WWTP effluents and a US estuary (Carpenter and Helbling, 2018; Munz et al., 2017; Muz et al., 2017; Ruff et al., 2015). The other identified pharmaceuticals showed similar intensities at both WWTP effluent sites (SI 3.6.1, Table C.1A). Methocarbamol is a muscle relaxant and irbesartan, olmesartan and valsartan (the latter detected as its transformation product valsartan acid) are used for treatment of hypertension. The high intensity in this study and detections in other studies can be explained by high consumption volumes of these widely used pharmaceuticals (Carpenter and Helbling, 2018; Munz et al., 2017). Irbesartan was detected in 100% of WWTP effluents in EU-wide study (Loos et al., 2013). Gabapentin-lactam is a human metabolite of the anticonvulsant gabapentin and is more stable under environmental conditions than the parent compound (Henning et al., 2018). Gabapentin was part of our target list and has been assigned to the WW2 sub-pattern showing a 50% higher intensity in the effluent of WWTP II than in the effluent of WWTP I, while the intensity of gabapentin-lactam was similar in both WWTP effluents. Thus, the lower gabapentin to gabapentin-lactam ratio in the effluent of WWTP I might be explained by a more efficient treatment in WWTP I.

Furthermore, 4-methyl-7-ethylaminocoumarin was identified by a reference standard as specific to WWTP I (SI 3.6.1, Table C.1A). Coumarin derivatives were identified as ecotoxicologically relevant compounds specifically emitted from this WWTP (Muschket et al., 2018). 4-Methyl-7-ethylaminocoumarin is the transformation product of 4-methyl-7-diethylaminocoumarin. Like the parent compound, it has an anti-androgenic effect. However it is less potent than its parent compound (Muschket et al., 2018). The sulfophenyl carboxylic acids (SPC) C6-SPC and C7-SPC were tentatively identified at confidence level 2b. Their identification matched the isotopologue and homologue patterns revealed in section 3.2 as representatives of a sulfur-containing homologue series. SPCs are main degradation products of linear alkylbenzene sulfonates (LAS) and have been detected in the aquatic environment and WWTP effluents (Lara-Martín et al., 2011). No records were available in MassBank spectral library for C6-SPC or C7-SPC. However, diagnostic fragments (183.0123 m/z and 197.0279 m/z) and ionization were matched to previous studies (SI 3.6.3, Figure C.34) (Gonsior et al., 2011; Lara-Martín et al., 2011). Moreover, the mass increment 14 m/z suggested a CH_2 - homologue series.

Seven out of 21 ions of interest were identified at level 4 in the WW pattern. By application of the pH-dependent LC retention method (Dann et al., 2016), we were able to separate two of these peaks with the same molecular formula with the m/z 274.2010 (SI 3.6.4, Figure C.35). Even though the two compounds could not be fully identified, one peak must belong to a carboxylic acid and the other one to a compound with a basic functional group with a basic pK_a between 2.6 and 6.4, e.g. primary, secondary, tertiary aromatic amines or triazine derivates.

The limits of proper pK_a calculation were exemplified for irbesartan, olmesartan and 4-methyl-7-ethylaminocoumarin. Here, the calculated pK_a did not correspond to the structures suggested by the pH-dependent LC retention (Table 1). Thus, care that has to be taken in the evaluation of calculated pK_a values. Only for two ions in the WW pattern, no unequivocal molecular formula could be determined.

The BR pattern was dominated by peaks which were predominantly showing ammonium adducts $[M+NH_4]^+$ but also the $[M+H]^+$ and $[M+Na]^+$ adducts. Five of these peaks were identified (level 1) as polyethylene glycols (PEGs) with the general molecular formula $C_{2n}H_{4n+2}O_{n+1}$. They are usually detected as these adducts (Alygizakis et al., 2019; Lara-Martín et al., 2011; Peter et al., 2018). PEGs have a broad field of application in industrial and household products and may enter via rain sewers during surface runoff (Peter et al., 2018) as well as via treated (Schymanski et al., 2014) and untreated (Gago-Ferrero et al., 2015) wastewater input. PEGs were also observed at other sampling sites at the Holtemme River, e.g. in urban regions and at the weir (SI 3.6.1, Table C.1B), but not as dominant as at the confluence with the Bode River. Moreover, the intensities of PEGs in the river samples dropped downstream of the WWTP effluents suggesting dilution by treated wastewater and a removal of PEGs by WWTPs in agreement with other studies (Freeling et al., 2019). The results coincided with the overall patterns revealed by isotopologue signatures and homologue series detection (section 3.2) which suggested a specific contribution of Bode River to the Holtemme River, e.g. by untreated wastewater or a specific point source. Moreover, other surfactants and industrial compounds were identified at this spot including triacetin, diethylene glycol monobutyl ether and azelaic acid (level 1). Triacetin was identified in surface waters and groundwater (Schwarzbauer and Ricking, 2010; Sorensen et al., 2015) and was previously linked to specific industrial effluents and proposed as an indicator for the production of paper and inks (Botalova et al., 2011). However, triacetin has a broad range of other industrial applications as a food additive, plasticizer and in pharmaceutical products suggesting a variety of sources. Azelaic acid was intensively studied in and associated with airborne organic particulate matter as a photochemical oxidation product of unsaturated fatty acids (e.g. Hyder et al., 2012; Wang et al., 2002). In our study, azelaic acid was only detected at the sampling site at the river mouth (SI 3.6.1, Table C.1B) which contradicts an input from atmospheric deposition. However, it is also used in personal care products (DrugBank, 2019), which might explain its local occurrence in the Holtemme River. Again, these specifically high occurrences in the BR pattern call for further in-depth investigations on sources in the Bode River and dynamics at this particular sampling site.

In the DRI pattern, five out of eight ions of interest could be identified to level 1 as constituents of cocamidopropylbetaine as well as n-lauroylethanolamine and triethylene glycol monomethyl ether. Cocamidopropylbetaine and n-lauroylethanolamine are surfactants mainly used in personal care products (ECHA, 2019a; ECHA, 2019b). These compounds were not related to the input of treated wastewater, as they are likely eliminated in WWTPs. They showed specifically high intensities in the urban area upstream of WWTP I (SI 3.6.1, Table C.1C) suggesting input of untreated wastewater via rain sewers (Beckers et al., 2018). Furthermore, they were clustered together with the target compound lauryl diethanolamide in the DRI pattern. In absence of a reference standard, lauryl sulfate was tentatively identified at level 2a (SI 3.6.5, Figure C.36). It was previously identified in untreated wastewater (Alygizakis et al., 2019). Triethylene glycol monomethyl ether and lauryl sulfate were related to point source pollution at a sampling site close to a rain sewer and at sampling site Holt36, which is at a weir (Figure A.1 and SI 3.6.1, Table C.1C). The site-specific detection of these compounds might suggest an input of raw wastewater and surface runoff via rain sewers, their quick removal from the water phase and a remobilization in the weir area from deposited sediments, respectively.

Conclusions

The analytical power of NTS is continuously increasing and the volume of NTS data produced is increasing exponentially. However, the availability of concepts and tools to structure and exploit these huge data sets is lagging behind. In the present study, we demonstrated how innovative analytical workflows integrating multivariate statistical approaches emerging from different areas of research help to identify pollution patterns and source-related fingerprints in highly complex pollutant mixtures. To our knowledge, this is the first study to apply a longitudinal cluster analysis on a non-target data set, which efficiently separated peaks originating from different sources. The identified patterns suggested a high abundance of natural background in environmental chemical mixtures which could be separated from clear anthropogenic inputs and require further investigation. The cluster analysis was robust enough to identify main pollution patterns despite many single detects in the data set. By means of isotopologue fingerprints and homologue series as well as detected target and identified non-target compounds, the patterns were related to inputs from WWTPs, specific pollutants at the river's mouth and point pollution of untreated wastewater. The proposed workflow is extendable to and should be tested in other settings (e.g. larger rivers, river stretches) to quickly identify pollution hotspots or pathways or identifying temporal dynamics. The exchange of identified patterns in environmental mixtures and source-related fingerprints is encouraged among researchers to test their validity in other water bodies and point sources and allow for their complementation. The approach presented here is an important building block in the ongoing attempts to understand, monitor, prioritize and manage complex environmental mixtures (Brack et al., 2018).

Figure legends

Figure 1: Main patterns (wastewater (WW), Bode River (BR), and diffuse and random (DRI) pattern) identified by cluster analysis on all peaks detected by non-target screening. Colored lines represent clusters identified by cluster analysis. Gray background represents longitudinal course across all sampling sites of intensities of individual peaks detected in LC-HRMS data set. Peak intensity was scaled to unit variance. The number of the sampling sites represents the river kilometer. Box above the plot indicates percentage of peaks of the data set assigned to a respective cluster.

Figure 2: Sub-patterns of main patterns (A) wastewater (WW), (B) Bode River (BR) and (C) diffuse and random input (DRI) identified by cluster analysis on all peaks included in the respective main pattern. Colored lines represent clusters identified by cluster analysis. Gray background represents longitudinal course across all sampling sites of intensities of individual peaks detected in LC-HRMS data set. Peak intensity was scaled to unit variance. The number of the sampling sites represents the river kilometer. Box above the plot indicates percentage of peaks of the data set assigned to a respective cluster.

Figure 3: Scatter plots of retention time [min] vs. mass-to-charge ratio of all peaks in the three main patterns (A) wastewater (WW), (B) Bode River (BR) and (C) diffuse and random input (DRI). Colored points represent isotopologues assigned to isotope peaks. Point size reflects the intensity of each peak.

Tables

Table 1: Results of structure elucidation for ions of interest

[#]level of confidence according to Schymanski et al. (2015), nr = no results obtained from experiments, nc= not calculable by JChem for Office

Acknowledgements

This study was supported by SOLUTIONS project funded by the European Union Seventh Framework Programme (FP7-ENV-2013-two-stage Collaborative project) under grant agreement number 603437. The authors thank the WWTP operators for providing effluent samples, discharge data and information on the WWTP catchments. The authors further acknowledge Christin Müller (UFZ) for providing the map of the study site and Andreas Musolff (UFZ) for calculating flow velocities of the river. A free academic license of JChem for Office (Excel) was used for structure based property calculation, JChem for Office 6.2.1, 2014, ChemAxon (http://www.chemaxon.com). The QExactive Plus LC-HRMS used is part of the major infrastructure initiative CITEPro (Chemicals in the Terrestrial Environment Profiler) funded by the Helmholtz Association.

References

Allen F, Pon A, Wilson M, Greiner R, Wishart D. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. Nucleic acids research 2014; 42: W94-W99.

Altenburger R, Ait-Aissa S, Antczak P, Backhaus T, Barceló D, Seiler T-B, et al. Future water quality monitoring — Adapting tools to deal with mixtures of pollutants in water resource management. Science of The Total Environment 2015; 512–513: 540-551.

Alygizakis NA, Gago-Ferrero P, Hollender J, Thomaidis NS. Untargeted time-pattern analysis of LC-HRMS data to detect spills and compounds with high fluctuation in influent wastewater. Journal of Hazardous Materials 2019; 361: 19-29.

Beckers L-M, Busch W, Krauss M, Schulze T, Brack W. Characterization and risk assessment of seasonal and weather dynamics in organic pollutant mixtures from discharge of a separate sewer system. Water Research 2018; 135: 122-133.

Botalova O, Schwarzbauer J, Sandouk Na. Identification and chemical characterization of specific organic indicators in the effluents from chemical production sites. Water Research 2011; 45: 3653-3664.

Brack W, Escher BI, Müller E, Schmitt-Jansen M, Schulze T, Slobodnik J, et al. Towards a holistic and solution-oriented monitoring of chemical status of European water bodies: how to support the EU strategy for a non-toxic environment? Environmental Sciences Europe 2018; 30: 33.

Brack W, Hollender J, de Alda ML, Müller C, Schulze T, Schymanski E, et al. High-resolution mass spectrometry to complement monitoring and track emerging chemicals and pollution trends in European water resources. Environmental Sciences Europe 2019; 31: 62.

Brunner AM, Bertelkamp C, Dingemans MML, Kolkman A, Wols B, Harmsen D, et al. Integration of target analyses, non-target screening and effect-based monitoring to assess OMP related water quality changes in drinking water treatment. Science of The Total Environment 2020; 705: 135779.

Buerge IJ, Buser H-R, Kahle M, Müller MD, Poiger T. Ubiquitous Occurrence of the Artificial Sweetener Acesulfame in the Aquatic Environment: An Ideal Chemical Marker of Domestic Wastewater in Groundwater. Environmental Science & Technology 2009; 43: 4381-4385.

Carpenter CMG, Helbling DE. Widespread Micropollutant Monitoring in the Hudson River Estuary Reveals Spatiotemporal Micropollutant Clusters and Their Sources. Environmental Science & Technology 2018; 52: 6187-6196.

Carpenter CMG, Wong LYJ, Johnson CA, Helbling DE. Fall Creek Monitoring Station: Highly Resolved Temporal Sampling to Prioritize the Identification of Nontarget Micropollutants in a Small Stream. Environmental Science & Technology 2019; 53: 77-87.

Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nature Biotechnology 2012; 30: 918.

Chiaia-Hernández AC, Günthardt BF, Frey MP, Hollender J. Unravelling Contaminants in the Anthropocene Using Statistical Analysis of Liquid Chromatography–High-Resolution Mass Spectrometry Nontarget Screening Data Recorded in Lake Sediments. Environmental Science & Technology 2017; 51: 12547-12556.

Dann JP, Brack W, Krauss M In *pH-Dependent Retention Time Measurement for the Identification of Unknown Substances with LC-HRMS*, NonTarget2016, Ascona, Switzerland, 30.05.2016, 2016; Ascona, Switzerland, 2016.

DrugBank. Azelaic acid. https://www.drugbank.ca/drugs/DB00548 (accessed: 24.04.2019)

ECHA. European Chemicals Agency.1-Propanaminium, 3-amino-N-(carboxymethyl)-N,N-dimethyl-, N-coco acyl derivs., hydroxides, inner salts.

https://echa.europa.eu/de/substance-information/-/substanceinfo/100.057.308 (accessed: 23.04.2019)

ECHA. European Chemicals Agency.N-(2-hydroxyethyl)dodecanamide. https://echa.europa.eu/de/substance-information/-/substanceinfo/100.005.055 (accessed: 23.04.2019)

Freeling F, Alygizakis NA, von der Ohe PC, Slobodnik J, Oswald P, Aalizadeh R, et al. Occurrence and potential environmental risk of surfactants and their transformation products discharged by wastewater treatment plants. Sci Total Environ 2019; 681: 475-487.

Gago-Ferrero P, Schymanski EL, Bletsou AA, Aalizadeh R, Hollender J, Thomaidis NS. Extended Suspect and Non-Target Strategies to Characterize Emerging Polar Organic Contaminants in Raw Wastewater with LC-HRMS/MS. Environmental Science & Technology 2015; 49: 12333-12341.

Genolini C, Alacoque X, Sentenac M, Arnaud C. kml and kml3d: R Packages to Cluster Longitudinal Data. Journal of Statistical Software 2015; 65: 34.

Gonsior M, Zwartjes M, Cooper WJ, Song W, Ishida KP, Tseng LY, et al. Molecular characterization of effluent organic matter identified by ultrahigh resolution mass spectrometry. Water Research 2011; 45: 2943-2953.

Greenwood PF, Berwick LJ, Croué JP. Molecular characterisation of the dissolved organic matter of wastewater effluents by MSSV pyrolysis GC–MS and search for source markers. Chemosphere 2012; 87: 504-512.

Henning N, Kunkel U, Wick A, Ternes TA. Biotransformation of gabapentin in surface water matrices under different redox conditions and the occurrence of one major TP in the aquatic environment. Water Research 2018; 137: 290-300.

Hollender J, Schymanski EL, Singer HP, Ferguson PL. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? Environmental Science & Technology 2017; 51: 11505-11512.

Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: a public repository for sharing mass spectral data for life sciences. Journal of Mass Spectrometry 2010; 45: 703-714.

Hug C, Ulrich N, Schulze T, Brack W, Krauss M. Identification of novel micropollutants in wastewater by a combination of suspect and nontarget screening. Environmental Pollution 2014; 184: 25-32.

Hyder M, Genberg J, Sandahl M, Swietlicki E, Jönsson JÅ. Yearly trend of dicarboxylic acids in organic aerosols from south of Sweden and source attribution. Atmospheric Environment 2012; 57: 197-204.

Kassambara A *ggpubr: 'ggplot2' Based Publication Ready Plots*, v.0.2; 2018.

Kolpin DW, Thurman EM, Linhart SM. Finding minimal herbicide concentrations in ground water? Try looking for their degradates. Science of The Total Environment 2000; 248: 115-122.

Lara-Martín PA, González-Mazo E, Brownawell BJ. Multi-residue method for the analysis of synthetic surfactants and their degradation metabolites in aquatic systems by liquid chromatography–time-of-flight-mass spectrometry. Journal of Chromatography A 2011; 1218: 4799-4807.

LHW. State Office for Flood Protection and Water Management Saxony-Anhalt.Datenportal Gewässerkundlicher Landesdienst Sachsen-Anhalt (Database State Waterways Service). http://gldweb.dhi-wasy.com/gld-portal/ (accessed: 28.01.2019)

Loos M, Singer H. Nontargeted homologue series extraction from hyphenated high resolution mass spectrometry data. Journal of Cheminformatics 2017; 9: 12.

Loos R, Carvalho R, António DC, Comero S, Locoro G, Tavazzi S, et al. EU-wide monitoring survey on emerging polar organic contaminants in wastewater treatment plant effluents. Water Research 2013; 47: 6475-6487.

Montero P, Vilar JA. TSclust: An R Package for Time Series Clustering. Journal of Statistical Software 2014; 62: 43.

Müller A, Schulz W, Ruck WKL, Weber WH. A new approach to data evaluation in the non-target screening of organic trace substances in water analysis. Chemosphere 2011; 85: 1211-1219.

Müller C, Musolff A, Strachauer U, Brauns M, Tarasova L, Merz R, et al. Tomography of anthropogenic nitrate contribution along a mesoscale river. Science of The Total Environment 2018; 615: 773-783.

Munz NA, Burdon FJ, de Zwart D, Junghans M, Melo L, Reyes M, et al. Pesticides drive risk of micropollutants in wastewater-impacted streams during low flow conditions. Water Research 2017; 110: 366-377.

Muschket M, Di Paolo C, Tindall AJ, Touak G, Phan A, Krauss M, et al. Identification of Unknown Antiandrogenic Compounds in Surface Waters by Effect-Directed Analysis (EDA) Using a Parallel Fractionation Approach. Environmental Science & Technology 2018; 52: 288-297.

Muz M, Krauss M, Kutsarova S, Schulze T, Brack W. Mutagenicity in Surface Waters: Synergistic Effects of Carboline Alkaloids and Aromatic Amines. Environmental Science & Technology 2017; 51: 1830-1839.

Neale PA, Antony A, Bartkow M, Farre M, Heitz A, Kristiana I, et al. Bioanalytical assessment of the formation of disinfection by-products in a drinking water treatment plant. Environmental Science & Technology 2012; 46: 10317–10325.

Nürenberg G, Schulz M, Kunkel U, Ternes TA. Development and validation of a generic nontarget method based on liquid chromatography – high resolution mass spectrometry analysis for the evaluation of different wastewater treatment options. Journal of Chromatography A 2015; 1426: 77-90.

Peter KT, Tian Z, Wu C, Lin P, White S, Du B, et al. Using High-Resolution Mass Spectrometry to Identify Organic Contaminants Linked to Urban Stormwater Mortality Syndrome in Coho Salmon. Environmental Science & Technology 2018; 52: 10317-10327.

Pignatello JJ, Xing B. Mechanisms of Slow Sorption of Organic Chemicals to Natural Particles. Environmental Science & Technology 1996; 30: 1-11.

Pluskal T, Castillo S, Villar-Briones A, Orešič M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 2010; 11: 395-395.

R Core Team *R: A language and environment for statistical computing*, v.3.4.3; R Foundation for Statistical Computing: Vienna, Austria, 2017.

Reemtsma T, Alder L, Banasiak U. Emerging pesticide metabolites in groundwater and surface water as determined by the application of a multimethod for 150 pesticide metabolites. Water Research 2013; 47: 5535-5545.

Richardson SD, Kimura SY. Emerging environmental contaminants: Challenges facing our next generation and potential engineering solutions. Environmental Technology & Innovation 2017; 8: 40-56.

Royal Society of Chemistry. ChemSpider, 2015.

Ruff M, Mueller MS, Loos M, Singer HP. Quantitative target and systematic non-target analysis of polar organic micro-pollutants along the river Rhine using high-resolution mass-spectrometry – Identification of unknown sources and compounds. Water Research 2015; 87: 145-154.

Ruttkies C, Schymanski EL, Strehmel N, Hollender J, Neumann S, Williams AJ, et al. Supporting non-target identification by adding hydrogen deuterium exchange MS/MS capabilities to MetFrag. Analytical and Bioanalytical Chemistry 2019.

Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. Journal of cheminformatics 2016; 8: 3-3.

Schollée JE, Schymanski EL, Avak SE, Loos M, Hollender J. Prioritizing Unknown Transformation Products from Biologically-Treated Wastewater Using High-Resolution Mass Spectrometry, Multivariate Statistics, and Metabolic Logic. Analytical Chemistry 2015; 87: 12121-12129.

Schwarzbauer J, Ricking M. Non-target screening analysis of river water as compound-related base for monitoring measures. Environmental Science and Pollution Research 2010; 17: 934-947.

Schymanski EL, Singer HP, Longrée P, Loos M, Ruff M, Stravs MA, et al. Strategies to Characterize Polar Organic Contamination in Wastewater: Exploring the Capability of High Resolution Mass Spectrometry. Environmental Science & Technology 2014; 48: 1811-1818.

Schymanski EL, Singer HP, Slobodnik J, Ipolyi IM, Oswald P, Krauss M, et al. Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. Analytical and Bioanalytical Chemistry 2015; 407: 6237-6255.

Sorensen JPR, Lapworth DJ, Nkhuwa DCW, Stuart ME, Gooddy DC, Bell RA, et al. Emerging contaminants in urban groundwater sources in Africa. Water Research 2015; 72: 51-63.

Stenson AC, Landing WM, Marshall AG, Cooper WT. Ionization and Fragmentation of Humic Substances in Electrospray Ionization Fourier Transform-Ion Cyclotron Resonance Mass Spectrometry. Analytical Chemistry 2002; 74: 4397-4409.

US EPA. United States Environmental Protection Agency.CompTox Chemicals Dashboard. https://comptox.epa.gov/dashboard (accessed: 2019)

Verkh Y, Rozman M, Petrovic M. A non-targeted high-resolution mass spectrometry data analysis of dissolved organic matter in wastewater treatment. Chemosphere 2018; 200: 397-404.

Wang G, Niu S, Liu C, Wang L. Identification of dicarboxylic acids and aldehydes of PM10 and PM2.5 aerosols in Nanjing, China. Atmospheric Environment 2002; 36: 1941-1950.

Wickham H *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York: 2016.

Zheng W, Wang X, Tian D, Zhang H, Tian W, Andersen ME, et al. Pollution Trees: Identifying Similarities among Complex Pollutant Mixtures in Water and Correlating Them to Mutagenicity. Environmental Science & Technology 2012; 46: 7274-7282.