Environmental Sciences Europe

RESEARCH Open Access

Prioritising site-specific micropollutants in surface water from LC-HRMS non-target screening data using a rarity score

Martin Krauss[1*], Christine Hug[1,2], Robert Bloch[1,2], Tobias Schulze[1] and Werner Brack[1,2]

Abstract

Background: Current compound prioritisation and monitoring approaches within Europe focus mainly on widely occurring priority and river basin specific pollutants but may overlook site-specific contamination from local emission sources. Thus, we propose a robust and semiautomated approach for the identification of site-specific chemicals and a prioritisation of water bodies with specific contamination based on non-target screening data from liquid chromatography coupled to high-resolution mass spectrometry.

Results: For prioritisation of site-specific contaminants, we calculated rarity scores for all peaks occurring in a set of 31 surface water samples, which combine the maximum signal intensity of a peak in a dataset with its frequency of occurrence in that dataset in one single number. These were a robust measure without the need to address the problems of missing data in more sophisticated multivariate statistical methods. For our dataset, site-specific compounds were defined for rarity scores > 1000, and the studied 31 sites showed a huge difference in the number of such peaks (0–91 in positive and 0–48 in negative ion mode). Together with isotopologue detection, the evaluation of mass defects and the occurrence of homologue series, which all could be obtained from automated data processing, a more detailed characterisation of these site-specific contaminations was possible. For three selected sites with a high number of site-specific peaks, novel or unexpected compounds could be identified, which stem from specific usage or (former) industrial production upstream of these sites.

Conclusions and outlook: The proposed approach allows for a rapid screening of large non-target screening datasets for site-specific contaminants, the prioritisation of sites with such a specific contamination and the subsequent identification of these compounds. Thus, the risk of overlooking possibly hazardous chemicals (including unknowns) which are not covered in conventional monitoring and prioritisation schemes is reduced.

Keywords: Non-target screening, Micropollutant prioritisation, LC-HRMS, High-resolution mass spectrometry, Surface water

Background

The occurrence of organic micropollutants in surface water has raised concerns due to their harmful effects on aquatic organisms and the possible entry into human water supply [51]. Over the last two decades, the compound spectrum analysed has steadily increased, although the number of compounds included in routine monitoring programs is still rather low compared to those compounds known to be present in environmental samples [5].

The European Water Framework Directive (WFD; European Union 2000) is currently the main basis for surface water monitoring activities in European countries. It has a specific focus on European scale Priority Substances, which are used to define the Chemical Status of a water body, together with varying lists of river basin specific pollutants (RBSPs). Thus, monitoring efforts are typically biased towards compounds relevant for larger-scale

*Correspondence: martin.krauss@ufz.de
[1] Department of Effect-Directed Analysis, UFZ–Helmholtz Centre for Environmental Research, Permoserstr.15, 04318 Leipzig, Germany
Full list of author information is available at the end of the article

catchments. However, also site-specific contamination might substantially contribute to the likelihood that surface water bodies fail to meet environmental quality objectives and thus a further assessment is required (WFD, Annex II, Sect. 1.5, Assessment of Impact).

Water pollution due to household effluents treated in and emitted via municipal wastewater treatment plants (WWTPs) is expected to be composed of a more or less consistent, typical set of substances from major human activities including laundry care, home care, health care, personal care and food [10]. Concentrations are mainly impacted by the type of wastewater treatment used, the number of inhabitants served by the WWTP, and the effluent dilution in the receiving water [29, 37]. Additionally, micropollutants from agricultural use (dominated by pesticides) reach surface water via diffuse inputs from leaching, and in particular by surface runoff during rain events [18, 33]. Surface waters may also be contaminated by local inputs from industrial production sites, landfills, or accidental releases, either directly or via WWTPs. These inputs might contain highly specific substances or substances in much higher concentrations as compared to municipal wastewater (e.g. [14, 41, 43]).

To select the relevant compounds to be monitored among these thousands of chemicals, different approaches for prioritisation have been developed. These use either predicted environmental concentrations from consumption and emissions models or measured concentrations and compare these to (eco)toxicological threshold values (e.g. [1, 11, 55, 57]). The outcomes using these methods depend strongly on the availability and quality of data, which might be very limited for certain compound classes [55]. Thus, they bias the spectrum towards well-known compounds, unknown or unexpected compounds such as metabolites and by-products are hardly considered. Furthermore, such prioritisation approaches based on general emissions scenarios will not rank site-specific chemicals among the top candidates.

With the introduction of liquid chromatography coupled to high-resolution mass spectrometry (LC-HRMS), it became possible to screen water samples for a more comprehensive spectrum of chemicals, provided that these are amendable to the individual analytical steps of the method [21, 31, 53, 54]. As a consequence, data-driven approaches based on LC-HRMS have been put forward for the discovery and prioritisation of compounds by mining LC-HRMS data for the presence of a large number of known chemicals in so-called suspect screening approaches [17, 55, 56]. However, in a suspect screening based on a list of known chemicals, unknown or unexpected compounds such as transformation and by-products are hardly considered and very large compound list have to be processed to cover all potentially relevant compounds.

Non-target screening approaches can be applied without any prior knowledge of the compounds present solely starting from the analytical data [31]. It has been successfully applied to prioritise so far unknown chemicals in rivers based on time series analysis [8, 22], their spatial trends in a river course [49], or in the context of fish mortality in a river [44]. It may be expected that NTS will be increasingly applied in water monitoring in Europe [6]. However, an exhaustive identification of all chemicals at all sites is not realistic.

In the present paper, we suggest a robust approach based on a semiautomated evaluation of non-target LC-HRMS data for the prioritisation of water bodies with site-specific contamination and the identification of the underlying chemicals. These could either be compounds which are found with the given detection limits at only one or a few sites or whose concentrations are several orders of magnitude higher at a particular site as compared to other sites or catchments due to local inputs. To express site-specific contamination in a single value, we calculate rarity scores for each detected peak in the dataset and prioritise for sites with a high number of these peaks. Additionally, peak attributes with a diagnostic value such as isotope patterns, mass defects, and occurrence of homologue series are considered.

Using a set of 31 samples from the catchments of the Rivers Saale and Mulde, Germany, the prioritisation approach is demonstrated for three sites with a high number of site-specific peaks, which were further characterised and identified.

Materials and methods

Sites and sampling

Surface water was sampled at 31 sites from the catchments of the rivers Saale and Mulde, which are major tributaries of the river Elbe in Germany (Additional file 1: Figure S1). Sites were selected at rivers and streams of different size, downstream of the discharge of industrial and municipal wastewater treatment plant (WWTP) effluents, or upstream of the first WWTP. Samples were taken in 5 L aluminium containers and stored at 4 °C until extraction. Details on the sites are given in Additional file 1: Table S1.

Estimation of discharge and wastewater fraction

Discharge data for the sampling sites were obtained for each sampling date either from associated gauging stations or from the size of the flow profile of small streams and flow velocity measurements over the profile. Details are given by Hug et al. [26] and in Additional file 1:

Table S2. The mean annual discharge was obtained from hydrological records for gauging stations. In general, the real wastewater fractions are likely larger than those calculated, as for WWTPs < 2000 person equivalents, no data were available and wastewater from decentralised treatment or untreated wastewater might contribute. In the study area, about 85% of the inhabitants are connected to centralised wastewater treatment, and particularly in rural areas on-site treatment, mainly through septic tanks, or discharge directly into surface waters can occasionally be found.

Chemical analyses

Information on chemicals used is given in Additional file 1: Sect. 1.1. Extraction of the samples was done by solid-phase extraction using multi-layer cartridges similar to those described elsewhere [28, 39]. Details are given in Additional file 1: Sect. 1.2. Within a previous study focusing on ecotoxicological characterisation of these extracts by an in vitro assay, also a target screening for 205 compounds was carried out [26]. For this study, the extracts were stored at −20 °C and analysed within 3 months after extraction and the data evaluation reported in this study is based on the archived data. Extracts concentrated 625-fold were analysed by LC-HRMS using reversed-phase separation and electrospray ionisation (ESI) in positive (ESI+) and negative ion mode (ESI−). A nominal resolving power of 100,000 referenced to *m/z* 400 was used (details see Additional file 1: Sect. 1.3). MS/MS spectra were obtained in additional runs using data-dependent MS2 on a precursor ion list of the prioritised peaks with collision-induced dissociation (CID) and higher energy collisional dissociation (HCD) at different collisions energies and a nominal resolving power of 15,000. Due to the biological analysis of the extracts [26], isotope-labelled internal standards were added only prior to LC-HRMS analysis and were used for quality control of peak detection in this study.

Automated peak detection

Raw HRMS full-scan chromatograms (*m/z* 100–1000) from ESI+ and ESI− runs were converted from profile to centroid mode and to .mzML format using ProteoWizard 3.0.6485 [30]. Afterwards, aligned peak lists containing 31 samples, one processing blank (i.e. from 10 mL of ultrapure water processed with the SPE cartridge as a sample) and one solvent blank (i.e. the type and amount of solvent used for eluting the cartridge processed as a sample) were generated by MZmine 2.20 [45]. We applied the steps mass detection, FTMS shoulder peak detection, chromatogram building, smoothing, peak deconvolution by local minimum search (minimum peak intensity 30,000 in ESI+ and 10,000 in ESI− mode), and alignment by the Join Aligner algorithm. Settings were slightly adjusted from Hu et al. [24] and are given in Additional file 1: Table S3. For further processing, ESI+ and ESI− peak lists with accurate *m/z*, retention time, peak intensity and area were exported from MZmine as.csv files. From the processing and solvent blanks, a combined blank peak list was generated in Excel by taking the maximum value in each of these two blanks. The sample and blank peak lists were imported into R, v3.3.0 (R [46] for further processing. All peaks with an area-to-height ratio > 50 were removed from the peak list to exclude signals coming from background noise (for details see [25]. Peaks with an intensity ratio < 10 between the surface water and the blank peak list were excluded from further analysis. Peak lists were exported for all individual samples as.csv files.

Determination of rarity scores

From the aligned peak lists, *m/z*, retention times (defining a unique peak in the dataset) and the intensities in all samples were used. To identify peaks which occur at a small number of the studied sites with high intensity as compared to the other sites, we calculated a rarity score for each peak *x* (RS_x) according to:

$$\mathrm{RS}_x = \frac{\text{maximum intensity across all sites}(x)}{\text{median intensity across all sites}(x)} \cdot \frac{\text{total number of samples}}{\text{number of positive detects}} \quad (1)$$

For the calculation of the median intensity, non-detects were replaced with the threshold intensity of the peak detection in MZmine (30,000 in ESI+ mode and 10,000 in ESI− mode).

The rarity scores combine a low frequency of occurrence of a peak in a dataset and its maximum signal intensity in relation to the median intensity in one single number. As for every value based on a statistical measure, the calculation of rarity scores is meaningful only for a larger dataset, although we cannot specify a minimum number of samples. While a value of 1 is the lowest possible one (with median = maximum intensity and present at all sites), the maximum values are given by the peak intensity range obtained by the instrument used and the number of samples.

An advantage of this univariate approach is that it can be applied even to datasets with a large fraction of non-detects or “zeroes” using the threshold signal intensity for the missing data. Such data may be difficult to handle by more sophisticated multivariate statistical methods, which are commonly used to prioritise peaks in metabolomics and occasionally in environmental studies (e.g. [50]). Such methods are either vulnerable to bias of the data or computation is time-consuming and requires expert knowledge [19, 20]. Common methods are the deletion of data records with missing values, imputation of single values (e.g. the detection limit or the half detection limit) or simple regression imputation. Deletion of records yields a small number of cases or variables and would remove such site-specific peaks occurring in maybe one single sample. Constant values infuse the data with unconditional and uncorrelated observations and thus bias the variance and correlation relationship as the distribution is changed. Regression imputation of perfectly correlated values is vulnerable to overestimation of the regression fit.

Determination of additional peak attributes

Mass defects (i.e. differences between the nominal mass and the accurate monoisotopic mass of an ion) were calculated using an R script, assuming that the mass defects span a range from $M-0.4$ to $M+0.6$ for a nominal mass M. Using the R package non-target version 1.8 [34], peak lists of all individual samples were screened for bounds of feasible isotope peaks (^{13}C, ^{15}N, ^{34}S, ^{37}Cl and ^{81}Br) with a rule-based algorithm. Homologue series detection [35] was carried out for four or more consecutive mass differences corresponding to CH_2, CH_2O, C_2H_4O, C_3H_6O, C_2H_6SiO, CF_2 and C_2H_4 units for singly and doubly charged ions. Peaks were finally grouped into components, i.e., the monoisotopic peak and its associated isotope or adduct peaks representing an individual chemical compound. Details of the R package non-target settings are given in Additional file 1: Table S4. Statistical analyses were conducted using R, v3.3.0 and Statistica 12 (Statsoft Inc.). For data visualisation, the R packages ggplot2 [61] and ggradar (https://rdrr.io/github/ricardo-bion/ggradar, last accessed 01/07/2019) were used.

Identification of prioritised peaks

For the identification of prioritised peaks, most plausible molecular formulas were determined from the raw data files based on accurate masses and isotope patterns using the QualBrowser of Xcalibur (Thermo Scientific) by visual comparison of measured and simulated mass spectra. Possible structures were searched in compound databases (Chemspider, Royal Society of Chemistry [48]; Pubchem, NCBI [42], and experimental MS/MS spectra were searched against MassBank [23]. Plausible candidate compounds were selected based on commercial/industrial relevance and additional literature search. For confirmation, reference standards were obtained if available. Confidence levels for the identification were assigned according to [52]. Marvin, InstantJChem and JChem for Excel (Chemaxon, Budapest, Hungary) were used for chemical structure drawing and handling and calculation of ion masses. Given the scope of this study on evaluating the approach for its potential to prioritise site-specific contamination, compound identification was based solely on compound database search of molecular formulas, a MassBank search and confirmation of plausible hits with reference standard, but we did not use additional approaches such as MS fragmentation prediction or retention time prediction to assist with the identification of candidate compounds within this study.

Results and discussion

Prioritisation of site-specific contamination based on rarity scores

The distribution of rarity scores was similar for ESI+ and ESI− mode, with about 80% of the detected peaks showing values between 10 and 100. About 1% of peaks had values above 1000, thus we prioritised these as representing site-specific compounds in our set of samples (Fig. 1). At this RS level, peaks down to a signal intensity of 10^6 (which would for example correspond to a concentration of 20 ng/L of a well ionising compound such as atrazine) might become classified as rare peaks if they occur in a low number of samples. These peaks would usually be missed if only signal intensity is used as a criterion for prioritisation. In contrast, many peaks with intensities between 10^6 and 10^7 occurring at many of the study sites get rarity scores below 50, thus these are not considered for site-specific contamination. Obviously, peaks detected at low intensities at only a few sites might potentially also represent a site-specific contamination; however, this cannot be assessed based on the data, as we do simply not know whether (or at which level) the compounds are present in samples where we could not detect these.

Figure 2 shows that most of the prioritised peaks with $RS > 1000$ actually occur in only one sample, with substantially smaller numbers in 2–7 samples in ESI+ and two or three samples in ESI−. Only one to three peaks with an $RS > 1000$ occur in 8–15 samples in ESI+, and in 4–10 samples in ESI−; these were mainly peaks with intensities $> 5 \times 10^6$ in one or a low number of samples and much lower intensities in other samples. These findings suggest that the rarity score is a suitable approach to prioritise site-specific contamination which is characterised either by high differences in intensity among the samples (as a proxy for concentration) or by a restricted frequency of occurrence. It should be noted, however, that the calculation of the RS might be problematic in rare cases: If a compound is detected in 15 out of 31 samples at peak intensities around 10^8, but not in the other 16 samples,

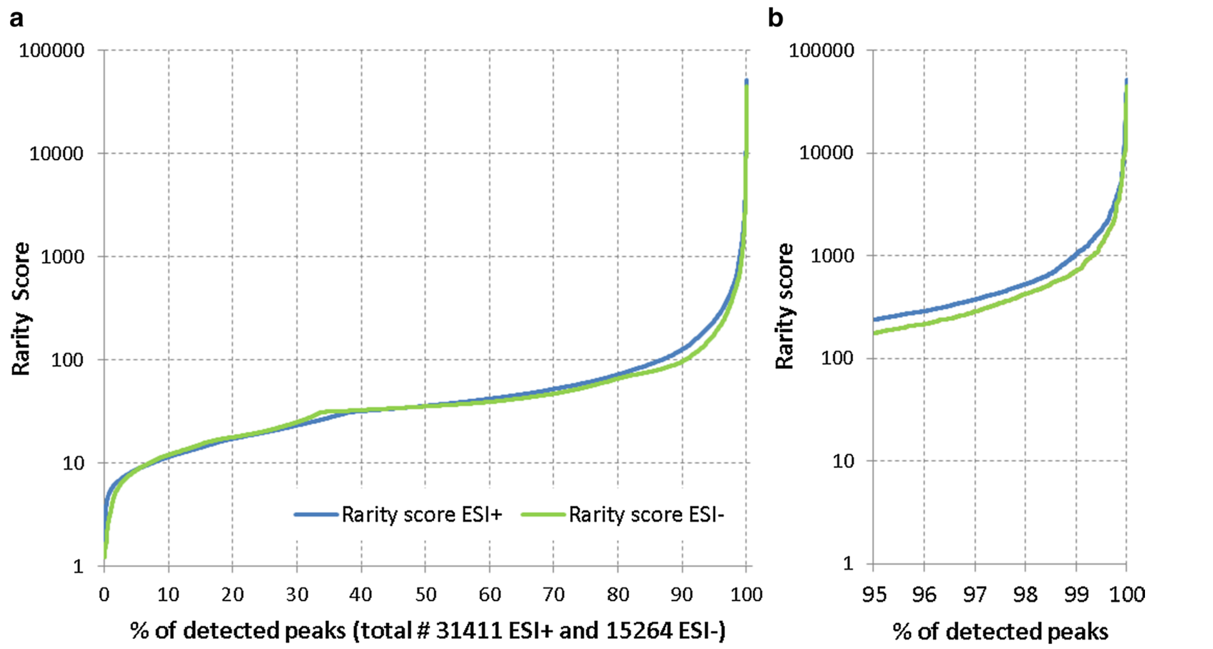

Fig. 1 **a** Rarity scores in ESI+ and ESI− mode of all detected peaks (sorted according to increasing rarity scores), which is enlarged in **b** for the range of 95–100%

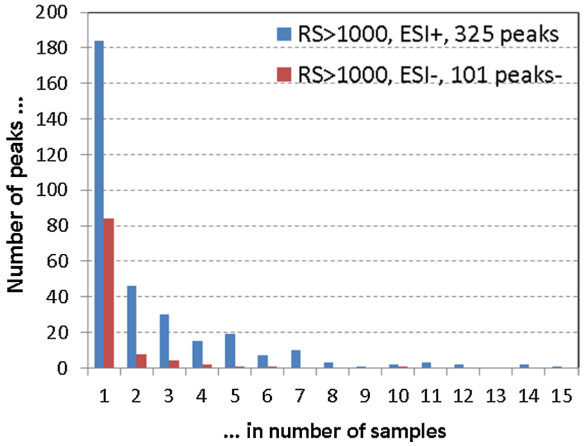

Fig. 2 Frequency of occurrence of peaks with rarity scores > 1000 in ESI+ and ESI− mode in the 31 samples

the resulting RS value would be about 25,000, as the median is at a threshold of 10,000. If the compound is detected in 16 out of 31 samples at peak intensities around 10^8, but not in the other 15 samples, the resulting RS value would be around 5, as the median is about 10^8.

To prioritise sites with a specific contamination, we compared the numbers of peaks with rarity scores above threshold levels of 5000 and 1000, respectively, among all sites as shown in Fig. 3. The number of peaks with high rarity scores showed large differences among the samples. A RS value of 5000 was exceeded by up to ten compounds in ESI+ and 13 compounds in ESI− mode in one sample, and a value of 1000 by up to 91 compounds in ESI+ and up to 48 compounds in ESI− mode in individual samples, while other samples had no single compound with rarity scores above these levels. In ESI+ mode, the sites with the largest number of rare peaks are B2, S8, DB, LN, H2 (RS > 5000) and DB, LN, LA, WE (RS > 1000). In ESI− mode site, SP shows clearly the site with the largest number of rare peaks (13 peaks with RS > 5000, 48 peaks with RS > 1000). Still considerable numbers of peaks with RS > 1000 in ESI− mode could be detected at the sites Sol (14 peaks) and DB (11 peaks). The occurrence of detected peaks with rarity scores > 1000 in the individual samples is given in Tables S5 (ESI+) and S6 (ESI−) in Additional file 2.

Using peak attributes to further characterise site-specific contamination

For a further characterisation of site-specific contamination, we used the percentage of peaks containing potentially Cl, Br and S (as inferred from the isotopologue detection), the percentage of peaks with negative mass defect and those being part of a homologue series which are visualised in Fig. 4 for all sites. A detailed discussion of the performance of the peak attribute

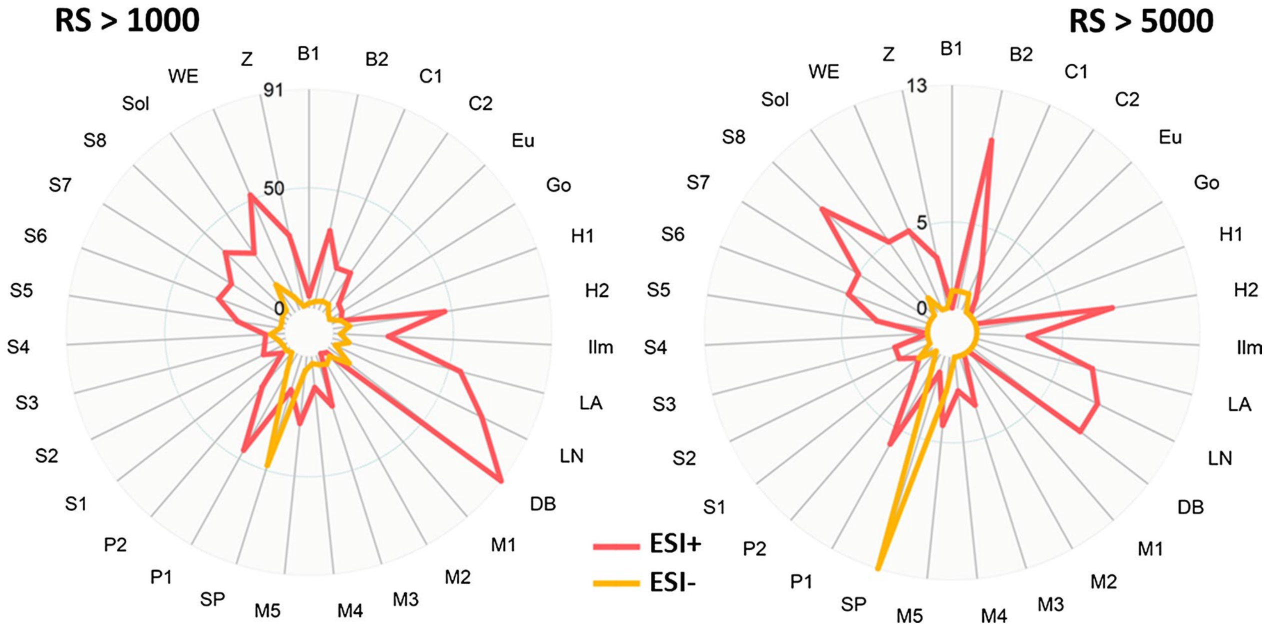

Fig. 3 Comparison of the studied sites for the number of peaks above rarity scores of 1000 and 5000, respectively

determination and consequences for the usage of this data is given in Additional file 1: Sect. 2.1.

For site B2, the large number of rare peaks coincides with the largest percentage of peaks with potential Cl and Br isotopologue peaks and with negative mass defects (Figs. 3, 4). Other sites with relatively high fractions of Cl or Br isotopologue peaks were H1 and H2, with S isotopologue peaks B1, H1 and P2. The highest fractions of peaks in homologue series were found at sites B1, H1, M1 and M2. In ESI− mode, the largest number of rare peaks at site SP coincides with the largest percentage of peaks with potential S isotopologue peaks and with negative mass defects. Site WE showed a similarly large fraction of S isotopologue peaks and peaks with negative mass defects as site SP, but no peaks with particularly high rarity scores. Similar as for ESI+ mode, site B2 had the largest percentage of Cl or Br isotopologue peaks, followed by site H1. The percentage of peaks contained in homologue series did not coincide with the occurrence of a large number of site-specific peaks, and in fact, no such peaks were among those with RS > 1000.

Characterisation and identification of site-specific compounds

The previous section showed a range of sites with a specific contamination pattern and a detection of a high number of peaks with a RS > 1000, for which the generated peak attribute information can be used within the identification process. This is exemplified here for the sites SP (in ESI− mode; RS > 1000 for 48 peaks), DB (in ESI+ mode; RS > 1000 for 91 peaks) and H2 (in ESI+ mode; RS > 1000 for 47 peaks). A detailed overview of all peaks with rarity scores > 1000 in all samples is given in Tables S7 (ESI+) and S8 (ESI−) in Additional file 2.

Site SP (Spittelwasser downstream of Bitterfeld)

The Spittelwasser showed a large number of peaks with high RS values in ESI− mode, along with a large percentage of compounds containing sulphur and with negative mass defects based on the determination of peak attributes (Fig. 4). In contrast, peak numbers and numbers of compounds with high RS values in ESI+ mode were not noticeably high. As evident from Figure S4C (Additional file 1), many of these high-intensity S-containing compounds eluted around 3 min, others around 18-20 min retention time. The two most intense peaks at RT 18.8 and 19.7 min could be identified as being 1- and 2-naphthalenesulphonic acid (*m/z* 207.0121, M−H$^-$) based on a reference standard of 2-naphthalenesulphonic acid. Most of the other S-containing compounds were tentatively assigned as closely related compounds such as naphthalenedisulphonic (*m/z* 286.9688, M−H$^-$) and naphthalenetrisulphonic acids (*m/z* 366.9256 for M−H$^-$ and *m/z* 182.9594 for [M−2H]$^{2-}$) hydroxy- and amino-naphthalenesulphonic acids (*m/z* 302.9637 and *m/z* 301.9796, respectively, both M−H$^-$) as well as naphthylsulphate (*m/z* 223.0069, M−H$^-$). These assignments were based on the match and similarity of MS/MS spectra and

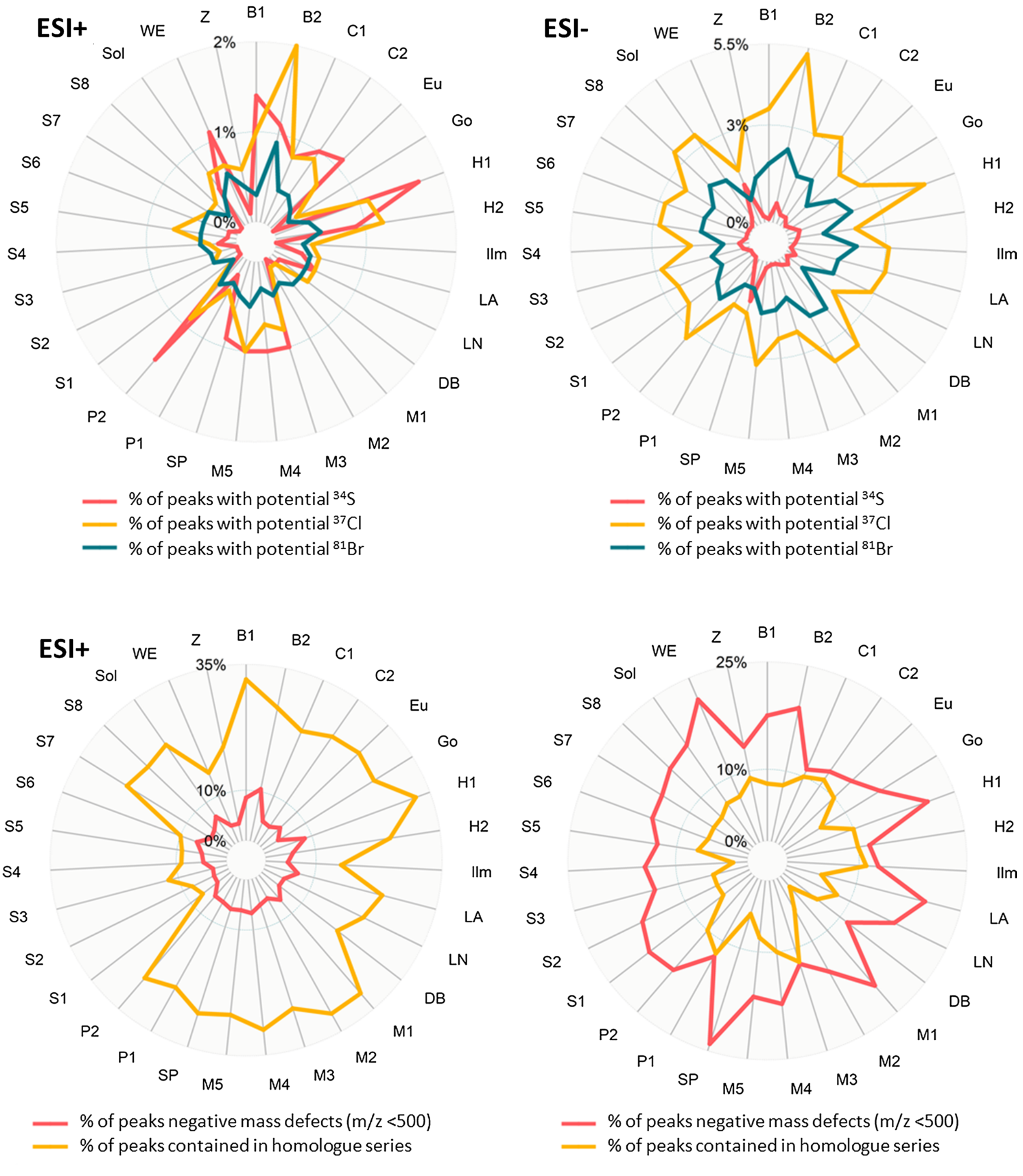

Fig. 4 Comparison of the studied sites for percentages of peaks with S/Cl/Br isotope pattern (top) and percentages of peaks with negative mass defect (*m/z* < 500) and contained in homologue series (bottom)

retention times to those of reference compounds. Details are given in Table S9 and Figures S5 to S11 (Additional file 1).

Thus, our non-target screening approach revealed that derivatives of naphthalenesulphonic acids are important water contaminants in the Spittelwasser. The occurrence of naphthalene sulphonic acids and their derivatives in high concentrations was demonstrated for textile and tannery wastewater [9], stemming from their use as dye precursors and use in dyeing processes, and in landfill leachates [47]. In case of the Spittelwasser, it is likely that the found compounds are a legacy contamination related to the former dye (and maybe other chemicals) production at Bitterfeld. The contamination of sediments of the Mulde river and its tributary Spittelwasser with arylsulphonic acid derivatives and alkylsulphonic acid aryl esters was previously recognised [4, 15, 16]. Sediments of the Spittelwasser and lower Mulde are also heavily contaminated by persistent chlorinated compounds from the former chemical industry [15]. However, we did not detect a large percentage of chlorinated compounds in the Spittelwasser water sample, suggesting that their occurrence is limited to compounds with a high affinity to sediments and/or a poor ionisation by ESI. The peaks with high RS values at the site SP were not detected at any other site studied, except for one peak of a naphthalenedisulphonic acid found at about 100-fold lower intensity at site S1 (*m/z* 182.9594, RT 2.3 min).

Site DB (Dorfbach Niederschindmaas)

The Dorfbach is a small brook receiving wastewater from the WWTP of a large car manufacturer with 8000 employees, which also treats municipal wastewater of about 3000 inhabitants from adjacent settlements. In ESI+ mode, this site shows the largest number of peaks with RS values above 1000 (Fig. 3). The most intense rare peak in ESI+ mode at *m/z* 391.2294 could be identified as hexa(methoxymethyl)melamine (HMMM) based on a reference standard (Additional file 1: Table S10 and Figure S13). The full-scan spectrum of HMMM shows a significant in-source fragmentation resulting in the loss of one to three CH_4O from the protonated molecule, which was partially assigned vice versa as a methanol adduct of the fragment by the non-target package (Additional file 1: Figure S14). Without a reference standard, it is indeed impossible to distinguish in-source fragmentation from methanol adduct formation in this case. Several other high-intensity peaks showed a similar full-scan mass spectral pattern (CH_4O losses difference), and molecular formulas suggested compounds related to HMMM. The same peaks were detected in an old HMMM reference standard stored for more than 2 years at 4 °C, where they obviously stem from hydrolysis (Additional file 1: Figure S13). Although these compounds showed a low fragment ion intensity (typically in the 10^4 intensity range despite 10^7 precursor ion intensity) resulting in poor MS/MS spectra (Additional file 1: Figure S14), these compounds were tentatively identified as penta- and tetra(methoxymethyl)melamine and *O*-demethylated HMMM. HMMM is one important precursor of melamine–formaldehyde resins used for durable coatings such as in beverage cans and car paint finishes. Thus, the car manufacturer releasing treated wastewater in the Dorfbach is a plausible source. The technical product contains a mixture of monomers and oligomers of HMMM as well as not fully methyl-methoxylated melamine (US EPA [58]. Thus, the observed demethylated and demethoxymethylated derivatives might stem both from transformation or these technical mixtures. HMMM and related compounds have been previously identified in wastewater and surface water [2, 3, 44]. A widespread presence of HMMM in German rivers was reported by Dsikowitzky and Schwarzbauer [13] with a huge temporal and spatial variation and maximum concentrations of up to 880 ng/L in the Mulde river. Furthermore, we detected several “rare” and high-intensity peaks with molecular formulas and retention times similar to those of HMMM (e.g. $C_{25}H_{28}O_8N_4$ at 22.8 min, $C_{12}H_{18}O_4N_6$ at 21.6 min, $C_{25}H_{33}O_5N_5$ at 20.3 min, $C_{22}H_{25}O_8N_6$ at 22.9 min, Table S10),which could be caused by the presence of similar substituted melamines used for production of melamine–formaldehyde resins [12]. The co-occurrence of several melamine-derivatives coincides with results from Peter et al. [44], who detected a (methoxymethyl)melamine “compound family” in urban stormwater runoff in the USA.

Most of these compounds could also be detected at the other studied sites (Fig. 5), among them sites LN and S6, with peak heights about two orders of magnitude lower than at site DB, pointing towards some specific sources there. At the other sites fewer compounds, mainly PMMM and tetra(methoxymethyl)melamine were found and peaks heights were in general about three orders of magnitude lower, confirming a widespread occurrence of this compound class in surface waters.

Site H2 (Holtemme downstream of WWTP Silstedt)

Site H2 on the Holtemme river receives municipal wastewater from one relatively large WWTP (80,000 person equivalents) serving the town of Wernigerode and surrounding villages, showing second largest number of peaks with RS values > 1000 in ESI+ mode. The twenty most intense peaks with RS > 1000 and their tentative identification and confirmation are shown in Additional file 1: Table S11 and Figures S16–S19.

Among these compounds were 7-diethylamino-4-methylcoumarin, 7-ethylamino-4-methylcoumarin and 7-amino-4-methylcoumarin, which were recently identified at this site as causative compounds for the observed anti-androgenicity [40]. The latter is used as

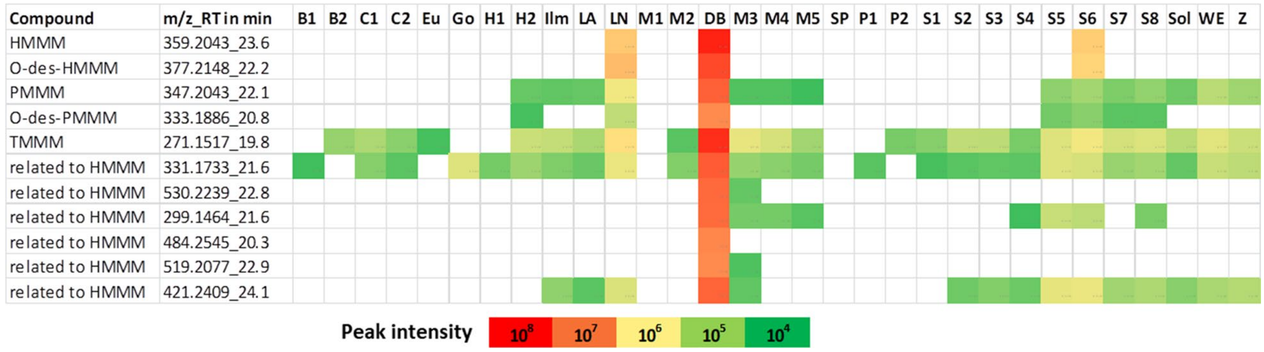

Fig. 5 Occurrence and peak intensities at all the 31 studied sites of hexa(methoxymethyl)melamine (HMMM), likely transformation products and related compounds detected as site-specific contaminants at site DB (PMMM: penta(methoxymethyl)melamine; TMMM: tetra(methoxymethyl) melamine; O-des: O-desmethylated compound). Details on the tentative identification of the compounds are given in Additional file 1

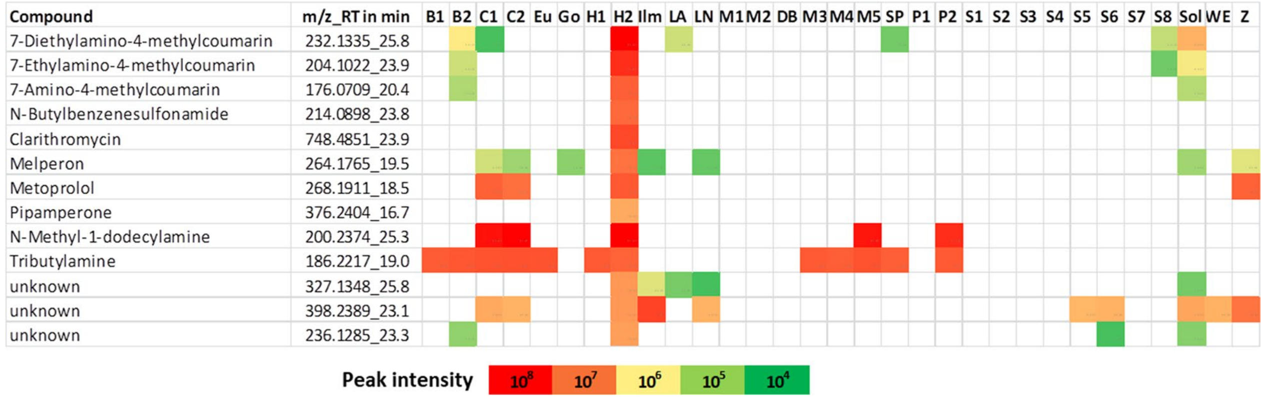

Fig. 6 Occurrence and peak intensities at all the 31 studied sites of the most intense compounds with RS values > 1000 at site H2. Details on the tentative identification of the compounds are given in Additional file 1

optical brightener (or fluorescent whitening agent) for textiles and a constituent in cleaning detergents and washing powders [27] and has not been found anywhere else in surface water or wastewater [36]. We could detect one to all three of these compounds at six other sites at much lower peak heights (Fig. 6). While at sites B2 and S8, which are located further downstream of site H2 at the Bode and Saale, respectively, the occurrence might be related to the input upstream of H2, an input occurs also into the Solgraben (site Sol), the Chemnitz (C1), the Alte Luppe (LA) and the Spittelwasser (SP).

Furthermore, we could identify the antipsychotic drugs melperon and pipamperone and the antibiotic clarithromycin (included in the target screening compound set of [26]). All compounds could be confirmed by reference standards. The high concentrations of pipamperone and melperon (estimated > 1 µg/L) and of clarithromycin of more than 5 µg/L [26] are not likely to stem from medical use. Pipamperone was previously analysed by Van De Steene et al. [59] and was detected in WWTP effluents typically at concentrations below 40 ng/L and in surface water below 20 ng/L. However, the authors found high pipamperone concentrations of up to 36 µg/L in the effluent of a WWTP treating wastewater from pharmaceutical and chemical industries. We did not detect pipamperone at any other site, while melperon was detected at six other sites at levels at least 20-fold lower. Clarithromycin concentrations in WWTP effluents are typically in the range of 50–500 ng/L [38, 53, 60], and the wastewater fraction in the Holtemme was calculated to be at about 27%. We did not detect clarithromycin or pipamperone at any other site. Thus, the most probable source of these compounds is the production by a pharmaceutical company located in Wernigerode. Emissions from drug manufacturing have been recognised as a significant source of pharmaceuticals at specific sites [7, 14, 25].

Metoprolol, *N*-methyl-1-dodecylamine and tributylamine were detected at similar peak intensities at three, four and eleven other sites, respectively, resulting in high

RS values for these compounds, but not at other sites although previous studies indicate a ubiquitous occurrence of metoprolol in the aquatic environment [26]. The manual re-evaluation of these compound peaks in Xcalibur indicated that this finding was based on artefacts related to peak picking with MZmine, as all three compounds were present in most samples with varying intensity. However, the peaks of these compounds bearing all an aliphatic amino group were typically more than 1.5 min wide with a significant tailing, which hampered the peak picking, and resulted in their misclassification as site-specific contaminants. Note that in this case a false negative in peak detection resulted in a false positive assignment of a site-specific peak.

Conclusions

A new approach to identify and prioritise samples with a significant site-specific contamination based on LC-HRMS non-target screening data without any prior knowledge of the chemicals present was proposed. It is based on a simple calculation of rarity scores (RS) for each detected peak, without the need that the dataset fulfils any prerequisites for more sophisticated statistical approaches. The data processing steps with a final prioritisation of site-specific peaks and determination of peak attributes can be accomplished within 3–4 h using freely available software and can be applied by users less experienced in non-target screening or statistical data evaluation. The obtained rarity scores can be used for both, ranking compounds for identification, but also for ranking sites with a large number of such peaks for further investigation. As the magnitude of RS values depends on the instrument used and the dataset itself, it is not possible to set a general threshold value for site-specific peaks; the selection of peaks should instead be guided by the ranking and the occurrence among the different sites, and—very pragmatically—by the time which can be spent on the subsequent identification. This second step is by far more laborious and time-consuming, although automated workflows including MS/MS fragmentation prediction and MS library search have been established (e.g. [8]). Nevertheless, some degree of expert knowledge is required, but efforts can be focused on the relevant compounds and are supported by the automated annotation of isotopologues, homologue series and mass defects.

LC-HRMS instrumentation is currently becoming more frequently available also at authorities carrying out regulatory monitoring (e.g. along the Rhine river; [22, 32]). The proposed approach to detect site-specific contamination can be used by such authorities in investigative monitoring of catchments and water bodies which fail to meet quality criteria, while monitoring of priority substances and RBSPs does not indicate a chemical pollution issue. It may also be directly applied for locations where a specific contamination is suspected rather than using targeted methods focusing on a limited set of compounds. This will significantly reduce the risk of overlooking possibly hazardous chemicals (including unknowns), for which detailed investigations on sources and toxicity can follow. Ultimately, it could guide compound- and source-specific mitigation measures at sites where problematic compounds are emitted.

Additional files

Additional file 1. Sampling sites, data processing and performance evaluation, as well as on compound identification for selected site-specific contaminants.

Additional file 2. Calculated peak attributes for all sites and the intensities of peaks with high rarity scores for all sites.

Abbreviations

ESI: electrospray ionisation; FTMS: Fourier transform mass spectrometry; HMMM: hexa(methoxymethyl)melamine; LC-HRMS: liquid chromatography–high-resolution mass spectrometry; *m/z*: mass-to-charge ratio; NTS: non-target screening; PMMM: penta(methoxymethyl)melamine; RBSP: River basin specific pollutant; RS: rarity score; RT: retention time; WFD: Water Framework Directive; WWTP: wastewater treatment plant.

Acknowledgements

We thank Peter von der Ohe for supporting the selection of sites, Marion Heinrich for help during sample collection, Margit Petre and Tim Reinicke for assistance with sample preparation. Emma Schymanski and Martin Loos of Eawag are acknowledged for helpful comments on a previous version of the manuscript. A free academic licence for Marvin, JChem for Excel and InstantJChem was kindly provided by ChemAxon (Budapest, Hungary).

Authors' contributions

MK, RB, WB and TS designed the sampling campaign. RB, MK, WB and TS collected samples. RB and CH performed sample extraction and instrumental analysis. MK developed the concept for prioritisation of site-specific contamination from non-target screening data, evaluated the data and wrote the draft. RB, CH and TS contributed to the data evaluation workflow. WB, CH and TS critically revised the manuscript. All authors read and approved the final manuscript.

Funding

We acknowledge financial support by the SOLUTIONS project within the European Union's Seventh Framework Programme for research, technological development and demonstration under Grant Agreement No. 603437.

Availability of data and materials

The calculated peak attributes for all sites are available in Additional file 2, as well as the intensities of peaks with high rarity scores for all sites.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details
[1] Department of Effect-Directed Analysis, UFZ–Helmholtz Centre for Environmental Research, Permoserstr.15, 04318 Leipzig, Germany. [2] Department of Ecosystem Analysis (ESA), Institute for Environmental Research (Biology V), RWTH Aachen University, Worringer Weg 1, 52074 Aachen, Germany.

Received: 4 March 2019 Accepted: 7 July 2019
Published online: 22 July 2019

References

1. Besse J-P, Kausch-Barreto C, Garric J (2008) Exposure assessment of pharmaceuticals and their metabolites in the aquatic environment: application to the French situation and preliminary prioritization. Hum Ecol Risk Assess 14:665–695
2. Bobeldijk I, Stoks PGM, Vissers JPC, Emke E, van Leerdam JA, Muilwijk B, Berbee R, Noij THM (2002) Surface and wastewater quality monitoring: combination of liquid chromatography with (geno)toxicity detection, diode array detection and tandem mass spectrometry for identification of pollutants. J Chromatogr A 970:167–181
3. Botalova O, Schwarzbauer J, Al-Sandoukdlx N (2011) Identification and chemical characterization of specific organic indicators in the effluents from chemical production sites. Water Res 45:3653–3664
4. Brack W, Altenburger R, Ensenbach U, Moder M, Segner H, Schuurmann G (1999) Bioassay-directed identification of organic toxicants in river sediment in the industrial region of Bitterfeld (Germany)—a contribution to hazard assessment. Arch Environ Contam Toxicol 37:164–174
5. Brack W, Escher BI, Müller E, Schmitt-Jansen M, Schulze T, Slobodnik J, Hollert H (2018) 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? Environ Sci Eur 30:33. https://doi.org/10.1186/s12302-018-0161-1
6. Brack W, Hollender J, López de Alda M, Müller C, Schulze T, Schymanski E, Slobodnik J, Krauss M (2019) High resolution mass spectrometry to complement monitoring and track emerging chemicals and pollution trends in European water resources. Environ Sci Eur Accept Publ 512:540–551
7. Cardoso O, Porcher J-M, Sanchez W (2014) Factory-discharged pharmaceuticals could be a relevant source of aquatic environment contamination: review of evidence and need for knowledge. Chemosphere 115:20–30. https://doi.org/10.1016/j.chemosphere.2014.02.004
8. Carpenter CMG, Wong LYJ, Johnson CA, Helbling DE (2019) Fall creek monitoring station: highly resolved temporal sampling to prioritize the identification of nontarget micropollutants in a small stream. Environ Sci Technol 53:77–87. https://doi.org/10.1021/acs.est.8b05320
9. Castillo M, Alonso MC, Riu J, Barceló D (1999) Identification of polar, ionic, and highly water soluble organic pollutants in untreated industrial wastewaters. Environ Sci Technol 33:1300–1306. https://doi.org/10.1021/es981012b
10. Diamond J, Altenburger R, Coors A, Dyer SD, Focazio M, Kidd K, Koelmans AA, Leung KMY, Servos MR, Snape J, Tolls J, Zhang XW (2018) Use of prospective and retrospective risk assessment methods that simplify chemical mixtures associated with treated domestic wastewater discharges. Environ Toxicol Chem 37:690–702. https://doi.org/10.1002/etc.4013
11. Diamond JM, Latimer HA, Munkittrick KR, Thornton KW, Bartell SM, Kidd KA (2011) Prioritizing contaminants of emerging concern for ecological screening assessments. Environ Toxicol Chem 30:2385–2394. https://doi.org/10.1002/etc.667
12. Diem H, Matthias G, Wagner RA (2010) Amino resins. Ullmann's encyclopedia of industrial chemistry. Wiley, New York. https://doi.org/10.1002/14356007.a02_115.pub2
13. Dsikowitzky L, Schwarzbauer J (2015) Hexa(methoxymethyl)melamine: an emerging contaminant in German rivers. Water Environ Res 87:461–469. https://doi.org/10.2175/106143014X14060523640919
14. Fick J, Söderström H, Lindberg RH, Phan C, Tysklind M, Larsson DGJ (2009) Contamination of surface, ground, and drinking water from pharmaceutical production. Environ Toxicol Chem 28:2522–2527. https://doi.org/10.1897/09-073.1
15. Franke S, Heinzel N, Specht M, Francke W (2005) Identification of organic pollutants in waters and sediments from the Lower Mulde river area. Acta Hydrochim Hydrobiol 33:519–542
16. Franke S, Schwarzbauer J, Francke W (1998) Arylesters of alkylsulfonic acids in sediments—part III of organic compounds as contaminants of the Elbe River and its tributaries. Fresen J Anal Chem 360:580–588. https://doi.org/10.1007/s002160050762
17. Gago-Ferrero P, Krettek A, Fischer S, Wiberg K, Ahrens L (2018) Suspect screening and regulatory databases: a powerful combination to identify emerging micropollutants. Environ Sci Technol 52:6881–6894. https://doi.org/10.1021/acs.est.7b06598
18. Gassmann M, Stamm C, Olsson O, Lange J, Kümmerer K, Weiler M (2013) Model-based estimation of pesticides and transformation products and their export pathways in a headwater catchment. Hydrol Earth Syst Sci 17:5213–5228. https://doi.org/10.5194/hess-17-5213-2013
19. Guida R, Engel J, Allwood JW, Weber RJM, Jones MR, Sommer U, Viant MR, Dunn WB (2016) Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. Metabolomics 12:1–14. https://doi.org/10.1007/s11306-016-1030-9
20. Hastie T, Tibshirani R, Friedman J, Hastie T, Friedman J, Tibshirani R (2009) The elements of statistical learning, vol 2. Springer, Berlin
21. Hérnández F, Pozo ÓJ, Sancho JV, López FJ, Marín JM, Ibáñez M (2005) Strategies for quantification and confirmation of multi-class polar pesticides and transformation products in water by LC-MS2 using triple quadrupole and hybrid quadrupole time-of-flight analyzers. Trends Anal Chem 24:596–612
22. Hollender J, Schymanski EL, Singer HP, Ferguson PL (2017) Nontarget screening with high resolution mass spectrometry in the environment: ready to go? Environ Sci Technol 51:11505–11512. https://doi.org/10.1021/acs.est.7b02184
23. Horai H et al (2010) MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 45:703–714
24. Hu M, Krauss M, Brack W, Schulze T (2016) Optimization of LC-Orbitrap-HRMS acquisition and MZmine 2 data processing for nontarget screening of environmental samples using design of experiments. Anal Bioanal Chem 408:7905–7915. https://doi.org/10.1007/s00216-016-9919-8
25. Hug C, Ulrich N, Schulze T, Brack W, Krauss M (2014) Identification of novel micropollutants in wastewater by a combination of suspect and nontarget screening. Environ Pollut 184:25–32. https://doi.org/10.1016/j.envpol.2013.07.048
26. Hug C, Zhang X, Guan M, Krauss M, Bloch R, Schulze T, Reinecke T, Hollert H, Brack W (2015) Microbial reporter gene assay as a diagnostic and early warning tool for the detection and characterization of toxic pollution in surface waters. Environ Toxicol Chem 34:2523–2532
27. Hunger K (2003) Industrial Dyes. Wiley, New York. https://doi.org/10.1002/3527602011.fmatter
28. Huntscha S, Singer HP, McArdell CS, Frank CE, Hollender J (2012) Multiresidue analysis of 88 polar organic micropollutants in ground, surface and wastewater using online mixed-bed multilayer solid-phase extraction coupled to high performance liquid chromatography-tandem mass spectrometry. J Chromatogr A 1268:74–83. https://doi.org/10.1016/j.chroma.2012.10.032
29. Karakurt S, Schmid L, Hübner U, Drewes JE (2019) Dynamics of wastewater effluent contributions in streams and impacts on drinking water supply via riverbank filtration in germany—a national reconnaissance. Environ Sci Technol 53:6154–6161. https://doi.org/10.1021/acs.est.8b07216
30. Kessner D, Chambers M, Burke R, Agusand D, Mallick P (2008) ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 24:2534–2536. https://doi.org/10.1093/bioinformatics/btn323
31. Krauss M, Singer H, Hollender J (2010) LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns. Anal Bioanal Chem 397:943–951
32. Kunkel U, Ruppe S, Loos M, Schlüsener M, Singer H, Brüggen S, Fügel D, Bader T, Zemmelink H, Pijnappels M, Scheurer M, Weiß S (2018) Vereinheitlichte mess-und auswertemethoden als grundlage für eine effiziente zeitnahe und grenzübergreifende gewässerüberwachung mittels non-target-analytik? Vom Wasser 116:109–113
33. Leu C, Singer H, Stamm C, Muller SR, Schwarzenbach RP (2004) Simultaneous assessment of sources, processes, and factors influencing

herbicide losses to surface waters in a small agricultural catchment. Environ Sci Technol 38:3827–3834

34. Loos M (2012) Nontarget: detecting, combining and filtering isotope, adduct and homologue series relations in high-resolution mass spectrometry (HRMS) data. R Package Version. 2012:1
35. Loos M, Singer H (2017) Nontargeted homologue series extraction from hyphenated high resolution mass spectrometry data. J Cheminform 9:12. https://doi.org/10.1186/s13321-017-0197-z
36. Loos R, Hanke G, Eisenreich SJ (2003) Multi-component analysis of polar water pollutants using sequential solid-phase extraction followed by LC-ESI-MS. J Environ Monit 5:384–394. https://doi.org/10.1039/B300440F
37. Luo Y, Guo W, Ngo HH, Nghiem LD, Hai FI, Zhang J, Liang S, Wang XC (2014) A review on the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment. Sci Total Environ 473–474:619–641. https://doi.org/10.1016/j.scitotenv.2013.12.065
38. Michael I, Rizzo L, McArdell CS, Manaia CM, Merlin C, Schwartz T, Dagot C, Fatta-Kassinos D (2013) Urban wastewater treatment plants as hotspots for the release of antibiotics in the environment: a review. Water Res 47:957–995. https://doi.org/10.1016/j.watres.2012.11.027
39. Moschet C, Piazzoli A, Singer H, Hollender J (2013) Alleviating the reference standard dilemma using a systematic exact mass suspect screening approach with liquid chromatography-high resolution mass spectrometry. Anal Chem 85:10312–10320. https://doi.org/10.1021/ac4021598
40. Muschket M, Di Paolo C, Tindall AJ, Touak G, Phan A, Krauss M, Kirchner K, Seiler T-B, Hollert H, Brack W (2018) Identification of unknown antiandrogenic compounds in surface waters by effect-directed analysis (EDA) using a parallel fractionation approach. Environ Sci Technol 52:288–297. https://doi.org/10.1021/acs.est.7b04994
41. Muz M, Dann JP, Jäger F, Brack W, Krauss M (2017) Identification of mutagenic aromatic amines in river samples with industrial wastewater impact. Environ Sci Technol 51:4681–4688. https://doi.org/10.1021/acs.est.7b00426
42. NCBI PubChem. National Center for Biotechnology Information. http://pubchem.ncbi.nlm.nih.gov/. Accessed 3 Jul 2015
43. Oliaei F, Kriens D, Weber R, Watson A (2013) PFOS and PFC releases and associated pollution from a PFC production plant in Minnesota (USA). Environ Sci Pollut Res 20:1977–1992. https://doi.org/10.1007/s11356-012-1275-4
44. Peter KT, Tian Z, Wu C, Lin P, White S, Du B, McIntyre JK, Scholz NL, Kolodziej EP (2018) Using high-resolution mass spectrometry to identify organic contaminants linked to urban stormwater mortality syndrome in Coho salmon. Environ Sci Technol 52:10317–10327. https://doi.org/10.1021/acs.est.8b03287
45. Pluskal T, Castillo S, Villar-Briones A, Orešic M (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinform 11:395
46. Development Core Team R (2010) R: A language and environment for statistical computing. Austria, Vienna
47. Riediker S, Suter MJF, Giger W (2000) Benzene- and naphthalenesulfonates in leachates and plumes of landfills. Water Res 34:2069–2079. https://doi.org/10.1016/S0043-1354(99)00368-1
48. Royal Society of Chemistry ChemSpider. Royal Society of Chemistry. http://www.chemspider.com. Accessed 3 Jul 2015.
49. Ruff M, Mueller MS, Loos M, Singer HP (2015) 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 Res 87:145–154. https://doi.org/10.1016/j.watres.2015.09.017
50. Schollée JE, Schymanski EL, Avak SE, Loos M, Hollender J (2015) Prioritizing unknown transformation products from biologically-treated wastewater using high-resolution mass spectrometry, multivariate statistics, and metabolic logic. Anal Chem 87:12121–12129. https://doi.org/10.1021/acs.analchem.5b02905
51. Schwarzenbach RP, Escher BI, Fenner K, Hofstetter TB, Johnson CA, von Gunten U, Wehrli B (2006) The challenge of micropollutants in aquatic systems. Science 313:1072–1077. https://doi.org/10.1126/science.1127291
52. Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP, Hollender J (2014) Identifying small molecules via high resolution mass spectrometry: communicating confidence. Environ Sci Technol 48:2097–2098. https://doi.org/10.1021/es5002105
53. Schymanski EL, Singer HP, Longrée P, Loos M, Ruff M, Stravs MA, Ripollés Vidal C, Hollender J (2014) Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. Environ Sci Technol 48:1811–1818. https://doi.org/10.1021/es4044374
54. Schymanski EL et al (2015) Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. Anal Bioanal Chem 2:1–19. https://doi.org/10.1007/s00216-015-8681-7
55. Singer HP, Wössner AE, McArdell CS, Fenner K (2016) Rapid screening for exposure to "non-target" pharmaceuticals from wastewater effluents by combining HRMS-based suspect screening and exposure modeling. Environ Sci Technol 50:6698–6707. https://doi.org/10.1021/acs.est.5b03332
56. Sjerps RMA, Vughs D, van Leerdam JA, ter Laak TL, van Wezel AP (2016) Data-driven prioritization of chemicals for various water types using suspect screening LC-HRMS. Water Res 93:254–264. https://doi.org/10.1016/j.watres.2016.02.034
57. Slobodnik J, Mrafkova L, Carere M, Ferrara F, Pennelli B, Schüürmann G, von der Ohe PC (2012) Identification of river basin specific pollutants and derivation of environmental quality standards: a case study in the Slovak Republic. TrAC Trends Anal Chem 41:133–145. https://doi.org/10.1016/j.trac.2012.08.008
58. US EPA (2004) Final submission for hexamethoxymethylmelamine
59. Van De Steene JC, Stove CP, Lambert WE (2010) A field study on 8 pharmaceuticals and 1 pesticide in Belgium: removal rates in waste water treatment plants and occurrence in surface water. Sci Total Environ 408:3448–3453. https://doi.org/10.1016/j.scitotenv.2010.04.037
60. Verlicchi P, Al Aukidy M, Zambello E (2012) Occurrence of pharmaceutical compounds in urban wastewater: removal, mass load and environmental risk after a secondary treatment—a review. Sci Total Environ 429:123–155. https://doi.org/10.1016/j.scitotenv.2012.04.028
61. Wickham H (2009) ggplot2—elegant graphics for data analysis. Springer, New York. https://doi.org/10.1007/978-0-387-98141-3_1

Publisher's Note