

Prioritization of Organic Compounds for Suspect Screening in (Drinking) Water and Biota

Dürig W.^{1*}, Tröger R.¹, Andersson P.L.², Rybacka A.², Fischer S.³, Wiberg K.¹, Ahrens L.¹

¹Dept. of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Box 7050, SE-750 07 Uppsala, Sweden

²Dept. of Chemistry, Umeå University, SE- 901 87 Umeå, Sweden

³Swedish Chemicals Agency, Box 2, SE-172 13 Sundbyberg, Sweden

*corresponding author:

e-mail: wiebke.durig@slu.se

Abstract

Suspect and non-target screening are powerful tools for identification of organic compounds using high resolution mass spectrometry (HRMS). However, there is a lack of smart suspect screening lists for the identification of organic compounds which are persistent, have high mobility in the environment and are potentially bioaccumulative and toxic. In this study, we created a prioritization strategy for top candidates among ~32,000 compounds including pesticides, pharmaceuticals, flame retardants and many other chemicals based on their physiochemical characteristics (e.g. $\log K_{ow}$, $\log K_{oc}$), predicted environmental fate characteristics (e.g. biodegradability, bioconcentration factor), human toxicity, emission sources and emission quantity for (drinking) water and biota. For charged compounds (18% of the list), a $\log D$ value was estimated based on the $\log K_{ow}$ at $pH 7$ to account for different behavior of charged compounds in the environment. Three lines of specialization were developed, one for water and two for biota (LC- and GC-HRMS, respectively). For biota, the driving factor for achieving a high score was high mobility towards biota and bioaccumulation potential, whereas for water, the mobility in the aquatic environment and human toxicity was of most significant importance. Typically, halogenated compounds ended up in the “(drinking) water-” (26,4%) and “biota-GC-HRMS” (19,4%) suspect lists due to their high bioaccumulation factor. In summary, we present three suspect screening lists developed for anthropogenic compounds that may end up in water and biota providing new opportunities to screen for environmentally relevant compounds by HRMS.

Keywords: Suspect screening; prioritization of compounds; physiochemical properties; environmental fate; emerging pollutants

1. Introduction

The introduction of conventions and directives has successfully phased out high risk organic pollutants.

However, the chemicals and product markets constantly introduce new compounds for replacement. The CAS registrySM contains more than 124,000,000 organic and inorganic substances with CAS numbers, of which 347,020 substances are regulated worldwide¹, and approximately 145,000 substances are pre-registered in REACH². The Swedish Chemicals Agency (KemI) has registered over 130,000 substances with CAS numbers. On a daily basis, ~50 new regulated substances and 60-75 patent citations are added to the CAS database³. Only a small fraction of these compounds is regulated; for example, the Stockholm Convention comprises 30 compounds/groups⁴ and the Water Framework Directive (WFD) comprises 33 priority compounds/groups⁵. However, the fate, bioaccumulation potential and toxicity to wildlife and humans of emerging compounds are largely unknown. Moreover, it is challenging to assess all these compounds, due to limitations of analytical methods and capacity.

Recent suspect and non-target screening studies⁶⁻¹⁰ successfully identified some emerging compounds and developed identification techniques. A well implemented approach to identify a broad range of compounds in environmental monitoring and assessment is suspect screening¹¹⁻¹⁴. Nowadays, suspect lists are mainly dominated by known, highly produced and/or trend compounds. In general, these pre-set suspect screening lists focus on certain compound groups or are based on previous identified compounds¹⁵⁻¹⁸. Moreover, the environmentally relevant charge of compounds is rarely considered. A systematic approach to develop suspect lists for monitoring and screening purposes in different environments, like water and biota, is desired.

In this study, a flexible tool was developed for the development of suspect screening lists, which can be used for screening of environmental samples for organic compounds using high resolution mass spectrometry (HRMS). The organic compounds are ranked based on their physiochemical properties, environmental fate characteristics, emission source, emission quantity range, and toxicity, while the criteria of the ranking can be easily adjusted and applied for different matrices. Three suspect screening lists were created, one focusing on water and two on biota (LC- and GC-HRMS, respectively).

2. Materials and Methods

2.1. Databases

The main database of organic compounds was created by merging three databases including chemical structures (SMILES) information (US EPA¹⁹, Farmaceutiska specialiteter i Sverige (FASS), and the Noman list of emerging substances). Merging these databases and excluding duplicates resulted in a list of about 32,000 compounds including pesticides, pharmaceuticals, flame retardants, industrial additives and many more chemicals and compound groups. The above-mentioned databases were chosen due to their completeness of information (number of compounds, SMILES notation, and CAS numbers) and relevance for the Swedish environment. Depending on the final purpose of the suspect list, other databases may be selected as a starting point.

2.2. Chemical Descriptors

All compounds in the final database were characterized by empirical data or estimated chemical descriptor data calculated in EPI SuiteTM 20, MarvinView 15.10.12.0²¹ and the On-line CHEMical database and Modeling environment²² from their SMILE structures. The descriptors comprised physicochemical properties ($\log K_{ow}$, $\log K_{oc}$, $\log K_{oa}$ and $\log S_w$), predicted environmental fate characteristics (biodegradation and $\log BCF$), predicted numerical *in vitro* response toxicity data for several receptors (ER= estrogen receptor, AR= androgen receptor and TR= thyroid hormone receptor), and index values (0-5) from the SPIN database²³ characterizing emission sources and emission quantity ranges (EI_{air} , EI_{water} , EI_{soil} , $EI_{sewage\ treatment}$, $EI_{consumer}$, $EI_{occupational}$ and QI = quantity index). These descriptors were selected for their relevance in describing mobility and bioaccumulation potential of the compounds in water and biota. Biodegradation was estimated via BIOWIN 3 (EPI SuiteTM), and the toxicity data was predicted via OCHEM. Physicochemical properties (empirical and estimated) and predicted environmental fate characteristics were retrieved via EPI SuiteTM. For some parameters ($\log K_{ow}$, $\log K_{oc}$, $\log S_w$ and $\log BCF$), empirical data were available. All parameters, showed high correlation between experimental and estimated data. This indicates that the empirical data are reliable and can be given preference over the estimated values.

2.3. Charged compounds

At given environmental conditions, compounds exist in positively charged, negatively charged, as zwitterions, or neutral form. According to Franco *et al.* (2010)²⁴, 41% of the REACH compounds are positively or negatively charged, 8% are zwitterions, and 51% occur in their natural form at typical environmental *pH* conditions. Charged ions may obviously behave different in the environment as compared to their uncharged counter compounds^{25, 26}. However, it has been shown that positively charged compounds behave relatively similar to their neutral

counter compounds²⁷. Therefore, in this work, positively charged ions were treated as neutral. If a compound is most likely negatively charged at typical environmental conditions, this compound should be assigned a $\log D$ value (estimated $\log K_{ow}$ at *pH* 7) rather than a $\log K_{ow}$ value. The pK_a values for all compounds were estimated in MarvinView. Since the typical environmental *pH* is around 7, we considered a compound with a $pK_a < 5$ negatively charged and $pK_a > 9$ uncharged. If a compound has a pK_a between 5 and 9, it was attributed an average value of $\log D$ and $\log K_{ow}$. (Figure 1).

3. Results and Discussion

3.1. Prioritization strategy

Scoring range. Before prioritization of the compounds, a scoring range for each descriptor was defined to create suspect screening lists for biota using LC-HRMS, biota using GC-HRMS and (drinking) water. The descriptors $\log K_{ow}/D$, $\log K_{oc}$, $\log S_w$, $\log BCF$, biodegradation, and quantity index were used for scoring of the final lists (Figure 1). For the biota suspect lists, the exposure indices air, water, and soil were included, whereas the water, sewage treatment and occupational exposure indices were scored for the (drinking) water suspect list. The toxicity data used here are relevant for human endpoints, and therefore only included in the prioritization for the (drinking) water suspect list. The suspect screening list for biota using GC-HRMS contained only chemicals with $\log K_{ow} > 3$ ²⁸ and got an additional score for $\log K_{oa}$.

Scoring. For the development of the suspect screening lists for water and biota, the compounds were prioritized by assigning a scoring within a range of 0 to 1 for all descriptors. A high score is associated with a high rank in the suspect lists. The mobility and bioaccumulation potential were two important factors for ranking of the compounds²⁹. For biota, a centered scoring system was applied for most descriptors ($\log K_{ow} / \log D$, $\log K_{oc}$ and $\log S_w$) in the biotic environment due to the fact that they describe both high mobility and bioaccumulation potential²⁹. In addition, the descriptors $\log BCF$, biodegradation, $\log K_{oa}$, exposure indices and emission quantity index are scored linear. For water, the mobility in the aquatic environment was the driving factor resulting in linear scoring of the descriptors $\log K_{ow} / \log D$, $\log K_{oc}$, $\log S_w$, $\log BCF$, biodegradation, toxicity, exposure indices and emission quantity.

Weighing. Before summing up the scores, a weighing factor was added to each descriptor to put more focus on certain descriptors which describe bioaccumulation (e.g. $\log BCF$, biodegradation) and mobility (e.g. $\log K_{ow} / \log D$, $\log S_w$) in the desired environment best. A weighing of the descriptors was necessary to get a sufficient spread in the end scores. This resulted in three suspect lists (two for biota and one for (drinking) water), each containing the top-1000 ranked compounds which can be used for suspect screening.

3.2. Sensitivity analysis

Some of the chosen physiochemical properties and predicted environmental fate characteristics have been shown to be significantly correlated. The parameters $\log K_{ow}$, $\log K_{oc}$ and $\log BCF$ correlate, and should therefore not be weighted equally in the prioritization step.

The toxicity data used in our approach were predicted using *in vitro* response toxicity data for several receptors (ER= estrogen receptor, AR= androgen receptor and TR= thyroid hormone receptor) because there is a lack of (available) species-specific measurement data for such a great number of compounds. For each predicted *in vitro* response, an estimation of the accuracy was made within the OCHEM prediction environment (high percentage = high accuracy). If the accuracy estimation was lower than 60%, the result was considered unreliable and discarded.

The descriptor biodegradation comprises semi-quantitative data, rather than half-lives. The software BIOWIN 3 generates data for compound classes rather than individual compounds. Certain compounds may get the same score for this parameter, even though they might differ in biodegradation potential. However, the BIOWIN 3 software is suitable for biodegradation in water and has been used successfully to search for drugs of potential environmental concern³⁰.

3.3. Suspect lists for biota and (drinking) water

The prioritization strategy as described in section 3.1 resulted in three lists of top-1000 highest scored

compounds for biota LC-HRMS, biota GC-HRMS and (drinking) water.

About 18% of all compounds in the used databases are negatively charged in the environment ($pK_a < 5$) with 8.9%, 4.2% and 13% in the generated suspects list biota LC-HRMS, biota GC-HRMS and (drinking) water, respectively (Table 1). Charged compounds are mainly present in the LC-HRMS and (drinking) water suspect lists, because ionizable compounds are highly mobile in aquatic solutions. Parameters estimated for these compounds should be well considered, because EPI SuiteTM estimated most descriptors based on the chemical's neutral form. In this work, we tackled this issue with replacing $\log K_{ow}$ with $\log D$ for negatively charged compounds.

Typically, halogenated compounds end up in the (drinking) water (26,4%) and Biota GC-HRMS (19,4%) suspect lists due to their high bioaccumulation factor.

4. Conclusion

The identification of emerging contaminants by non-target screening is time-consuming and an accepted common approach is lacking^{31, 32}. Suspect screening reduces the work tremendously by screening for certain group of compounds¹². Ultimately, we developed a tool to create suspect screening lists for various matrices (e.g. biota, water) for the identification of anthropogenic compounds providing new opportunities to screen for environmentally relevant compounds using HRMS.

Table 1. The percentage of halogenated and negatively charged compounds in the three suspect lists for biota using GC-HRMS ($n = 1000$), biota using LC-HRMS ($n = 1000$), and (drinking) water ($n = 1000$)

	Halogenated			Negatively charged			
	Biota LC-HRMS	Biota GC-HRMS	(Drinking) Water		Biota LC-HRMS	Biota GC-HRMS	(Drinking) Water
F	2,6%	6,3%	4,7%	$pK_a < 5$	8,9%	4,2%	13,3%
Br	1,3%	1,7%	2,6%	$5 < pK_a < 9$	3,1%	1,4%	6,9%
Cl	10,9%	11,1%	18,7%				
I	0,2%	0,3%	0,4%				
F + Br	0,2%	0,8%	0,1%				
Cl + F	1,7%	1,8%	0,6%				
Cl + Br	0,1%	0%	0,7%				
Cl + I	0%	0,1%	0%				
F + I	0,2%	0,9%	0%				
Br+ I	0%	0%	0%				
Total	15,0%	19,4%	26,4%		12,0%	5,6%	20,2%

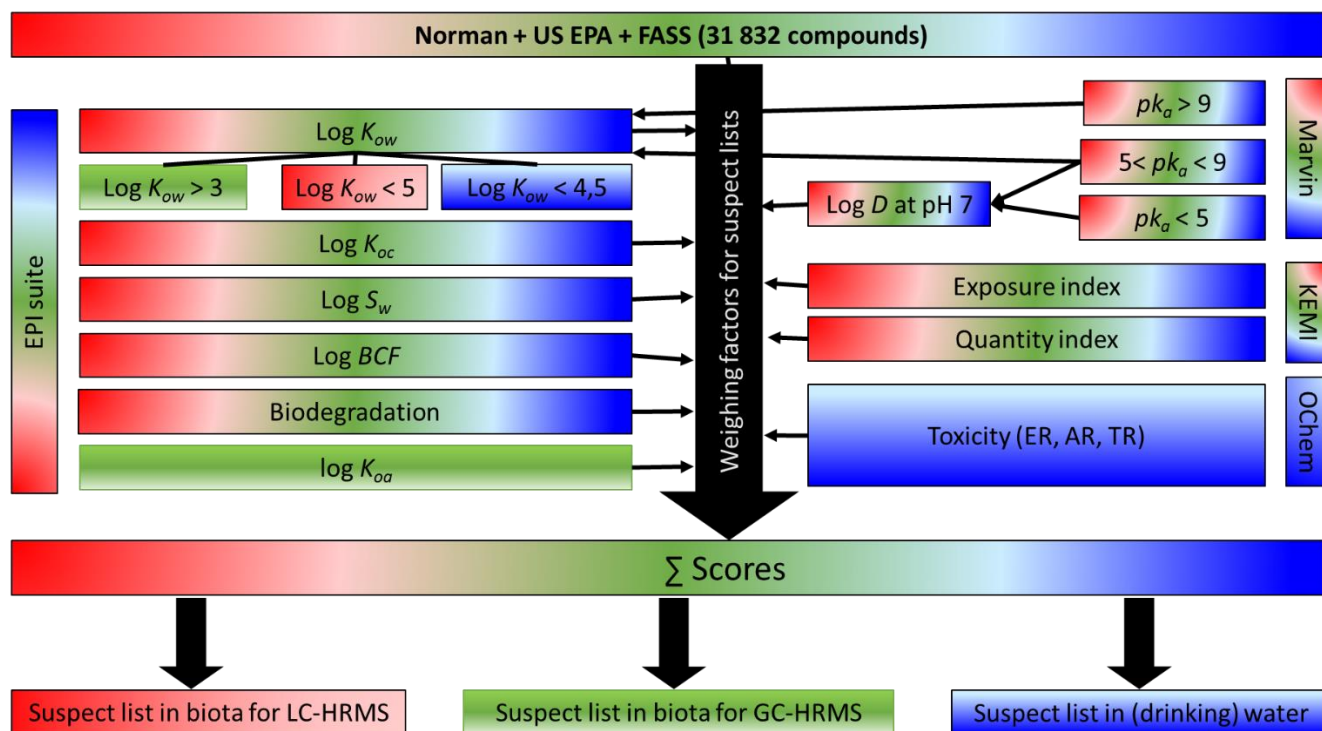


Figure 1. Overview of the prioritization strategy of organic compounds for the development of suspect screening lists of environmentally relevant compounds in (drinking) water and biota. Blue = (drinking) water, Red = biota LC- HRMS, green= biota GC-HRM. K_{ow} = octanol-water partitioning coefficient; D = adjusted K_{ow} at pH 7; K_{oc} = organic carbon-water partitioning coefficient; S_w = water solubility; BCF = bioconcentration factor; K_{oa} = octanol-air partitioning coefficient; ER= estrogen receptor; AR= androgen receptor; TR= thyroid hormone receptor. Exposure index and quantity index according to the SPIN database²³.

References

- CAS; 2016a; Devision of the american Chemical society home page; Available: <http://www.cas.org/content/regulated-chemicals> [accessed 23 November 2016].
- REACH; 2016; European Chemicals Agency home page; Available: <https://echa.europa.eu/information-on-chemicals/pre-registered-substances> [accessed 23 November 2016].
- CAS; 2016b; Devision of the american Chemical society home page; Available: <http://www.cas.org/content> [accessed 23 November 2016].
- Stockholm Convention; 2016; Stockholm Convention home page; Available: <http://chm.pops.int/TheConvention/ThePOPs/ListingofPOPs> [Accesses 14 February 2017].
- WFD; 2017; European commission home page; Available: http://ec.europa.eu/environment/water/water-dangersub/pri_substances.htm [Accessed 14 February 2017].
- Bletsou, A. A.; Jeon, J.; Hollender, J.; Archontaki, E.; Thomaidis, N. S., Targeted and non-targeted liquid chromatography-mass spectrometric workflows for identification of transformation products of emerging pollutants in the aquatic environment. *TrAC Trends in Analytical Chemistry* **2015**, *66*, 32-44.
- Ibanez, M.; Sancho, J. V.; Pozo, O. J.; Niessen, W.; Hernandez, F., Use of quadrupole time-of-flight mass spectrometry in the elucidation of unknown compounds present in environmental water. *Rapid communications in mass spectrometry : RCM* **2005**, *19*, (2), 169-78.
- 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.
- 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. *Analytical and bioanalytical chemistry* **2015**.
- Hug, C.; Ulrich, N.; Schulze, T.; Brack, W.; Krauss, M., Identification of novel micropollutants in wastewater by a combination of suspect and nontarget screening. *Environmental pollution* **2014**, *184*, 25-32.
- Plassmann, M. M.; Brack, W.; Krauss, M., Extending analysis of environmental pollutants in human urine towards screening for suspected compounds. *Journal of chromatography. A* **2015**, *1394*, 18-25.
- Krauss, M.; Singer, H.; Hollender, J., LC-high resolution MS in environmental analysis: from target screening to the identification of unknowns. *Analytical and bioanalytical chemistry* **2010**, *397*, (3), 943-51.

13. Brown, T. N. a. W., F., Screening Chemicals for the potential to be persistent organic pollutants: a case study of arctic contaminants. *Environ Sci Technol* **2008**, *42*, 5202-5209.
14. Chiaia-Hernandez, A. C.; Schymanski, E. L.; Kumar, P.; Singer, H. P.; Hollender, J., Suspect and nontarget screening approaches to identify organic contaminant records in lake sediments. *Analytical and bioanalytical chemistry* **2014**, *406*, (28), 7323-7335.
15. Fernandez-Sanjuan, M.; Meyer, J.; Damasio, J.; Faria, M.; Barata, C.; Lacorte, S., Screening of perfluorinated chemicals (PFCs) in various aquatic organisms. *Analytical and bioanalytical chemistry* **2010**, *398*, (3), 1447-56.
16. Masia, A.; Campo, J.; Vazquez-Roig, P.; Blasco, C.; Pico, Y., Screening of currently used pesticides in water, sediments and biota of the Guadalquivir River Basin (Spain). *Journal of hazardous materials* **2013**, *263 Pt 1*, 95-104.
17. Singer, H. P.; Wossner, A. E.; Mc Ardell, C. S.; Fenner, K., Rapid Screening for Exposure to "Non-Target" Pharmaceuticals from Wastewater Effluents by Combining HRMS-Based Suspect Screening and Exposure Modeling. *Environ Sci Technol* **2016**, *50*, (13), 6698-707.
18. Avagyan, R.; Westerholm, R., Target and suspect screening of OH-PAHs in air particulates using liquid chromatography-orbitrap high resolution mass spectrometry. *Talanta* **2017**, *165*, 702-708.
19. Judson, R. S.; Hu, X.; Slavov, S.; Rosenberg, S. A.; Farag, S.; Huang, R.; Todeschini, R.; Beger, R. D.; Politi, R.; Zang, Q.; Andersson, P. L.; Nicolotti, O.; Nikolov, N. G.; Cassotti, M.; Nicklaus, M.; Burton, J.; Shen, J.; Kancherla, J.; Balabin, I.; Tetko, I. V.; Ng, H. W.; Hong, H.; Incisivo, G. M.; Mangiatordi, G. F.; Grisoni, F.; Wedebye, E. B.; Muratov, E.; Benfenati, E.; Horvath, D.; Fourches, D.; Trisciuzzi, D.; Grulke, C. M.; Richard, A. M.; Worth, A.; Zakharov, A.; Varnek, A.; Tropsha, A.; Roncaglioni, A.; Rybacka, A.; Abdelaziz, A.; Mansouri, K., CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environmental Health Perspectives* **2016**.
20. EPI Suite™; Copyright 2000-2012; U.S. Environmental Protection Agency.
21. MarvinView 15.10.12.0; Copyright 1998-2015; ChemAxon Ltd., <http://www.chemaxon.com>.
22. On-line CHEmical database and Modeling environment; OCHEM; Available: <http://www.ochem.eu> [Accessed 23 November 2016].
23. SPIN database; Swedish Chemicals Agency; Available: <http://195.215.202.233/DotNetNuke/> [Accessed 23 November 2016].
24. Franco, A.; Ferranti, A.; Davidsen, C.; Trapp, S., An unexpected challenge: ionizable compounds in the REACH chemical space. *The International Journal of Life Cycle Assessment* **2010**, *15*, (4), 321-325.
25. Franco, A. a. T. S., Estimation of the soil-water partition coefficient normalized to organic carbon for ionizable organic chemicals. *Environmental toxicology and chemistry / SETAC* **2008**, *27*, (10), 1995-2004.
26. Fu, W. F., Antonio; Trapp, Stefan, Methods for Estimating The Bioconcentration Factor of Ionizable Organic Chemicals. *Environmental toxicology and chemistry / SETAC* **2009**, *28*, (7), 1372-1379.
27. Scheytt, T. M., P.; Lindstädt, R.; Heberer, T., 1-Octanol/Water Partition Coefficients of 5 Pharmaceuticals From Human Medical Care: Carbamazepine, Clofibrac acid, Diclofenac, Ibuprofen, And Propyphenazone. *Water, Air, and Soil Pollution* **2005**, *165*, 3-11.
28. Baduel, C.; Mueller, J. F.; Tsai, H.; Gomez Ramos, M. J., Development of sample extraction and clean-up strategies for target and non-target analysis of environmental contaminants in biological matrices. *Journal of Chromatography A* **2015**.
29. Kalberlah, F. O., J.; Schwarz, M. *Guidance for the precautionary protection of raw water destined for drinking water extraction from contaminants regulated under REACH*; 2014; p 204.
30. Andersson, P. L.; Fick, J.; Rännar, S., A Multivariate Chemical Similarity Approach to Search for Drugs of Potential Environmental Concern. *Journal of Chemical Information and Modeling* **2011**, *51*, (8), 1788-1794.
31. Anna, S.; Sofia, B.; Christina, R.; Magnus, B., The dilemma in prioritizing chemicals for environmental analysis: known versus unknown hazards. *Environmental science. Processes & impacts* **2016**.
32. Bader, T.; Schulz, W.; Kummerer, K.; Winzenbacher, R., General strategies to increase the repeatability in non-target screening by liquid chromatography-high resolution mass spectrometry. *Analytica chimica acta* **2016**, *935*, 173-86.