EXPLORING THE POTENTIAL OF A GLOBAL EMERGING CONTAMINANT EARLY WARNING NETWORK THROUGH THE USE OF RETROSPECTIVE SUSPECT SCREENING WITH HIGH-RESOLUTION MASS SPECTROMETRY

Nikiforos A. Alygizakis[1,2†], Saer Samanipour[3†], Juliane Hollender[4,5], María Ibáñez[6], Sarit Kaserzon[7], Varvara Kokkali[8], Jan A. van Leerdam[9], Jochen F. Mueller[7], Martijn Pijnappels[10], Malcolm J. Reid[3], Emma L. Schymanski[4,11], Jaroslav Slobodnik[2], Nikolaos S. Thomaidis[1], Kevin V. Thomas[3,7]*

[1]Laboratory of Analytical Chemistry, Department of Chemistry, University of Athens, Panepistimiopolis Zografou, 15771 Athens, Greece

[2]Environmental Institute, s.r.o., Okružná 784/42, 972 41 Koš, Slovak Republic

[3]Norwegian Institute for Water Research (NIVA), Gaustadalléen 21, 0349 Oslo, Norway

[4]Eawag: Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland

[5]Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland

[6]Research Institute for Pesticides and Water, University Jaume I, Avda. Sos Baynat s/n, 12071 Castellón de la Plana, Spain

[7]Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, 20 Cornwall Street, Woolloongabba, Queensland, 4102 Australia

[8]Vitens Laboratory, Snekertrekweg 61, 8912 AA Leeuwarden, The Netherlands

[9]KWR Watercycle Research Institute, P.O. Box 1072, 3430 BB, Nieuwegein, The Netherlands

[10]Rijkswaterstaat, Ministry of Infrastructure and the Environment, Zuiderwagenplein 2, 8224 AD, Lelystad, The Netherlands

[11]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 7, Avenue des Hauts Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg

[†]Authors contributed equally.

*Corresponding author

Kevin V Thomas

Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, 20 Cornwall Street, Woolloongabba, Queensland, 4102 Australia.

Email: kevin.thomas@uq.edu.au

Phone: 0061 417287582

Manuscript details

Word count abstract: 225

Word count text: 6735

Keywords: suspect screening, high-resolution mass spectrometry, retrospective screening, early warning system, contaminants of emerging concern

Abstract

A key challenge in the environmental and exposure sciences is to establish experimental evidence of the role of chemical exposure in human and environmental systems. High resolution and accurate tandem mass spectrometry (HRMS) is increasingly being used for the analysis of environmental samples. One lauded benefit of HRMS is the possibility to retrospectively process data for (previously omitted) compounds that has led to the archiving of HRMS data. Archived HRMS data affords the possibility of exploiting historical data to rapidly and effectively establish the temporal and spatial occurrence of newly identified contaminants through retrospective suspect screening. We propose to establish a global emerging contaminant early warning network to rapidly assess the spatial and temporal distribution of contaminants of emerging concern in environmental samples through performing retrospective analysis on HRMS data. The effectiveness of such a network is demonstrated through a pilot study, where eight reference laboratories with available archived HRMS data retrospectively screened data acquired from aqueous environmental samples collected in 14 countries on 3 different continents. The widespread spatial occurrence of several surfactants (e.g. PEGs and C12AEO-PEGs), transformation products of selected drugs (e.g. gabapentin-lactam, metoprolol-acid, carbamazepine-10-hydroxy, omeprazole-4-hydroxy-sulphide, 2-benzothiazole-sulfonic-acid), and industrial chemicals (3-nitrobenzenesulfonate and bisphenol-S) was revealed. Obtaining identifications of increased reliability through retrospective suspect screening is challenging and recommendations for dealing with issues such as broad chromatographic peaks, data acquisition, and sensitivity are provided.

Introduction

One of the key challenges in the environmental and exposure sciences is to establish experimental evidence of the role of chemical exposure in human and environmental systems.[1,2] Our 'chemosphere' is continuously changing and most chemicals that are indexed in the Chemical Abstract Service (CAS) are not characterized with respect to their potential effects on human safety and environmental health.[3] Non-target analysis employing high-resolution mass spectrometers has been established over the past years as one of the key approaches for tackling this complexity. High resolution and accurate hybrid tandem mass spectrometers, such as time-of-flight and Orbitrap instruments have facilitated increased reliability in target analysis (using reference standards), enabled suspect screening (without reference standards) and screening for unknowns.[4-6] Substantial research effort has been placed on developing tools and workflows that expedite these three approaches, with the overall outcome that the contemporary analyst is able to obtain large amount of accurate mass data for a particular sample. For example, in 2013 the NORMAN Network of reference laboratories, research centres and related organisations for monitoring of emerging environmental substances (www.norman-network.net) organized a non-target screening collaborative trial employing target, suspect, and non-target workflows to identify substances in water samples.[7] This trial revealed that non-target techniques are in general substantially harmonized between practitioners and that although data processing can be time consuming and remains a major bottleneck, suspect screening approaches are very popular. However it recognized that *"better integration and connection of desired features into software packages, the exchange of target and suspect lists, and the contribution of more spectra from standard substances into (openly accessible) database"* are necessary for the technique to reach maturity.[4] The archiving of HRMS data also allows for data to be processed retrospectively, for example to investigate the occurrence of a newly identified compound or simply one that was not considered at the time of analysis.[8] This possibility has led to researchers working in this field to digitally archive data in preparation for future retrospective analysis and has even led to proposals for the establishment of data repositories, akin to environmental data banks, where digital information can be safely stored for future retrospective analysis.

Non-target HRMS full scan data allows the potential for rapid and cost-effective screening of the occurrence of newly identified contaminants in previously archived HRMS data; often referred to as retrospective analysis. Typically, it refers to the application of suspect screening workflows to archived data as reference standard measurements are not available for the analytical settings. Whilst retrospective analysis with HRMS in environmental sciences has been discussed for some time [7,8,9,10] there are few published studies that actually apply the approach[11,12]. As far as we are aware there have not been coordinated studies to investigate the spatial and temporal distribution of contaminants of emerging concern in environmental samples through performing retrospective analysis on HRMS data acquired using different instrumental platforms and data processing software. This has the potential to be an improved and effective strategy for establishing the extent of a newly identified contaminant's occurrence rather than the traditional approach of a new contaminant(s) being reported in the scientific literature and individual research groups subsequently validating targeted methods and reporting their own data. In order to test this hypothesis, a pilot study was performed where eight reference laboratories with available archived HRMS data were recruited with the goal of exploring the potential of a contaminant of emerging concern early warning network through the use of retrospective suspect screening employing HRMS. The pilot study was referred to as the NORMAN Early Warning System, abbreviated to NormaNEWS.[13]

Materials and Methods

Participants and samples

The participants of the NormaNEWS exercise (8 reference laboratories; Eawag, KWR, NIVA, QAEHS, RWS, UJI, UoA, and Vitens) submitted samples from 14 countries and 3 continents. In total 48 sets of data from the analysis of environmental samples were evaluated. Detailed information on sample matrix, sampling date, instrument type, chromatographic separation (flow, column, gradient programs, and solvents), mass spectrometric method (acquisition mode and calibration method) are presented in the "**Sample Information**" sheet in the supporting information (SI) excel spreadsheet. Further, a more detailed description of the samples and methods used are presented in the SI spreadsheet, including information on any previously published datasets.

A wide variety of environmental samples were included in this study. The majority of the samples were wastewater (effluent and influent), surface water, and groundwater samples. More than half of the samples (26 out of 48) were wastewater samples (mainly effluent wastewater samples). Wastewater sample data sets were from Switzerland (various locations)[14], Norway, Sweden, Finland, Denmark, Iceland, Spain, Greece, Mexico and Australia. Fifteen of the 48 samples were samples from ecologically important large rivers such as Danube (station JDS57 Bulgarian/Romanian boarders)[7] and Rhine[15], smaller rivers such as

Swiss rivers (Furtbach and Doubs)[16], Dutch rivers (Meuse and Vecht) and the Logan river in Australia. Four groundwater samples were included from Spain and the Netherlands. One primary sludge sample from the wastewater treatment plant (WWTP) in Athens (Greece)[17] as well as one seawater sample affected by treated wastewater[18] were also evaluated. Finally, two drinking water samples from Ridderkerk and Lekkerkerk in The Netherlands were included in the study. All the participants were asked to provide only the absolute intensity of the identified features that were blank subtracted in order to avoid the false positive identification.

Participating laboratories analyzed their samples using their own routines (i.e. sample preparation and data processing) for all the analytes included in the NormaNEWS suspect list ("**NormaNEWS compounds**" sheet in the SI, on the NORMAN Suspect Exchange and in the CompTox Chemistry Dashboard). No specific method (i.e. chromatographic, ion source, and polarity) was recommended to the participants. This was in order to test the applicability of this approach for the data generated via different methods. For these analyses, a wide range of mass analyzers as well as chromatographic conditions was employed by different participants ("**Sample Information**" sheet in the SI). All of the reported results were further examined, through a quality control assessment, to produce harmonized and comparable results (see section 'Quality control criteria'). Finally, each identified peak was assigned with an appropriate confidence level.[19] These quality assurance steps were deemed necessary for interpretation of the results.

NormaNEWS suspect list

The final chemical screening suspect list consisted of 156 analytes including: 74 surfactants i.e. PEGs, C12AEO-PEGs, glycol ether sulfates (GES), linear alkylbenzyl sulfonates (LAS), sulfophenyl alkyl carboxylic acids (SPACs), and fluorosurfactants (PFAS, from several classes); 54 pharmaceuticals and their transformation products (e.g. carbamazepine, carbamazepine-10-hydroxy, diltiazem, diltiazem-desacetyl, and diltiazem-N-desmethyl); 17 bisphenols; and finally 11 industrial chemicals. We considered the surfactants and the industrial chemicals as two separate families of compounds, even though a lot of surfactants may have industrial source. This distinction was made due to multiple sources for surfactants. The suspect list compounds (name, molecular formula, CAS number, SMILES, InChI and InChIKey), qualifier fragment ions and lipophilic properties (logP and log K_{OW}) are included in the SI "**NormaNEWS compounds**" sheet and are available online on the NORMAN Suspect Exchange and in the CompTox Chemistry Dashboard. The list was formed from compounds suggested by participants and typically included novel emerging substances with limited environmental occurrence as well as established widely occurring environmental contaminants (e.g. carbamazepine), which was included to assess the overall concept. A high number of the proposed substances were transformation products (TPs) of parent drugs that were detected through suspect and non-target screening from bio-transformation experiments. In these cases, parent drugs (e.g. citalopram and atenolol) were also included so that detection rates of the parent drugs and their TPs could be investigated. Novel surfactant compounds were also included to verify their wide-spread occurrence. In addition, the inclusion of a group of bisphenols as well as 3-nitrobenzenesulfonate, specified as an industrial chemical, were a result of non-target screening identifications. The purpose of the NormaNEWs suspect list is to provide a dynamic list of potential environmentally relevant and novel chemicals, which is enriched using expert knowledge and non-target analysis results as new data become available. The list is available at the NORMAN Suspect List Exchange (http://www.norman-network.com/?q=node/236) and on the CompTox Chemistry Dashboard (https://comptox.epa.gov/dashboard/chemical_lists/normanews).

Quality control criteria

All participants of NormaNEWS exercise were requested to submit their results together with their raw LC-HRMS chromatograms. Raw chromatograms were converted to mzML using ProteoWizard (msconvert module v.3.0.10827).[20] For data acquired in data-independent acquisition mode, different collision energy channels were separated using an in-house script (provided in the SI), while lock mass scans were removed. For data-dependent acquisition mode, MS/MS spectra were exported as text files (named "precursor mass retention time") and were removed from the mzML files. Treated mzML files were converted to CDF files, which are readable from various data analysis software including Bruker DataAnalysis v.4.3. (Bruker Daltonics, Bremen, Germany), which was used here.

The performance of the following parameters was checked; mass accuracy of HRMS, stability of chromatography and presence of qualifier fragments of identified compounds in higher collision energy. A combination of an expert panel and literature information was used in order to set the threshold of each quality control criterion.

The quality control step enabled us to minimize the effect of analyst expertise and the instrumentation on the final results given that the evaluation of the analysts and/or the instrumentation was not within the goals of this exercise. Therefore, the data points that did not meet the quality control criteria were excluded from the finally reported results.

RESULTS AND DISCUSSION

Quality control assessment

Quality control was performed to ensure that data were generated from well-calibrated instruments and that the data submitted were reliable. The first and most important step of the procedure was to check that the mass difference between the experimental and theoretical mass did not exceed ±5 mDa, which was considered the maximum tolerable mass error in the provided complex environmental samples.[21, 22] This was highly relevant in assessing the confidence level assigned to each identified analyte in the list.

The mass accuracy quality control is summarized in the SI "**QC_mass accuracy_ppm/ QC_mass accuracy_Da**" sheet and the results presented in Figure 1. According to the submitted datasets, Orbitrap mass analyzers showed better mass accuracy performance (absolute average mass error 0.55 mDa) comparing to other TOF instruments (absolute average mass error 0.91 mDa), based on successfully identified compounds. Mass errors are caused by the complexity of the samples, saturation of the detector (see section challenges and recommendations), and the instrument itself (i.e. the age and hardware). LC-HRMS data obtained using LTQ Orbitrap instruments showed lower mass accuracy (absolute average mass error 1.1 mDa) when compared with the LTQ Orbitrap XL (absolute average mass error 0.52 mDa), which showed lower mass accuracy in comparison with the QExactive. We further investigated the effect of instrumentation used on the observed mass accuracies through a non-parametric statistical test Kruskal-Wallis. [23]A Kruskal-Wallis *p* value > 0.01 indicated the rejection of null-hypothesis and statistical significance of the observed differences in the measured averaged masses. The method used to calibrate each instrument was also considered. LC-HRMS data obtained using a Bruker QTOF were calibrated by injecting the calibrant substance at the beginning of the chromatogram, while data from Waters QTOF (in both cases) were calibrated by lock-mass every 0.5 or 2 minutes (injecting, recording and recalibrating based on calibrant peaks appearing every 0.5/2 minutes). High mass accuracy is an extremely crucial parameter to achieve high quality results during the suspect analysis. Especially, high accuracy measurements enable a decreased number of false positive detections.

The chromatographic stability of the LC separation was also assessed. All participants submitted at least 3 datasets for evaluation. Retention time data from the same instrumental set-up (and same partner) were grouped together and the normalized standard deviations (NSD) of the retention times of the detected substances were calculated (retention times of the detected substances in seconds can be found in the SI "**QC_observed_ret.time_Minutes**" sheet). A criterion of the maximum tolerable NSD of 10% was adopted for accepting the detection of a single compound across samples in data coming from the same partner. The average normalized standard deviation of retention times in all samples was < 2% (Figure S1). The largest variability of 8.6 % was observed for analyte valsartan, whereas the lowest variability (<0.1%) was observed for acesulfame in samples from Netherlands, GES-07 in samples from Australia, and GES-09 and GES-06 in samples from Greece. Retention time stability was considered as another extremely important parameter, which has a direct effect on the identification confidence. The low deviation observed in all the submitted datasets indicated the high quality and reliability of the LC separation of the participating laboratories.

The third QC criterion related to the presence of qualifier ions (QI) in the MS/MS spectra (SI "**NormaNEWS compounds**" sheet). These ions are fragments of the parent ion and are observable at higher collision energy or even at low collision energy as in-source fragments. The criterion was set on the presence of the QIs as either an in-source fragment or at higher collision energy. The identification level of compounds that did not comply with the third QC criterion were regarded as questionable and were marked accordingly.[19] As these QIs proved to be a very efficient way of improving the confidence of the suspect hit, Top 3 fragments have now been extracted from all mass spectra submitted to MassBank.EU and also put on the NORMAN Suspect Exchange (direct download) and the CompTox Chemistry Dashboard Downloads (direct link) for community use. The QC stage was used to exclude the features that did not meet the previously set criteria, thus harmonization. Consequently, we have reported only the features that met these mentioned criteria.

Overview of the retrospective screening

PolyEthylene Glycol 09 (PEG-09) was the most frequently detected compound, being present in 41 out of the 48 samples (85%) analyzed. Several bisphenols, transformation products of perfluorooctane sulfonate, and the pharmaceutical omeprazole were not detected in any of the samples analyzed ("**Max. Absolute Intensity_counts**" sheet in the SI and Figures 2, XS, X1S, X2S). Series of surfactants, such as PEGs, C12AEO-PEGs, and GES, resulted in a higher detection frequency for compounds with masses varying between 400 and 600 Da compared to both smaller and larger molecules from the same families (Figure S2.A). Schymanski et al and Gago-Ferrero et al. have previously observed a similar trend for these surfactants.[14, 24] The observed trend may be explained by the efficient ionization of mid-size molecules compared to other compounds and potentially the fact that they are used as technical mixtures.[25] LAS had an average frequency of detection of around 50%. The largest measured LAS, in terms of mass (i.e. C14-LAS), were detected in only 4 samples out of 48 samples. Based on the estimated retention time for LAS-C14, we interpret that the chromatographic run times used by different partners were not sufficiently long to successfully detect this suspect analyte in the evaluated samples. Only 3 of the 5 suspect fluorinated surfactants were detected with perfluorooctane sulfonate (PFOS) having the highest detection frequency of ~ 35%. For industrial chemicals and pharmaceuticals, venlafaxine was the suspect analyte with the highest frequency of detection (68%), while several bisphenols were not detected in any of the samples. Additionally, we observed a higher occurrence frequency of the suspect analytes in the locations with higher population density such as Spain, Switzerland, and Greece compared to locations such as Scandinavia and Australia with lower population density, Figures 2 and S3. The observed trend was consistent across all the analyzed matrices. However, it should be noted that considering the limited data set for this pilot study, further interpretation of the spatial and temporal distribution of pollutants is not possible. The future implementation of this approach will provide larger datasets for comprehensive spatial and temporal assessment of CEC occurrence across the globe.

The presence of a large number of successfully detected surfactants and industrial chemicals in both wastewater influents, effluents, and surface waters suggests the wide spread occurrence of these CECs in the environment across the globe, Figure 2. Although modern wastewater treatment plants are to some extent equipped to remove these pollutants[26-29], the high production/consumption volumes of these chemicals used in households and industrial applications translates into their release into the environment. The environmental occurrence, fate and behavior of surfactants have been widely investigated, however more reliable environmental data for these pollutants are necessary.[30-32] Collective exercises such as NormaNEWS are therefore an important step forward towards producing a comprehensive and reliable database on the environmental occurrence of surfactants and/or other chemicals of emerging concern (CEC), which can be used for better understanding of their environmental fate and behavior. Furthermore, this exercise, through the provided QC criteria, metadata template (i.e. SI spreadsheet), provides all necessary information and guidelines for laboratories across the globe for the reliable detection, identification, and reporting of CECs in different environmental compartments.

Challenges and recommendations

For analysts to obtain high-confidence identifications through retrospective suspect screening they face several challenges. Here, recommendations for dealing with difficulties such as broad peaks, data acquisition, and sensitivity are provided in the following.

The presence of broad peaks in the chromatograms of complex samples is often caused by the physico-chemical properties of that compound and the selected chromatographic method is unavoidable. For example, the LAS surfactants that elute at the end of the gradient of a typical reverse phase chromatographic run result in characteristic broad peaks (Figure 3A). Many peak picking algorithms are unable to detect such broad peaks. Therefore, employing peak picking independent approaches[33, 34], prior knowledge of those analytes, and visualization tools, even though not comprehensive, may be useful in dealing with broad peaks.

Data-dependent acquisition is often used in non-target analysis. Certain limitations with data-dependent acquisition may potentially cause false identification of features due to its limitations. This acquisition mode isolates and provides MS/MS spectra of some of the most abundant ions per full scan. Even though this approach is the ideal acquisition mode during identification of peaks with the most abundant ions, this mode is not suitable for retrospective screening, due to the limited number of MS/MS spectra obtained. In case the peak of an environmentally relevant compound is not one of those most abundant ions, the MS/MS spectra of this chemical would not be recorded (Figure 3B). Therefore, confident identification of that peak would not be possible. As a solution, it is highly recommended that samples are injected in data-independent acquisition mode which is the ideal acquisition mode for retrospective screening. In data-independent acquisition, HRMS is recording full scan and MS/MS spectra without prior isolation of any mass. Therefore, all fragments (and fragments of fragments in case of in-source fragments) of all co-eluting compounds are recorded, resulting in complex but information-rich MS/MS spectra that requires adequate data processing tools for confident identification of features. However, to our knowledge this is the most effective acquisition method for the samples that are meant for retrospective analysis. As different compounds have different fragmentation behavior depending on the different collision energies, the use of multiple (e.g. low, medium, high) or ramped collision energies should be considered during acquisition of data for retrospective screening to cover as many compounds as possible. As different instruments have different settings and acquisition speeds, a compromise may need to be found to provide sufficient resolution in the full scan while obtaining as much fragmentation information as possible. Pilot studies such as these and the upload of corresponding suspect lists and fragment information to public resources greatly help exchange experience to find these ideal compromises for future investigations.

Another inherent concern about LC-HRMS data is sensitivity. Among other reasons, one possible case for non-detection of pollutants is that current HRMS instruments operated in full scan are sensitive depending on the frequency with which they acquire full scans.[35] This means that low abundant or poorly ionized chemicals are not detected in case HRMS instrument records full scans at a high frequency rate. For example, recording full-scans at low frequency (2 Hz) will enable the detection of more compounds in comparison with a higher frequency rate (i.e. 20 Hz). Therefore, the analysts should try to find a compromise between the sampling speed and the sensitivity required for the analyses. For the samples, that are meant to be analyzed via retrospective screening a lower sampling frequency is recommended given that under these conditions a higher sensitivity is achieved.

Substances at high concentration levels in extracts and/or having high ionization efficiency can often result in the detector becoming saturated (Figure 3C). In this case, the peak reaches a plateau, which makes peak picking and determination of exact mass and retention time very difficult. For example, surfactants such as PEGs and C12AEO-PEGs were affected by detector saturation due to their high concentrations in the evaluated samples. The mentioned uncertainties in the exact mass and retention time are caused by the fact that saturation reduces the mass accuracy of the measurements for certain instruments, which is of extreme importance when performing identification. However, increasing the mass extraction window may solve these issues. On the other hand, such less strict mass accuracy criterion may increase the likelihood of false positive detection.

Another open issue in mass spectrometry is related to structural isomers (Figure 3D). Isomers are structurally similar compounds with the same molecular formula (same mass and isotopic profile) and share very similar fragmentation. This happened in the case of the detection of bisphenol S in the surface waters of the Netherlands. Two peaks, with different retention times, with acceptable mass accuracy, isotopic fit and same qualifier ions seem to belong to two different isomers of bisphenol S. In such cases, deeper knowledge of fragmentation behavior and/or retention time prediction could help to identify the peak that belongs to the suspected substance. Ion ratio (ratio of the intensity of a fragment to the intensity of another fragment) can be also considered. However, this information should be carefully examined, because of ion suppression caused by high background signal produced by complex sample's matrix. Classes of substances such as the surfactants mentioned here also contain many structurally related substances that cannot be distinguished easily with mass spectrometry. These are now being grouped as "related substances" in the CompTox Chemistry Dashboard (see hyperlinks for the different surfactant classes throughout this manuscript) as a first step in working towards computational solutions to deal with the extremely complex challenge of chemical substances of Unknown or Variable Composition, Complex Reaction Products and Biological Materials (UVCBs).[36, 37] Finally, all the samples need to be analyzed both in positive and negative mode in order to cover a wider chemical space compared to only single polarity.

The early warning system and its potential

This exercise confirmed the high occurrence frequency of several surfactants (e.g. PEGs and C12AEO-PEGs), transformation products of selected drugs (e.g. gabapentin-lactam, metoprolol-acid, carbamazepine-10-hydroxy, omeprazole-4-hydroxy-sulphide, 2-benzothiazole-sulfonic-acid), and industrial chemicals such as 3-nitrobenzenesulfonate and bisphenol S. These chemicals are not typically included in target/suspect lists used for surface water monitoring programs. Subsequently, there are limited environmental occurrence data available for these pollutants.[38-40] This clearly demonstrates that an early warning network such as NormaNEWS enables the efficient and reliable detection and identification of novel CECs in different environmental compartments at both a temporal and spatial scale. Consequently, a reasonably large and diverse dataset on the environmental occurrence of novel CECs in different matrices has been generated during this pilot project. Clearly, this study was a proof of concept to test the applicability of such an approach to a diverse global dataset. Further development and larger global coverage is necessary in order to generate a dataset suitable for both environmental interpretation and policy making practices. Such a dataset provides an initial screen that can be used to inform contaminant prioritization exercises leading to further monitoring, fate and effect studies and subsequent risk assessment. Furthermore, given that the data are harmonized across a large number of laboratories and the confidence level of each identification is provided, the inherent reliability of each identification becomes more intuitive to non-experts. The purpose of this network activity would not be to replace ongoing targeted monitoring and screening programs, but to provide a robust and comprehensive complementary collaborative approach for informing the refinement of priority substance lists. This also shows the great potential for screening much larger lists in the future, although the manual verification of the results is still a demanding task. More computationally efficient methods will be needed before this can be expanded to potentially lists of tens of thousands of substances.

The NormaNEWS pilot was performed using a very simple approach where all participants manually submitted data on their CECs of interest in order to create a suspect screening list for the collaborative exercise. This enabled researchers to easily obtain additional data on the CECs that they are particularly interested in. Future lists could be generated by a number of different approaches including from open resources, such as massbank.eu. As highlighted recently by Schymanski and Williams,[36] open resources will be instrumental in defining the evolution of suspect screening. The community-wide sharing of CECs through the exchange of suspect lists (e.g. the NORMAN Suspect Exchange and the Chemistry Dashboard lists) as well as tentatively and unequivocally identified spectra and sharing the related fragments is therefore key to the success of a global early warning network. Also key will be the willingness of the scientific community to share their HRMS data in an open MS format (e.g. mzML[41], mzXML[42], and netCDF[43]).

The Global Natural Products Social Molecular Networking (GNPS; http://gnps.ucsd.edu/) provides a vision as to how global collaboration and social cooperation can be used to address major scientific challenges in the sharing and community curation of MS data.[44] Taking inspiration from GNPS, we propose that HRMS data are made available (through a virtual repository and with necessary metadata) in order to facilitate living data along with periodic automated re-analysis of data (i.e. with updates to the suspect list or the addition of new data sets). Ideally, this repository will be easily accessible through a web-application and free of the aforementioned challenges. The environmental and exposure sciences currently lag behind other fields, such as proteomics[45], metabolomics[46] and natural product research[47] in globally collaborating and sharing data through open/social platforms in order to revolutionize the way data are processed to achieve significant outcomes. We acknowledge that not all the data tools are currently in place to make our proposal a reality, however progress is being made in this area[33, 34, 48, 49]. For example, within the NORMAN Network (http://www.norman-network.net/) there is an initiative to develop a digital sample freezing platform. A global emerging contaminant early warning network based on adopting the successful practices of other similar networks will play a pivotal role in identifying chemicals using HRMS that has the potential to possess significant outcomes in protecting human and environmental health.

SUPPORTING INFORMATION

Text, tables and figures with detailed information on experimental methods, QA/QC procedures and supplemental data (xls, PDF).

ACKNOWLEDGEMENTS

Support by the NORMAN Network of reference laboratories, research centers and related organizations for the monitoring of emerging environmental substances facilitated this work under its Joint Program of Activities (http://www.norman-network.net) and is greatly appreciated. We are especially grateful to Antony J. Williams (US EPA, CompTox Chemistry Dashboard) for integrating the NormaNEWS Suspect List, other NORMAN lists and the fragment data into the Dashboard, as well as for all his efforts in cross-linking the surfactants. JH, HPS and ELS acknowledge the efforts of Philipp Longrée, Heinz Singer and other colleagues at Eawag who contributed to the analysis of the Eawag data. SS and KVT were supported in part by the Research Council of Norway (Project Number 243720). KVT, JFM and SK gratefully acknowledge the financial support of the Queensland Department of Health. ELS was supported in part by the EU FP7 Project SOLUTIONS (Grant Number 603437).

REFERENCES

1. Kortenkamp, A.; Faust, M.; Scholze, M.; Backhaus, T., Low-level exposure to multiple chemicals: reason for human health concerns? *Environ Health Perspect* **2007,** *115 Suppl 1*, 106-114.

2. Pleil, J. D., Categorizing Biomarkers of the Human Exposome and Developing Metrics for Assessing Environmental Sustainability. *Journal of Toxicology and Environmental Health, Part B* **2012,** *15*, (4), 264-280.

3. Muir, D. C. G.; Howard, P. H., Are There Other Persistent Organic Pollutants? A Challenge for Environmental Chemists. *Environmental Science & Technology* **2006,** *40*, (23), 7157-7166.

4. Rager, J. E.; Strynar, M. J.; Liang, S.; McMahen, R. L.; Richard, A. M.; Grulke, C. M.; Wambaugh, J. F.; Isaacs, K. K.; Judson, R.; Williams, A. J.; Sobus, J. R., Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring. *Environ Int* **2016,** *88*, 269-280.

5. Andra, S. S.; Austin, C.; Patel, D.; Dolios, G.; Awawda, M.; Arora, M., Trends in the application of high-resolution mass spectrometry for human biomonitoring: An analytical primer to studying the environmental chemical space of the human exposome. *Environ Int* **2017,** *100*, 32-61.

6. Leendert, V.; Van Langenhove, H.; Demeestere, K., Trends in liquid chromatography coupled to high-resolution mass spectrometry for multi-residue analysis of organic micropollutants in aquatic environments. *TrAC Trends in Analytical Chemistry* **2015,** *67*, 192-208.

7. Schymanski, E. L.; Singer, H. P.; Slobodnik, J.; Ipolyi, I. M.; Oswald, P.; Krauss, M.; Schulze, T.; Haglund, P.; Letzel, T.; Grosse, S.; Thomaidis, N. S.; Bletsou, A.; Zwiener, C.; Ibanez, M.; Portoles, T.; de Boer, R.; Reid, M. J.; Onghena, M.; Kunkel, U.; Schulz, W.; Guillon, A.; Noyon, N.; Leroy, G.; Bados, P.; Bogialli, S.; Stipanicev, D.; Rostkowski, P.; Hollender, J., Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. *Anal Bioanal Chem* **2015,** *407*, (21), 6237-55.

8. Krauss, M.; Singer, H.; Hollender, J., LC-high resolution MS in environmental analysis: from target screening to the identification of unknowns. *Anal Bioanal Chem* **2010,** *397*, (3), 943-51.

9. Hernandez, F.; Sancho, J. V.; Ibanez, M.; Abad, E.; Portoles, T.; Mattioli, L., Current use of high-resolution mass spectrometry in the environmental sciences. *Anal Bioanal Chem* **2012,** *403*, (5), 1251-64.

10. Gomez-Ramos, M. M.; Ferrer, C.; Malato, O.; Aguera, A.; Fernandez-Alba, A. R., Liquid chromatography-high-resolution mass spectrometry for pesticide residue analysis in fruit and vegetables: screening and quantitative studies. *J Chromatogr A* **2013,** *1287*, 24-37.

11. Polgar, L.; Garcia-Reyes, J. F.; Fodor, P.; Gyepes, A.; Dernovics, M.; Abranko, L.; Gilbert-Lopez, B.; Molina-Diaz, A., Retrospective screening of relevant pesticide metabolites in food using liquid chromatography high resolution mass spectrometry and accurate-mass databases of parent molecules and diagnostic fragment ions. *J Chromatogr A* **2012,** *1249*, 83-91.

12. Chiaia-Hernandez, A. C.; Krauss, M.; Hollender, J., Screening of lake sediments for emerging contaminants by liquid chromatography atmospheric pressure photoionization and electrospray ionization coupled to high resolution mass spectrometry. *Environ Sci Technol* **2013,** *47*, (2), 976-86.

13. Gomez-Ramos Mdel, M.; Perez-Parada, A.; Garcia-Reyes, J. F.; Fernandez-Alba, A. R.; Aguera, A., Use of an accurate-mass database for the systematic identification of transformation products of organic contaminants in wastewater effluents. *Journal of chromatography. A* **2011,** *1218*, (44), 8002-12.

14. Schymanski, E. L.; Singer, H. P.; Longree, P.; Loos, M.; Ruff, M.; Stravs, M. A.; Ripolles Vidal, C.; Hollender, J., Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. *Environ Sci Technol* **2014,** *48*, (3), 1811-8.

15. Ruff, M.; Mueller, M. S.; Loos, M.; Singer, H. P., 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* **2015,** *87*, 145-54.

16. Moschet, C.; Wittmer, I.; Simovic, J.; Junghans, M.; Piazzoli, A.; Singer, H.; Stamm, C.; Leu, C.; Hollender, J., How a complete pesticide screening changes the assessment of surface water quality. *Environ Sci Technol* **2014,** *48*, (10), 5423-32.

17. Gago-Ferrero, P.; Borova, V.; Dasenaki, M. E.; Tauhomaidis Nu, S., Simultaneous determination of 148 pharmaceuticals and illicit drugs in sewage sludge based on ultrasound-assisted extraction and liquid chromatography-tandem mass spectrometry. *Anal Bioanal Chem* **2015,** *407*, (15), 4287-97.

18. Alygizakis, N. A.; Gago-Ferrero, P.; Borova, V. L.; Pavlidou, A.; Hatzianestis, I.; Thomaidis, N. S., Occurrence and spatial distribution of 158 pharmaceuticals, drugs of abuse and related metabolites in offshore seawater. *Sci Total Environ* **2016,** *541*, 1097-105.

19. Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J., Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* **2014,** *48*, (4), 2097-8.

20. Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P., A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **2012,** *30*, (10), 918-20.

21. Zedda, M.; Zwiener, C., Is nontarget screening of emerging contaminants by LC-HRMS successful? A plea for compound libraries and computer tools. *Anal Bioanal Chem* **2012,** *403*, (9), 2493-502.

22. Kaufmann, A.; Walker, S., Evaluation of the interrelationship between mass resolving power and mass error tolerances for targeted bioanalysis using liquid chromatography coupled to high-resolution mass spectrometry. *Rapid Commun Mass Spectrom* **2013,** *27*, (2), 347-56.

23. Breslow, N., A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrika* **1970,** *57*, (3), 579-594.

24. Gago-Ferrero, P.; Schymanski, E. L.; Bletsou, A. A.; Aalizadeh, R.; Hollender, J.; Thomaidis, N. S., Extended Suspect and Non-Target Strategies to Characterize Emerging Polar Organic Contaminants in Raw Wastewater with LC-HRMS/MS. *Environ Sci Technol* **2015,** *49*, (20), 12333-41.

25. Mazzoni, M.; Rusconi, M.; Valsecchi, S.; Martins, C. P.; Polesello, S., An on-line solid phase extraction-liquid chromatography-tandem mass spectrometry method for the determination of perfluoroalkyl acids in drinking and surface waters. *J Anal Methods Chem* **2015,** *2015*, 942016.

26. Prats, D.; Ruiz, F.; Vazquez, B.; M., R.-P., Removal of anionic and nonionic surfactants in a wastewater treatment plant with anaerobic digestion. A comparative study. *Water Res* **1997,** *31*, (8), 1925-1930.

27. Aboulhassan, M. A.; Souabi, S.; Yaacoubi, A.; Baudu, M., Removal of surfactant from industrial wastewaters by coagulation flocculation process. *Int J Environ Sci Tech* **2006,** *3*, (4), 327-332.

28. Gonzalez, S.; Petrovic, M.; Barcelo, D., Removal of a broad range of surfactants from municipal wastewater--comparison between membrane bioreactor and conventional activated sludge treatment. *Chemosphere* **2007,** *67*, (2), 335-43.

29. Luo, Y.; Guo, W.; Ngo, H. H.; Nghiem, L. D.; Hai, F. I.; Zhang, J.; Liang, S.; Wang, X. C., A review on the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment. *Sci Total Environ* **2014,** *473-474*, 619-41.

30. Jackson, M.; Eadsforth, C.; Schowanek, D.; Delfosse, T.; Riddle, A.; Budgen, N., Comprehensive review of several surfactants in marine environments: Fate and ecotoxicity. *Environ Toxicol Chem* **2016,** *35*, (5), 1077-86.

31. Jardak, K.; Drogui, P.; Daghrir, R., Surfactants in aquatic and terrestrial environment: occurrence, behavior, and treatment processes. *Environ Sci Pollut Res Int* **2016,** *23*, (4), 3195-216.

32. Ying, G.-G., Fate, behavior and effects of surfactants and their degradation products in the environment. *Environment International* **2006,** *32*, (3), 417-431.

33. Samanipour, S.; Langford, K.; Reid, M. J.; Thomas, K. V., A two stage algorithm for target and suspect analysis of produced water via gas chromatography coupled with high resolution time of flight mass spectrometry. *J Chromatogr A* **2016,** *1463*, 153-61.

34. Samanipour, S.; Baz-Lomba, J. A.; Alygizakis, N. A.; Reid, M. J.; Thomaidis, N. S.; Thomas, K. V., Two stage algorithm vs commonly used approaches for the suspect screening of complex environmental samples analyzed via liquid chromatography high resolution time of flight mass spectroscopy: A test study. *J Chromatogr A* **2017,** *1501*, 68-78.

35. Acena, J.; Stampachiacchiere, S.; Perez, S.; Barcelo, D., Advances in liquid chromatography-high-resolution mass spectrometry for quantitative and qualitative environmental analysis. *Anal Bioanal Chem* **2015,** *407*, (21), 6289-99.

36. Schymanski, E. L.; Williams, A. J., Open Science for Identifying "Known Unknown" Chemicals. *Environ Sci Technol* **2017,** *51*, (10), 5357-5359.

37. Williams A.; Grulke, C. M.; McEachran A; Richard, A.; Jolley R; Dunne J; Edmiston E; J, E. Comptox Chemistry Dashboard: Web-based data integration hub for environmental chemistry and toxicology data. https://www.slideshare.net/AntonyWilliams?utm_campaign=profiletracking&utm_medium=sssite&utm_source=ssslideview

38. Beretsou, V. G.; Psoma, A. K.; Gago-Ferrero, P.; Aalizadeh, R.; Fenner, K.; Thomaidis, N. S., Identification of biotransformation products of citalopram formed in activated sludge. *Water Res* **2016,** *103*, 205-14.

39. Nika, M. C.; Bletsou, A. A.; Koumaki, E.; Noutsopoulos, C.; Mamais, D.; Stasinakis, A. S.; Thomaidis, N. S., Chlorination of benzothiazoles and benzotriazoles and transformation products identification by LC-HR-MS/MS. *J Hazard Mater* **2017,** *323*, (Pt A), 400-413.

40. Christophoridis, C.; Nika, M. C.; Aalizadeh, R.; Thomaidis, N. S., Ozonation of ranitidine: Effect of experimental parameters and identification of transformation products. *Sci Total Environ* **2016,** *557-558*, 170-82.

41. Martens, L.; Chambers, M. C.; Sturm, M.; Kessner, D.; Levander, D.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P.-A.; Deutsch, E. W., mzML - a Community Stadard for Mass Spectometry Data. *Mol Cell Proteomics* **2011,** *10*, (1).

42. Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R., A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* **2004,** *22*, (11), 1459-66.

43. Erickson, B., ANDI MS standard finalized. *Anal Chem* **2000,** *72*, (3), 103 A–103 A.

44. Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W. T.; Crusemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderon, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C. C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C. C.; Yang, Y. L.; Humpf, H. U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; P, C. A. B.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodriguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P. M.; Phapale, P.; Nothias, L. F.; Alexandrov, T.; Litaudon, M.; Wolfender, J. L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D. T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Muller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. O.; Pogliano, K.; Linington, R. G.; Gutierrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N., Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* **2016,** *34*, (8), 828-837.

45. Sturm, M.; Bertsch, A.; Gropl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O., OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics* **2008,** *9*, 163.

46. Uppal, K.; Walker, D. I.; Liu, K.; Shuzhao, L.; G., Y.-M.; P., J. D., Computational Metabolomics: A Framework for the Million Metabolome. *Chem Res Toxicol* **2016,** *29*, (12), 1956-1975.

47. Allard, P. M.; Genta-Jouve, G.; Wolfender, J. L., Deep metabolome annotation in natural products research: towards a virtuous cycle in metabolite identification. *Curr Opin Chem Biol* **2017,** *36*, 40-49.

48. Samanipour, S.; Reid, M. J.; Thomas, K. V., Statistical Variable Selection: An Alternative Prioritization Strategy during the Nontarget Analysis of LC-HR-MS Data. *Anal Chem* **2017,** *89*, (10), 5585-5591.

49. Samanipour, S.; Reid, M.; Baek, K.; Thomas, K. V., Combining a deconvolution and a universal library search algorithm for the non-target analysis of data independent LC-HRMS spectra. *Environ Sci Technol* **2018**, In Press.

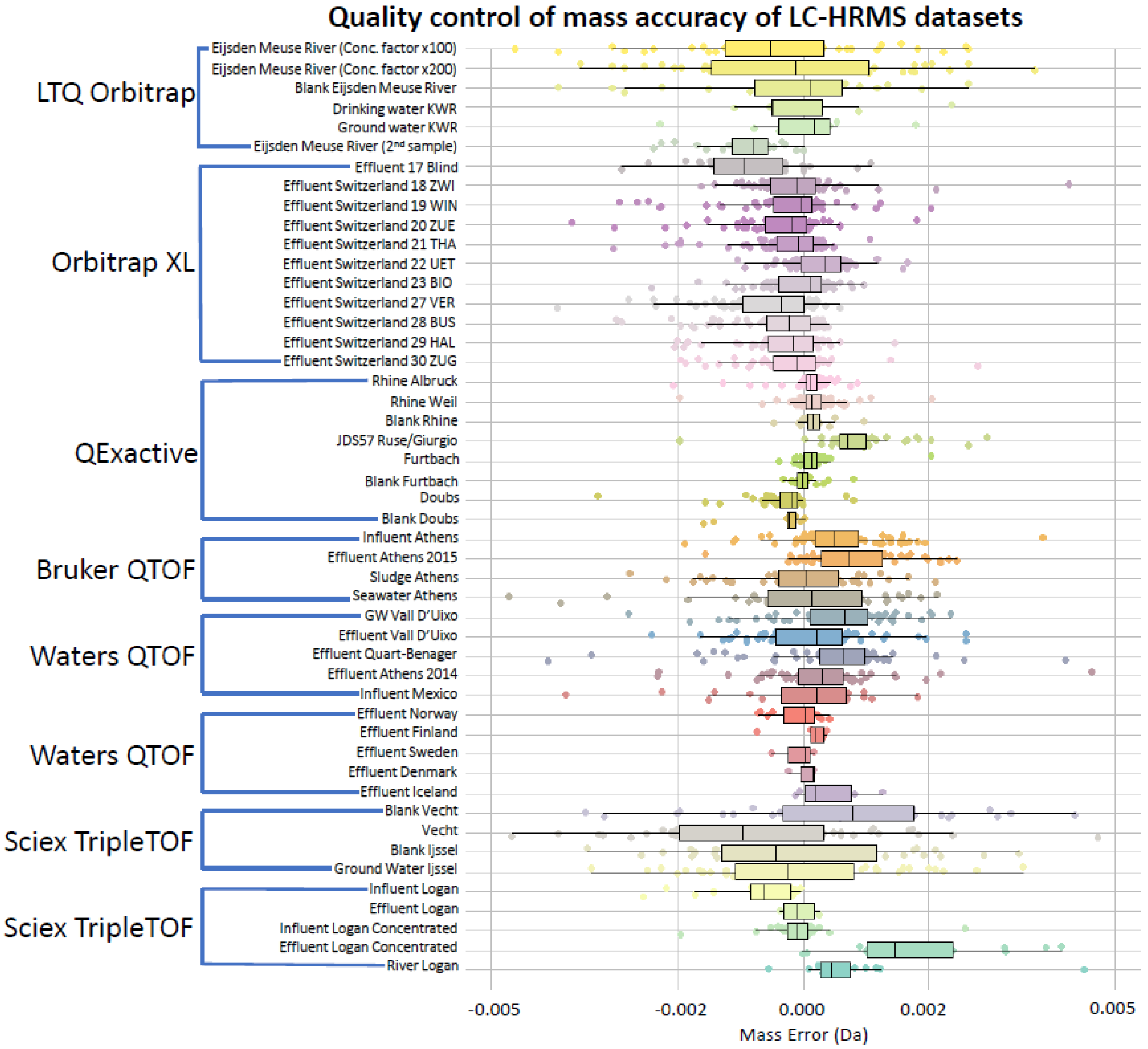

Figure 1. Quality control of mass accuracy of the submitted datasets based on the identified compounds. Type of mass analyzer, calibration type of the mass analyzer as well as other factors (age of equipment, scan sampling rate of the detector) affect the performance and the quality of the results.

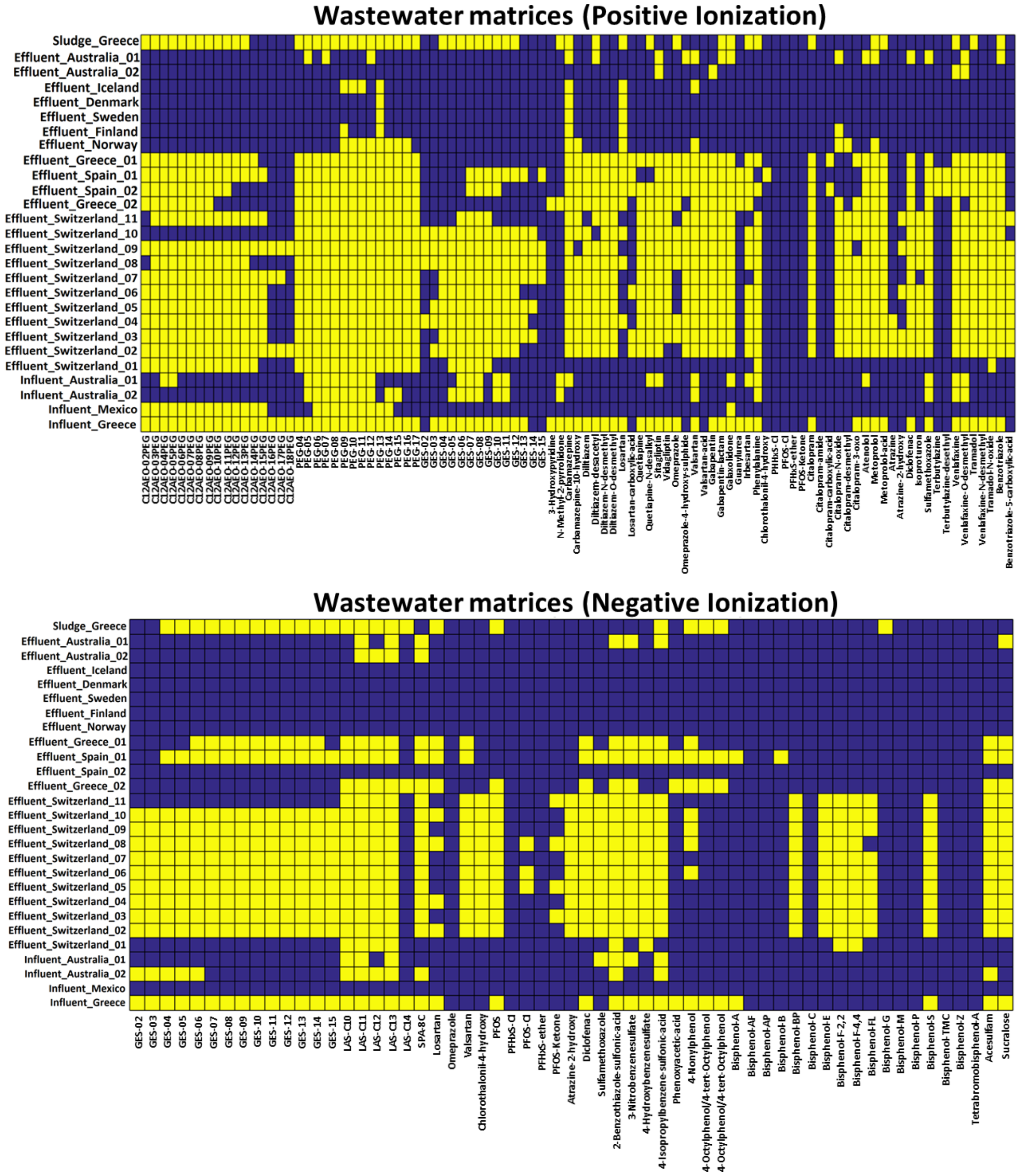

Figure 2. Heat map showing the occurrence of the selected substances in the retrospectively screened samples (primary sludge from WWTP of Athens, Greece, effluent wastewater samples from Australia, Iceland, Spain, Denmark, Sweden, Finland, Norway, Greece and Switzerland) and influent wastewater samples (Australia, Mexico, Greece) for positive and negative ionization. Successfully identified compounds are marked in yellow.

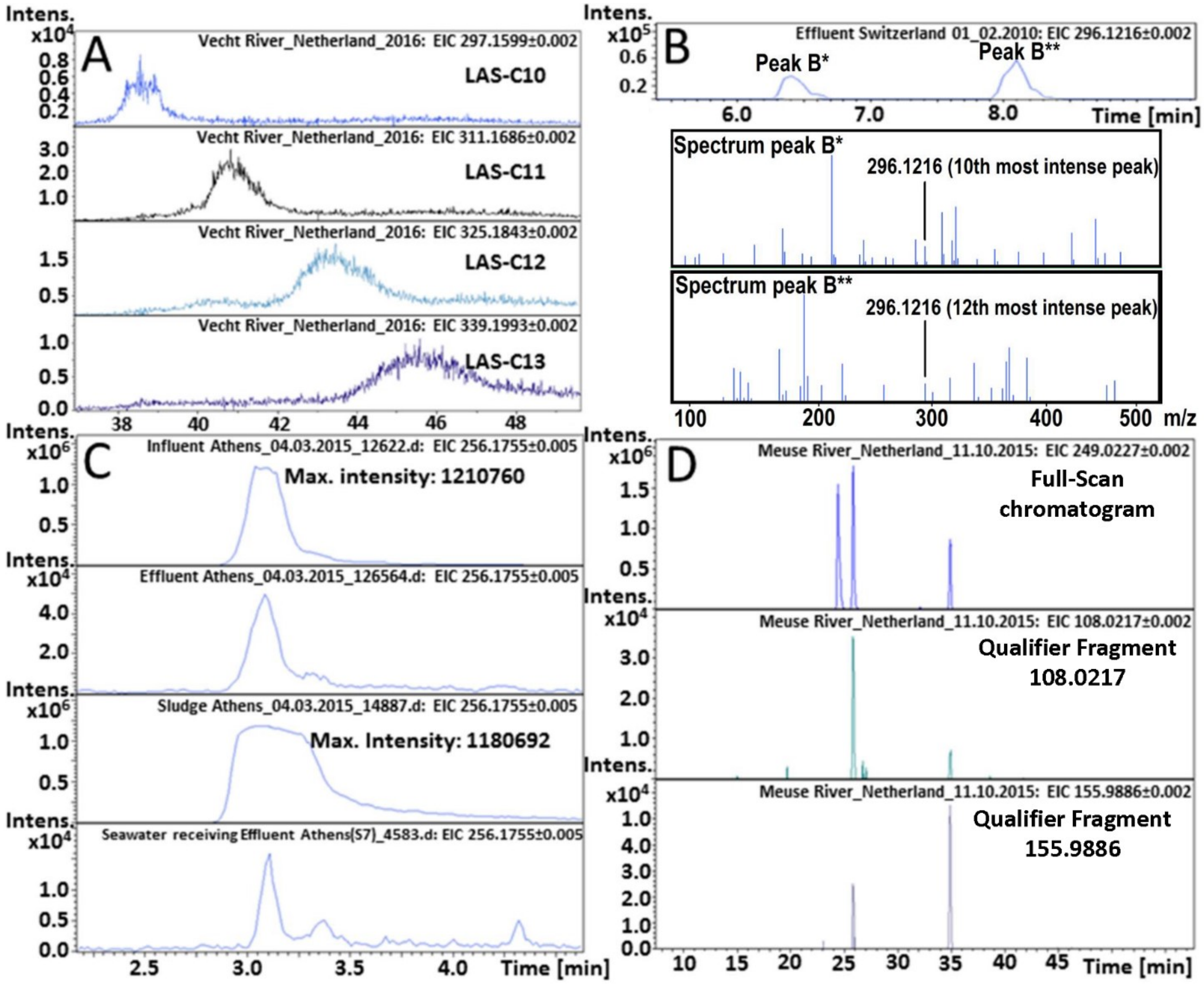

Figure 3. Challenges faced during evaluation of the results; A. Broad peaks of Linear alkylbenzene sulphonate (LAS) surfactants makes peak-picking challenging, B. Missing fragmentation information (MS/MS) of compound of interest decreases identification confidence, because data-dependent acquisition is capable to capture MS/MS only for preselected or few most abundant spectral peaks per scan (marked with red rhombus). Peaks are mass accuracy and isotopic profile consistent but not abundant enough so that MS/MS spectra have not been acquired (case of Quetiapine-N-desalkyl), C. Saturation of detector deteriorates mass accuracy, affects peak-picking and causes quantification mistakes when quantification is done by maximum intensity and not by peak area (case of PEG-05), D. Bisphenol S isomers cannot be distinguished, because in both cases qualifier fragment ions (m/z 108.0217 and 155.9886) are present in both peaks in the high collision energy channel.

EXPLORING THE POTENTIAL OF A GLOBAL EMERGING CONTAMINANT EARLY WARNING NETWORK THROUGH THE USE OF RETROSPECTIVE SUSPECT SCREENING WITH HIGH-RESOLUTION MASS SPECTROMETRY

Nikiforos A. Alygizakis[1,2†], Saer Samanipour[3†], Juliane Hollender[4,5], María Ibáñez[6], Sarit Kaserzon[7], Varvara Kokkali[8], Jan A. van Leerdam[9], Jochen F. Mueller[7], Martijn Pijnappels[10], Malcolm J. Reid[3], Emma L. Schymanski[4,11], Jaroslav Slobodnik[2], Nikolaos S. Thomaidis[1], Kevin V. Thomas[3,7]*

[1]Laboratory of Analytical Chemistry, Department of Chemistry, University of Athens, Panepistimiopolis Zografou, 15771 Athens, Greece

[2]Environmental Institute, s.r.o., Okružná 784/42, 972 41 Koš, Slovak Republic

[3]Norwegian Institute for Water Research (NIVA), Gaustadalléen 21, 0349 Oslo, Norway

[4]Eawag: Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland

[5]Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland

[6]Research Institute for Pesticides and Water, University Jaume I, Avda. Sos Baynat s/n, 12071 Castellón de la Plana, Spain

[7]Queensland Alliance for Environmental Health Sciences (QAEHS), 39 Kessels Road, Coopers Plains, Queensland, 4108 Australia

[8]Vitens Laboratory, Snekertrekweg 61, 8912 AA Leeuwarden, The Netherlands

[9]KWR Watercycle Research Institute, P.O. Box 1072, 3430 BB, Nieuwegein, The Netherlands

[10]Rijkswaterstaat, Ministry of Infrastructure and the Environment, Zuiderwagenplein 2, 8224 AD, Lelystad, The Netherlands

[11]Universitè Du Luxembourg, Luxembourg Centre for Systems Biomedicine, 6, avenue du Swing

L-4367 Belvaux, Luxembourg

[†]Authors contributed equally.

*Corresponding author

Kevin V Thomas

Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, 39 Kessels Road, Coopers Plains, Queensland, 4108 Australia.

Email: kevin.thomas@uq.edu.au

Phone: 0061 417287582

Supplementary spreadsheet

Investigated substances, sample information, experimental set-up and identifications are all summarized in the supplementary spreadsheet. The spreadsheet consists of 5 tabs; "*Sample information*", *"NormaNEWS compounds*", "Max. Absolute Intensity_counts ", "*QC_mass accuracy_Da*", "*QC_mass accuracy_ppm*" and "QC_observed_ret.time_Minutes ".

Sample information tab contains information about the samples (location, sampling date, matrix type), instrument type, model and chromatographic conditions (column, flow, gradient solvents and program). For each dataset, mzML files are attached.

In *NormaNEWS compounds* tab are the investigated substances (full name, short name and molecular formula), chemical identifiers (CAS, SMILES, InChi and InChIKey), preferable ionization type for detection of the compounds, fragments qualifying the identity of the compounds and predicted LogP (source: ACD/Labs) and logKow (source: EPI Suite)

Max. Absolute Intensity_counts tab contains all the identifications. Compounds are represented as rows while samples are represented as columns. If the chemical was detected in the sample, the maximum intensity value is marked otherwise is marked as N.D. (standing for Not Detected). If no data were available to evaluate the presence or absence of the compound (e.g. no data are available for negative ionization), then the cell contents is marked as NA (standing for Not Available). Red color in the tab corresponds unequivocal molecular formula while dark red color corresponds to mass of interest.

QC_mass accuracy_Da and *QC_mass accuracy_ppm* contain the mass accuracy error in Dalton and ppm respectively. The mass accuracy was used as quality control parameter of the chromatograms.

QC_observed_ret.time_Minutes contains the observed experimental retention time in minutes. Datasets coming from the same instrument and obtained under the same experimental conditions should have consistent stable retention time for the identified substances. Chromatographic drift was also considered as another important quality control parameter.

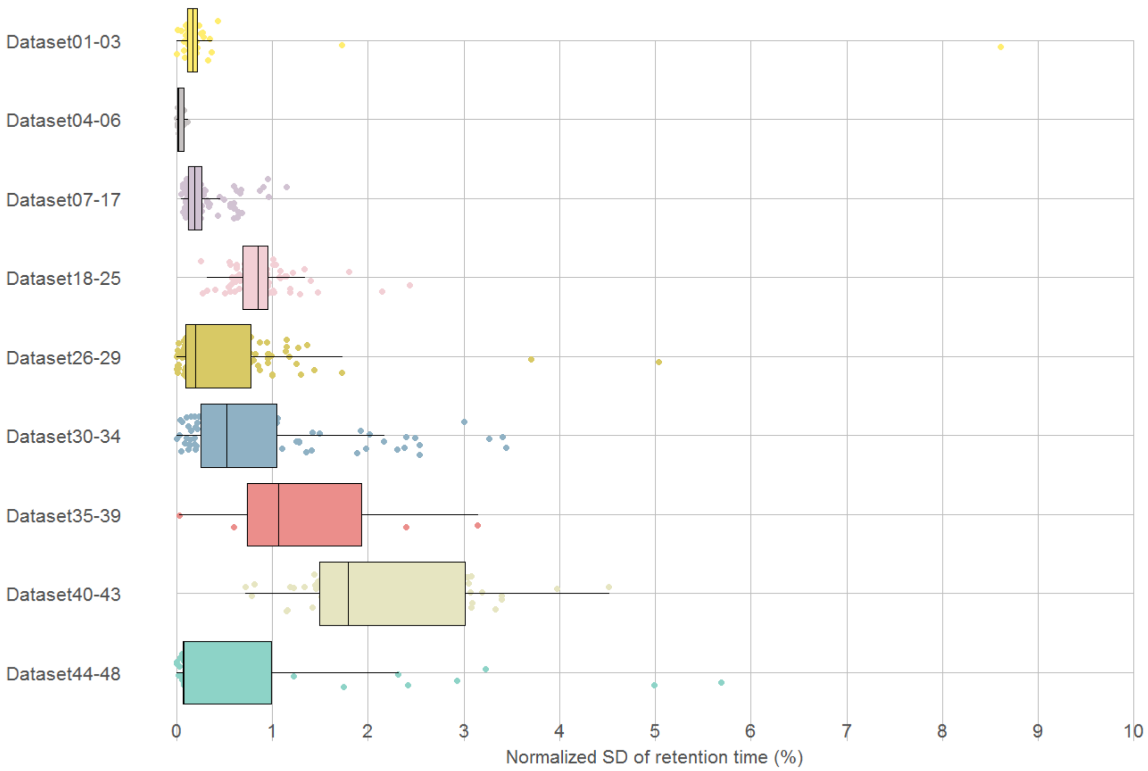

Figure S1. Quality control of chromatographic stability of the submitted datasets

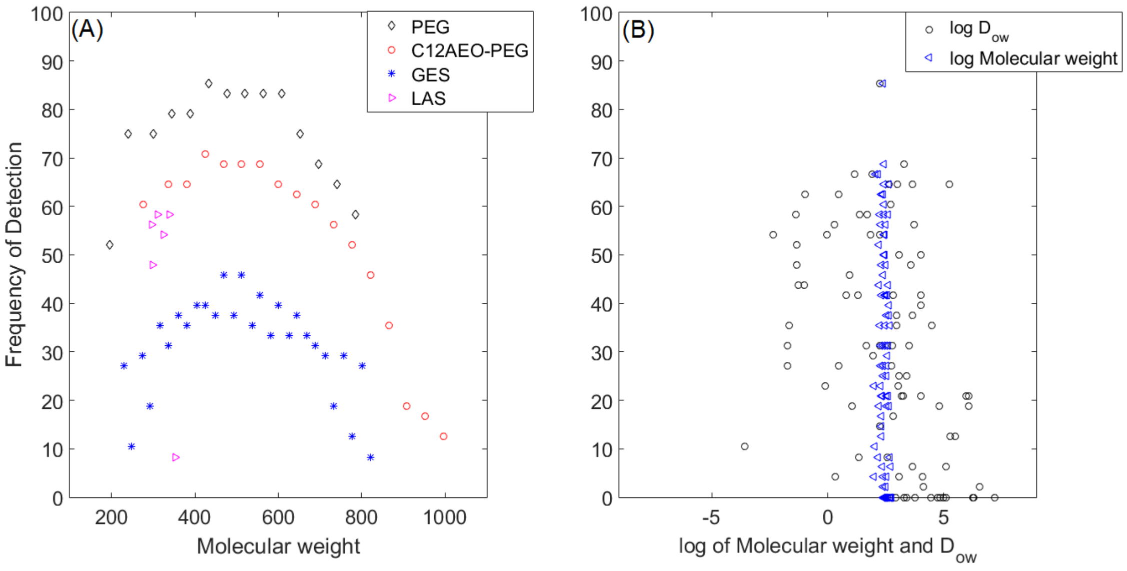

Figure S2.A. Frequency of detection of surfactants against molecular weight; S2B. Frequency of detection of identified substances against exact mass and D_{ow}.

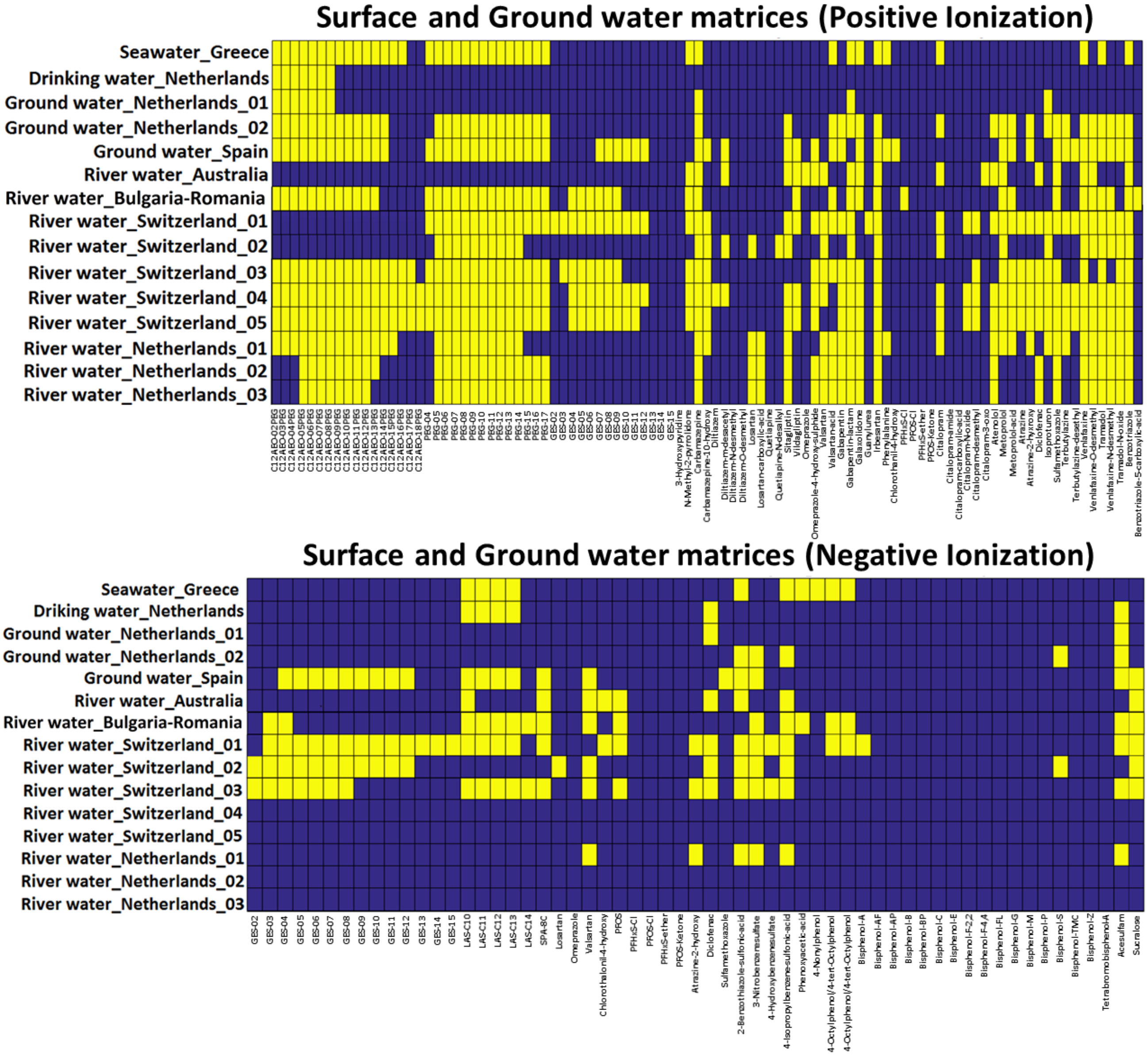

Figure S3. Heat map showing the occurrence of the selected substances in the retrospectively screened samples (seawater receiving effluent wastewater, drinking water, ground water from the Netherlands and Spain and river water from Switzerland, the Netherlands, Danube river water from Romanian-Bulgarian borders) for positive and negative ionization. Successfully identified compounds are marked in yellow.

Scripts for QA/QC

```
# Script by Nikiforos Alygizakis, University of Athens, 01/12/2016

# Load the functions part


file<-"F:/Black Sea/SEAWATER_POS/S7_BlackSea.mzXML"


mzxml<-read.mzXML(file)

information<-getinfo(mzxml)


#Case of data-independent

lowcollision<-4

highcollision<-25

ms1<-removescans(mzxml,scansORtime=information$scan[as.numeric(information$CE)==highcollision],time=F)

mse<-removescans(mzxml,scansORtime=information$scan[as.numeric(information$CE)==lowcollision],time=F)


write.mzXML(ms1,paste("MS1",strsplit(file,"/")[[1]][length(strsplit(file,"/")[[1]])]))

write.mzXML(mse,paste("MS2",strsplit(file,"/")[[1]][length(strsplit(file,"/")[[1]])]))



#Case of data-dependent

file_data_dependent<-"F:/Black Sea/SEAWATER_POS/S7_BlackSea_DataDependent5precursors.mzXML"


mzxml<-read.mzXML(file_data_dependent)

mzxml_no_MSn<-removeMSn(mzxml)

write.mzXML(mzxml_no_MSn,paste("MS1",strsplit(file_data_dependent,"/")[[1]][length(strsplit(file_data_dependent,"/")[[1]])]))


##Functions
```

```
#'@title This function reads mzXML files
#'
#'@description Reads a mzXML file and returns a mzXML list object in the global environment.
#'
#'@usage read.mzXML(filename)
#'@param filename The directory in the hard drive that the mzXML files is stored.
#'
#'@details This functions reads a mzXML file and stores it as a list in the variables global environment.
#'
#'@return
#'Returns a list object, which contains the following elements;
#'\item{header}{Stores header of <mzXML> section containing information about namespace and schema file location.}
#'\item{parentFile}{Path to all the ancestor files. Stored as XML.}
#'\item{dataProcessing}{Description of any data manipulation. Stored as XML.}
#'\item{msInstrument}{General information about the MS instrument. Stored as XML.}
#'\item{scan}{List of Mass Spectra scans. Each element of the list contain the following elements;}
#'\item{peaks}{ peak intensities of the scan}
#'\item{mass}{ masses (m/z) corresponding to \code{peaks}. Vectors \code{mass} and \code{peaks} have the same length.}
#'\item{num}{ scan number}
#'\item{parentNum}{ scan number of parent scan in case of recursively stored scans (\code{msLevel>1})}
#'\item{msLevel}{ Level 1 means MS1, while level 2 means MS2, etc.}
#'\item{scanAttr}{ Other useful information, such as retention time, polarity, collision energy, total ion current}
#'\item{maldi}{ acquisition dependent properties of a MALDI experiment (optional)}
#'\item{scanOrigin}{ name of parent file}
#'\item{precursorMz}{ information about the precursor ion}
#'\item{nameValue}{ properties of the scan not included elsewhere}
#'
```

```
#'@author
#'Codes maintained by Nikiforos Alygizakis <nalygizakis@chem.uoa.gr>
#'
#'@examples
#'#Donot run
#'library("peakTrams")
#'sample<-read.mzXML(filename=c:\R_working_directory\sample.mzXML)
#'
#' @references
#'Definition of \code{mzXML} format:
#'\url{http://tools.proteomecenter.org/mzXMLschema.php}
#' @references
#'Documentation of \code{mzXML} format:
#'\url{http://sashimi.sourceforge.net/schema_revision/mzXML_2.1/Doc/mzXML_2.1_tutorial.pdf}
#' @references
#'More Documentation of \code{mzXML} format:
#'  \url{http://sashimi.sourceforge.net/software_glossolalia.html}
#'
#'@export
read.mzXML<-function(filename)
{
  Paste = function(...) paste(..., sep="", collapse="")

  strtrunc = function(Str,Sub) {
    lp = attr(regexpr(paste(".*",Sub,sep=""),Str),'match.length')
    return( substring(Str, 1, lp) )
    #y = unlist(strsplit(Str,Sub)) # other way of doing it
    #return( paste(y[-length(y)], sub,  sep="", collapse="") )
  }
```

```
 fregexpr = function(pattern, filename)
 { # similar to gregexpr but operating on files not strings
  buf.size=1024
  n  = file.info(filename)$size
  pos = NULL
  fp = file(filename, "rb")
  for (d in seq(1,n,by=buf.size)) {
    m = if (n-d>buf.size) buf.size else n-d
    p = gregexpr(pattern, readChar(fp, m))[[1]]
    if(p[1]>0) pos=c(pos, p+d-1)
  }
  close(fp)
  if (is.null(pos)) pos=-1
  return (pos)
 }
```

```
 new.mzXML = function(){
  object = list(
    header         = NULL, # required - list    - Path to all the ancestor files (up to the native acquisition file) used to generate the current XML instance document.
    parentFile     = NULL, # required - list    - Path to all the ancestor files (up to the native acquisition file) used to generate the current XML instance document.
    dataProcessing = NULL, # required - list    - Description of any manipulation (from the first conversion to mzXML format until the creation of the current mzXML instance document) applied to the data.
    msInstrument   = NULL, # optional - element - General information about the MS instrument.
    separation     = NULL, # optional - element - Information about the separation technique, if any, used right before the acquisition.
    spotting     = NULL, # optional - element - Acquisition independent properties of a MALDI experiment.
    scan         = vector(mode="list")
  )
  class(object) <- "mzXML"
  return(object)
}
#------------------------------
# define XML handler function
#------------------------------
mzXMLhandlers <- function()
{
  #--------------------------------------------------------
  # local variables
  #--------------------------------------------------------
  obj      = new.mzXML() # create new mzXML object
  iScan    = 0
  ParentID = vector(mode="integer")
  sha1     = vector(mode="list", length=2) # optional - element - sha-1 sums
  sha1[1] <- sha1[2] <- 0
  # Optional attributes that might come with a scan that will be stored
  OptScanAttr = c("polarity", "scanType", "centroided", "deisotoped",
          "chargeDeconvoluted", "retentionTime", "ionisationEnergy",
          "collisionEnergy", "cidGasPressure", "totIonCurrent")

  #------------------------------
  # local functions
  #------------------------------
  ToString = function(x, indent = 1)
```

```
{ # converts content of a node to a string
  if (is.null(x)) return(NULL);
  spaces = if (indent>0) Paste(rep("  ", indent)) else ""
  Name = xmlName(x, TRUE)
  val  = xmlValue(x)
  if (Name=="text") return( Paste(spaces, val, "\n") )
  if (!is.null(xmlAttrs(x))) {
    att = paste(names(xmlAttrs(x)), paste("\"", xmlAttrs(x),
                                        "\"", sep = ""), sep = "=", collapse = " ")
    att = paste(" ", att, sep="")
  } else att = ""
  chl = ""
  for (i in xmlChildren(x)) chl = Paste(chl, ToString(i, indent+1))
  if (chl=="") Str = Paste(spaces, "<" , Name, att, "/>\n")
  else Str = Paste(spaces, "<" , Name, att, ">\n", chl, spaces, "</", Name, ">\n")
  return(Str)
}

CatNodes = function(x,Name, indent = 2)
{ # concatinate strings of several nodes
  Str=NULL
  for (y in xmlElementsByTagName(x, Name))
    Str = paste(Str, ToString(y,indent), sep="")
  return(Str)
}

read.mzXML.scan = function(x)
{ # process scan section of mzXML file
  if (is.null(x)) return(NULL)
if (xmlName(x) != "scan") return(NULL)
scanOrigin <- precursorMz <- nameValue <- maldi <- mass <- peaks <- NULL
att        = xmlAttrs(x)
num        = as.integer(att["num"])
msLevel    = as.integer(att["msLevel"])
peaksCount  = as.integer(att["peaksCount"]) # Total number of m/z-intensity pairs in the scan
msk        = names(att) %in% OptScanAttr
if (sum(msk)==0) scanAttr = ""
else {
  scanAttr = paste( names(att[msk]), paste("\"", att[msk],
                                    "\"", sep = ""), sep = "=", collapse = " ")
  scanAttr = paste(" ", scanAttr, sep="")
}
maldi      = ToString(x[["maldi"]])
scanOrigin  = CatNodes(x, "scanOrigin", 3)
nameValue   = CatNodes(x, "nameValue", 3)
precursorMz = CatNodes(x, "precursorMz", 3)
precursorMz = gsub("\n      " , " " , precursorMz)
for (y in xmlElementsByTagName(x, "scan"))
  ParentID[as.integer(xmlAttrs(y)["num"])] <<- num
y          = x[["peaks"]]
att        = xmlAttrs(y)
peaks      = xmlValue(y) # This is the actual data encoded using base64
precision   = att["precision"] # nr of bits used by each component (32 or 64)
byteOrder   = att["byteOrder"] # Byte order of the encoded binary information (must be network)
pairOrder   = att["pairOrder"] # Order of the m/z intensity pairs (must be m/z-int
endian      = if(byteOrder=="network") "big" else "little"
if(precision=="32") size=4
else if(precision=="64") size=8
```

```
  else stop("read.mzXML.scan: incorrect precision attribute of peaks field")
  #if (pairOrder!="m/z-int")
  #warning("read.mzXML.scan: incorrect pairOrder attribute of peaks field")
  if (peaksCount>0) {
    p = base64decode(peaks, "double", endian=endian, size=size)
    np = length(p) %/% 2
    if (np != peaksCount)
      warning("read.mzXML.scan: incorrect 'peakCount' attribute of 'peaks' field: expected ",
              peaksCount, ", found ", np, "  ",(3*((nchar(peaks)*size)/4))/2, " (scan #",num,")")
    dim(p)=c(2, np)
    mass =p[1,]
    peaks=p[2,]
  }
  #x$children=NULL; # needed to capture the header
  #header <<- toString(x)
  return( list(mass=mass, peaks=peaks, num=num, parentNum=num,
               msLevel=msLevel, scanAttr=scanAttr, maldi=maldi,
               scanOrigin=scanOrigin, precursorMz=precursorMz, nameValue=nameValue) )
}

#---------------------------------------------------------
# the instructions how to parse each section of mzXML file
#---------------------------------------------------------
list(
  mzXML  = function(x, ...) {
    y = x[["sha1"]]
    sha1[1]    <<- if (!is.null(y)) xmlValue(y) else 0
    x$children =  NULL
    obj$header <<- toString(x,terminate=FALSE)
```

```
      NULL
    },

    msRun = function(x, ...) {
      y = x[["sha1"]]
      sha1[2]             <<- if (!is.null(y)) xmlValue(y) else 0
      obj$msInstrument   <<- ToString(x[["msInstrument"]],2)
      obj$separation     <<- ToString(x[["separation"]],2)
      obj$spotting       <<- ToString(x[["spotting"]],2)
      obj$parentFile     <<- CatNodes(x,"parentFile")
      obj$dataProcessing <<- CatNodes(x,"dataProcessing")
      NULL
    },

    scan  = function(x, ...) {
      iScan <<- iScan+1
      obj$scan[[iScan]] <<- read.mzXML.scan(x)
      x$children=NULL
      x
    },

    data = function() {
      if (is.null(obj$header)) NULL
      else list(mzXML=obj, ParentID=ParentID, sha1=sha1)
    }
  ) #end of list of handler functions
} # done with local functions

#---------------------------------
```

```
# begining of read.mzXML function

#---------------------------------

library(XML)

library(digest)

library(caTools)

if (!is.character(filename)) stop("read.mzXML: 'filename' has to be a string")

if (length(filename)>1) filename = paste(filename, collapse = "")  # combine characters into a string


sha1File = digest(filename, algo="sha1", file=TRUE)

x = xmlTreeParse(file=filename, handlers=mzXMLhandlers(),

                addAttributeNamespaces=TRUE) $ data()

if (is.null(x)) # is this file a mzXML file ?

  stop("read.mzXML: This is not mzXML file");

mzXML    = x$mzXML

sha1Read = x$sha1


# sort scans into correct order; find parent numbers of recursive nodes

n = length(mzXML$scan)

NumID = integer(n)

for (i in 1:n) {

  NumID[i] = mzXML$scan[[i]]$num

  mzXML$scan[[i]]$scanOrigin = paste("<scanOrigin parentFileID='",sha1File,

                                  "' num='",NumID[i],"'/>\n", sep="");

}


i<-1

rt<-c()

for(i in 1:length(mzXML$scan)){
  rt[i]<-as.numeric(strsplit(strsplit(strsplit(mzXML$scan[[i]]$scanAttr,
"retentionTime=[\"]PT")[[1]][2],"[\"]")[[1]][1],"S")[[1]][1])
  }

  mzXML$scan = mzXML$scan[ order(rt) ]
  for (i in 1:n)
    if(!is.na(x$ParentID[i])) mzXML$scan[[i]]$parentNum = x$ParentID[i]
  else x$ParentID[i] = mzXML$scan[[i]]$parentNum
#  mzXML$scan = mzXML$scan[ order(x$ParentID) ]

  ## read sha1 section
  n = sum(as.integer(lapply(sha1Read, is.character))) # how many sha1 were found
  if( n>0 ) {
    ## sha1 - sha-1 sum for this file (from the beginning of the file up to
    ## (and including) the opening tag of sha1
    if (is.null(sha1Read[[1]])) sha1Read[[1]]=sha1Read[[2]]
    sha1Pos = fregexpr("<sha1>", filename) + 6 # 6 = length("<sha1>")
    for(i in n) { # multiple sha1 sections are possible
      sha1Calc = digest(filename, algo="sha1", file=TRUE, length=sha1Pos[i]-1)
      if (sha1Read[[i]]!=sha1Calc)
        warning("Stored and calculated Sha-1 sums do not match (stored '",
                sha1Read[[i]],"'; calculated '", sha1Calc,"')")
    }
  }

  # strip mzXML terminator from header section
  mzXML$header = gsub("/>", ">\n", mzXML$header)
  mzXML$header = gsub("^ +", "", mzXML$header)
  # Remove incorrect "-quotes inserted in 2.10.0
  mzXML$header = gsub("[\u0093\u0094\u201C\u201D]", '"', mzXML$header)
  # add info about parent file (the file we just read)
#  mzXML$parentFile = Paste(mzXML$parentFile, "    <parentFile filename='file://",
#                          filename, "' fileType='processedData' fileSha1='", sha1File, "'/>\n")
  return( mzXML )
}


#'Gets retention time and number of peaks of full scans of a mzXML list
#'
#'Takes in a raw sample and returns a data frame with retention time of each full scan
#'@param sample mzXML list created from read.mzXML function
#'@return A data frame with number of scan, with retention time of each full scan, mslevel, number of
spectral peaks and in case
#'of MS/MS full scan precursor mass and precurson intensity.
#'
#'@examples
#'sample_mzXML<-
read.mzXML(list.files(paste(find.package(package="peakTrams"),"data",sep="/"),pattern = ".mzXML",
full.names = TRUE))
#'getrt(sample=sample_mzXML)
#'
#'@author Nikiforos Alygizakis <nalygizakis@chem.uoa.gr>
#'
#'@export
getinfo<-function(sample){
  numscan<-sample$scan[[1]]$num

  info<-data.frame(scan=numscan:length(sample[[5]]),timeofscan=0)
  for(numscan in 1:c(length(sample[[5]])-numscan+1)){
   if(length(strsplit(try(sample[[5]][[numscan]][[6]], silent=T),"Error")[[1]])!=2){
      info$timeofscan[numscan]<-sample[[5]][[numscan]][[6]]

      info[numscan,2]<-as.numeric(strsplit(strsplit(info[numscan,2],split="S")[[1]][1],split="PT")[[1]][2])

      #  info$basePeakMz[numscan]<-
sprintf("%.5f",sample$scan[[i]]$mass[which.max(sample$scan[[numscan]]$mass)])

      #  info$basePeakIntensity[numscan]<-
as.numeric(sprintf("%.0f",max(sample$scan[[numscan]]$peaks)))

    }

  }

  info$timeofscan<-as.numeric(info$timeofscan)


  numscan<-sample$scan[[1]]$num

  for(numscan in 1:c(length(sample[[5]])-numscan+1)){

    if(length(strsplit(try(sample[[5]][[numscan]][[6]], silent=T),"Error")[[1]])!=2){

      info$mslevel[numscan]<-(sample[[5]][[numscan]][[5]])

      info$numofpeaks[numscan]<-length(sample[[5]][[numscan]][[1]])

      info$CE[numscan]<-strsplit(strsplit(sample$scan[[numscan]]$scanAttr, "collisionEnergy=[\"]")[[1]][2],
"[\"]")[[1]][1]

    }

  }

  info$precursor<-NA

  info$precursorIntensity<-NA


  i<-1

  for(i in 1:length(info[,1])){

    if(length(strsplit(try(sample[[5]][[numscan]][[6]], silent=T),"Error")[[1]])!=2){

      if(info$mslevel[i]!=1){

        info$precursor[i]<-as.numeric(strsplit(strsplit(sample$scan[[i]]$precursorMz,"
</precursorMz>\n")[[1]][1],">   ")[[1]][2])

        info$precursorIntensity[i]<-as.numeric(strsplit(sample$scan[[i]]$precursorMz,"[\"]")[[1]][2])

      }
    }
  }
  sprintf("Done")

  info<-info[info$timeofscan!=0,]
  return(info)
}

#'Removes selected full scans from a mzXML list object
#'
#'Takes as input an object which was created by read.mzXML function
#'and returns an object without selected scans passed in scansORtime argument.
#'In case time is set to TRUE then scanORtime should be a vector of two elements containing
#'retention time in minutes. The selected full scans with retention time within this interval will
#'be deleted from the mzXML list.
#'@param mzXML file produced from read.mzXML function
#'@param scansORtime Selected scans (or scans with retention time if time=TRUE) to be removed
#'@param time Logical value TRUE or FALSE
#'@return a mzXML list without selected full scans
#'@author Nikiforos Alygizakis <nalygizakis@chem.uoa.gr>
#'@export
removescans<-function(mzXML=blank_HILIC,scansORtime=c(17,25),time=TRUE){

  if(scansORtime[2]=="end" & time==FALSE) scansORtime[2]<-max(getinfo(mzXML)$scan)
  if(scansORtime[2]=="end" & time==TRUE) scansORtime[2]<-max(getinfo(mzXML)$timeofscan)/60
  if(scansORtime[1]=="beginning" & time==FALSE) scansORtime[1]<-min(getinfo(mzXML)$scan)
  if(scansORtime[1]=="beginning" & time==TRUE) scansORtime[1]<-min(getinfo(mzXML)$timeofscan)/60
  scansORtime<-as.numeric(scansORtime)
  scansORtime2<-scansORtime
```

```
 info<-getinfo(mzXML)
 k<-which.min(abs(info$timeofscan-scansORtime[2]*60))
 if(info$mslevel[which.min(abs(info$timeofscan-scansORtime[2]*60))]!=1 & k!=length(info[,1])){
  while(info$mslevel[k]!=1) {
   k <- k+1
   scansORtime[2]<-info$timeofscan[k]/60
  }
  k<-k-1
  cat("Ending point was set at", paste(round(c(info$timeofscan[k]/60),4)), "because given ending retention time", scansORtime2[2] ,"corresponds to scan at MS2 level","\n")
 }


 u<-which.min(abs(info$timeofscan-scansORtime[1]*60))
 if(info$mslevel[which.min(abs(info$timeofscan-scansORtime[1]*60))]!=1){
  info<-getinfo(mzXML)
  while(info$mslevel[u]!=1) {
   u <- u-1
   scansORtime[1]<-info$timeofscan[u]/60
  }
  u<-u-1
  cat("Beginning point was set at", paste(round(c(info$timeofscan[u]/60),4)), "because given ending retention time", scansORtime2[1] ,"corresponds to scan at MS2 level","\n")
 }

 if(!is.na(info$mslevel[which.min(abs(info$timeofscan-scansORtime[2]*60))+1]!=1)){
  if(info$mslevel[which.min(abs(info$timeofscan-scansORtime[2]*60))]==1 & info$mslevel[which.min(abs(info$timeofscan-scansORtime[2]*60))+1]!=1) k<-which.min(abs(info$timeofscan-scansORtime[2]*60))-1
  }
  if(!is.na(info$mslevel[which.min(abs(info$timeofscan-scansORtime[1]*60))+1]!=1)){
    if(info$mslevel[which.min(abs(info$timeofscan-scansORtime[1]*60))]==1 &
info$mslevel[which.min(abs(info$timeofscan-scansORtime[1]*60))+1]!=1) u<-
which.min(abs(info$timeofscan-scansORtime[1]*60))-1
  }



  if(time==TRUE){
    info<-getinfo(mzXML)[u:k,]
    stayordelete<-c(rep(TRUE,length(mzXML[[5]])))
    stayordelete[info$scan]<-FALSE
  } else {
    stayordelete<-c(rep(TRUE,length(mzXML[[5]])))
    stayordelete[scansORtime]<-FALSE
  }

  new_sample<-list()
  new_sample<-mzXML[1:4]
  new_sample$scan<-mzXML[[5]][c(stayordelete)]
  attr(new_sample, "class") = "mzXML"



  i<-1
  for(i in 1:length(new_sample$scan)) new_sample[[5]][[i]]$num<-i


  return(new_sample)
}
```

```
#'Removes all MSn scan events from a mzXML list
#'
#'Takes as input an object which was created by read.mzXML function
#'and returns an object without the MSn scans.
#'@param sample mzXML list object
#'@return sample mzXML list object without MS/MS spectra
#'@author Nikiforos Alygizakis <nalygizakis@chem.uoa.gr>
#'@export
removeMSn<-function(sample){

  i<-1; removescanevents<-c();
  for(i in 1:length(sample$scan)) removescanevents[i]<-sample$scan[i][[1]]$msLevel

  removescanevents2<-c(); i<-1;
  for(i in 1:length(removescanevents))  if(removescanevents[i]!=1) removescanevents2[i]<-i



  new_sample<-list()
  new_sample<-sample[1:4]
  new_sample$scan<-sample[[5]][-removescanevents2[!is.na(removescanevents2)]]

  attr(new_sample, "class") = "mzXML"

  i<-1
  for(i in 1:length(new_sample$scan)) new_sample[[5]][[i]]$num<-i

  return(new_sample)
}
```