Journal Name

ARTICLE

Ranking REACH registered neutral, ionizable and ionic organic chemicals based on their aquatic persistency and mobility

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

H. P. H. Arp,*[a] T. N. Brown,[b] U. Berger [c] and S. E. Hale [a]

The contaminants that have the greatest chances of appearing in drinking water are those that are mobile enough in the aquatic environment to enter drinking water sources and persistent enough to survive treatment processes. Herein a screening procedure to rank neutral, ionizable and ionic organic compounds for being persistent and mobile organic compounds (PMOCs) was developed and applied to the list of industrial substances registered under the EU REACH legislation as of December 2014. This comprised 5155 identifiable, unique organic structures. The minimum cut-off criteria considered for PMOC classification herein are a freshwaterhalf-life > 40 days, which is consistent with the REACH definition of freshwater persistency, and a log D_{OC} < 4.5 between pH 4-10 (where D_{OC} is the organic carbon-water distribution coefficient). Experimental data were given the highest priority, followed by data from an array of available quantitative structure-activity relationships (QSARs), and as a third resort, an original Iterative Fragment Selection (IFS) QSAR. In total, 52% of the unique REACH structures made the minimum criteria to be considered a PMOC, and 21% achieved the highest PMOC ranking (half-life > 40 days, log D_{OC} < 1.0 between pH 4-10). Only 9% of neutral substances received the highest PMOC ranking, compared to 30% of ionizable compounds and 44% of ionic compounds. Predicted hydrolysis products for all REACH parents (contributing 5043 additional structures) were found to have higher PMOC rankings than their parents, due to increased mobility but not persistence. The fewest experimental data available were for ionic compounds; therefore, their ranking is more uncertain than neutral and ionizable compounds. The most sensitive parameter for the PMOC ranking was freshwater persistency, which was also the parameter that QSARs performed the most poorly at predicting. Several prioritized drinking water contaminants in the EU and USA, and other contaminants of concern, were identified as PMOCs. This identification and ranking procedure for PMOCs can be part of a strategy to better identify contaminants that pose a threat to drinking water sources.

Environmental Impact

A procedure to identify and rank organic substances for their ability to be persistent and mobile in the aquatic environment was developed and applied to REACH registered substances and their hydrolysis products. This is the first general screening approach to identify organic substances that may appear in drinking water based on substance properties and molecular structure. This procedure could be used for other chemical inventories, or proposed substances, as part of efforts to identify emerging or unknown drinking water contaminants. Several of the REACH registered substances identified as persistent and mobile in this study are known drinking water contaminants, though there are also many others that have not yet been investigated in drinking water resources.

1. Introduction

Ensuring that drinking water resources are secure from unwanted and toxic chemicals is a central goal of human health protection and human rights.[1-3] An under-investigated threat to drinking water resources is the plethora of new chemicals that are appearing on the market, as the chemical industry continues to innovate new, useful products and technologies. Some of these new and existing substances may possess certain intrinsic, physico-chemical properties that make them readily able to contaminate drinking water sources, , if they are used in a way that leads to substantial environmental emissions.[4] If the same substances are toxic, this could lead to serious health consequences. The intrinsic properties that enable a chemical to potentially contaminate drinking water resources are its aquatic persistency (P) and mobility (M). Organic compounds (OC) that have substantial P and M characteristics, so called PMOCs, can transport through river banks, groundwater aquifers, and other natural and urban barriers to reach sources

[a] Norwegian Geotechnical Institute, Postboks 3930 Ullevål Stadion, NO-0806 Oslo, Norway. Email: hpa@ngi.no, Tel: + 47 950 20 667
[b] ARC Arnot Research and Consulting Inc., 5536 Sackville St., Halifax, Nova Scotia, Canada.
[c] Department of Analytical Chemistry, Helmholtz Centre for Environmental Research – UFZ, Permoserstr. 15, DE-04318 Leipzig, Germany.

Electronic Supplementary Information (ESI) available in two parts. Part S1 is a document containing extra-information on the methods; Part S2 a spreadsheet containing relevant information on all substances screened in this study. See DOI: 10.1039/x0xx00000x

Please do not adjust margins

of drinking water. When PMOCs first appear in drinking water, it is difficult for them to be removed. These compounds can recirculate within the drinking water cycle, particularly in urban and drought-prone areas where waste water is recycled to drinking water. Drinking water treatment processes can only be a partial help, as compounds with substantial P and M properties may also survive treatment technologies like ozonation, chlorination, filtration by activated carbon, or even reverse osmosis.[5-7] Therefore, any contamination of drinking water with PMOCs can be long-lasting.

A central focus of the European Union's (EU) drinking water directive (Council directive 98/83/EC) is to prevent drinking water contamination that may adversely affect human health. The current focus of the EU's chemical regulation and in particular the REACH legislation (Regulation EC No 1907/2006), on the other hand, has not been to prevent drinking water contamination, but rather to have better control of substances exhibiting environmental persistency (P), bioaccumulation (B), and toxicity (T), so-called PBT substances.[8] This is largely because of the growing concern over the past five decades [9, 10] that PBT substances like DDT and PCB can have on human health and the environment. A PMOC that meets the REACH criteria for toxicity can be considered a PMT-type substance.[11] PBT and PMT substances bear some similarity. Both can accumulate in the environment, such that the risk of exposure to humans and ecosystems can increase with emissions. The key difference is the route of exposure. PBT substances accumulate predominantly through the food chain, in contrast to PMOC/PMT compounds, which recirculate and may accumulate through water cycles, including drinking water cycles. Further, as bioaccumulation and mobility are not inherently exclusive, a subset of PBT substances would be also PMT substances. Screening approaches to predict or identify PBT compounds from lists of existing substances have been applied,[8, 12, 13] and human exposure models that include drinking water as an exposure pathway have been developed,[14-16] but to our knowledge no similar screening tools have been implemented specifically to identify PMOC/PMT substances.

Mobility in the aquatic environment is associated with substances having a very high water solubility (S_{water}, µg/L) or substances having very low capacity for sorption to soils and other natural media. Sorption in this manner is typically quantified with an organic carbon-water partition coefficient (K_{oc}), defined as the ratio of a substance sorbed to natural soil or sediment organic carbon (µg/kg) vs that in surrounding water (µg/L) at equilibrium; for ionizable substances it is quantified with the pH-dependent organic carbon-water distribution coefficient (D_{oc}), which accounts for the total sum of neutral and charged species sorbed and dissolved. In general, the lower the K_{oc}/D_{oc} value, the more readily a substance can reach the aquatic environment, without sorbing substantially to surfaces.

One essential difficulty in conducting risk assessments for highly mobile substances is that, particularly for the most mobile substances, we often lack analytical approaches to measure them.[7] Standard gas-chromatographic and liquid-chromatographic techniques are poor at analysing substances with a K_{oc}/ D_{oc} value < 1 (i.e. substances that have a higher concentration in water than soil organic carbon at equilibrium). This lack of analytical methods has recently been referred to as the "analytical gap".[7] Techniques to measure these substances are few, though new methods are emerging. Therefore, many of these chemicals may already be in drinking water, going unnoticed.[7]

In this current study, we present a screening procedure that can be used to identify and rank existing and future PMOCs for their potential ability to permeate drinking water sources. The screening approach was designed specifically to be compatible with existing definitions and chemical properties used in the EU REACH legislation, to facilitate identification of PMOCs using substances properties included during the REACH registration process. This included the definition of persistent and very persistent in fresh- or estuarine water as presented in Annex XIII of REACH, and the qualitative description of mobility in Annex II of REACH (i.e. "MOBILITY: The potential of the substance or the appropriate constituents of a preparation, if released to the environment, to transport to groundwater or far from the site of release"). In this study we did not explicitly consider the definition toxicity in REACH, to identify PMT substances, in part because a study of PMOCs in the environment is of relevance in its own right; however, a sub-goal of this study is to compare identified PMOCs with previously identified PBT substances.

This screening approach can be used to help industry, environmental chemists and water regulators identify what chemicals have a potential to be rapidly distributed in the aquatic environment (i.e. pose a potential hazard). Most PBT screening studies heretofore have focussed primarily on neutral compounds.[8] In this study, it was essential to include ionizable and ionic species as well, due to their propensity to be mobile, despite the low accuracy of prediction tools currently available for ionic compounds. In addition, we also performed this assessment on predicted hydrolysis products of the REACH registered substances, so that not only the parent compounds are considered but environmental transformation products as well.

2. Methods

2.1. REACH List

The publically available list of REACH registered substances (https://echa.europa.eu/information-on-chemicals/registered-substances) was accessed on 19 December 2014, which at that time contained 14076 substance entries. Of these, 7313 had a *unique* Chemical Abstracts Services (CAS) number, 1172 had replicate CAS numbers, 5455 had a European Community (EC) number but no CAS, and 136 entries had neither a CAS nor EC number. Only the 7313 compounds with unique CAS numbers were considered (corresponding to 8485 individual substance

 This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

entries), as these were the easiest to link to available chemical property databases.

2.2. SMILES codes

For the 7313 unique CAS entries, SMILES codes (SMILES = simplified molecular input line-entry system)[17, 18] were obtained from various databases that linked CAS to SMILES. These included Chemaxon (https://www.chemaxon.com/), QSARToolbox v3.3 (http://echa.europa.eu/support/oecd-qsar-toolbox), PubChem (https://pubchem.ncbi.nlm.nih.gov/) and ChemSpider (www.chemspider.com/); all websites and databases were accessed January - March 2015. If none of the above databases contained a structure, SMILES were obtained manually from the structures presented in the REACH dossier. The SMILES from multiple sources listed above were compared, when available. Discrepancies were flagged, and the best SMILES was manually chosen or reformulated to have a net charge of zero and be in "dative bond" notation (e.g. a nitro group is often represented as [O-][N+]=O, but N(=O)=O in a neutral dative bond structure). Some of the aforementioned databases may provide SMILES without a charge of zero, such as by not adding the counter-ions (in the case of salts), or presenting the acidic / basic form of a neutral species despite the CAS being for the neutral species. Alternatively, some SMILES sources ignored charges when there should be one. To ensure a net charge of zero and the correct notation, counter-ions were manually added when they were missing. As an example, for chemicals like magnesium acetate, some databases would remove the counter-ion "CC(=O)[O-]", some like QSARToolbox would remove the charge to make it look neutral " CC(=O)O", but the correct structure used here (and most typically used by PubChem) was CC(=O)[O-].CC(=O)[O-].[Mg+2] (the "." in the SMILES means that the structures are not connected by covalent bonds).

2.3. Organic Compound Definition

Herein organic compounds are defined as those containing a C-H, C-C, Si-C bond, or 2 carbons along with any combination of the elements H, C, O, N, P, S, F, Cl, Br and I. Organic compounds that contained one or more B atom were classified as organoboranes, those that contained one or more Si atoms were classified as organosilanes and those containing another element than listed above were classified as organometallics. Compounds containing a single C in combination with one or more of the elements H, O, N, Si, P, S, F, Cl, Br and I were categorized as pseudo-organics, and were included in the screening as well. Other types of REACH registered substances either lacked a clear chemical structure (e.g. reaction products, natural products, complex mixtures) or contained other combinations of atoms than those listed above (i.e. completely inorganic structures), and were therefore not considered. Following this classification, there were 5530 unique organic and pseudo-organic substances with CAS numbers remaining.

A closer examination of these 5530 organic substances, however, revealed there were only 5155 unique REACH registered organic compound (REACH OC) structures, after accounting for reoccurring structures across different CAS numbers and CAS entries containing multiple organic structures. Reoccurring structures across different CAS numbers included common pseudo-organic counter-ions (e.g. carbonate occurred in 34 CAS entries). CAS entries could contain more than one organic structure due to mixtures with organic cations or anions (e.g. 126-97-6 (2-hydroxyethyl)ammonium mercaptoacetate), or at times blends of neutral molecules (e.g. 1319-77-3 refers to a mixture of three neutral isomers of cresol).

2.4. Classification by Charge and Ionizability

Whether structures were neutral, ionizable or ionic was classified as follows. First, a simple reading of the SMILES code in dative bond notation was used to categorise the ionic charge of the substance. Substances were initially classified as a "single anion" or "single cation" if one "-" or "+" was present in the SMILES, respectively. Otherwise, if multiple "-" or "+" were present, the substance was classified as a "multiple anion" or "multiple cation", respectively. If both "-" and "+" were provided in the same structure, the compound was classified as a "zwitterion". If no charges were present the compound was initially considered "neutral".

Next, the pH dependence of each structure, between a pH range from 4 to 10, was considered by estimating acidic and basic pK_a values (of A-H and BH+ moieties in the molecule, respectively) using the following commercial QSAR packages: JChem for Office along the Protonation Calculator Plugin from Chemaxon (www.chemaxon.com), Insights for Excel 2.3 by accelrys® (www.accelrys.com), the ADMET Predictor 7.1 software by Simulations-plus (www.simulations-plus.com/) and the SPARC v6.0 standalone calculator from Archem (www.archemcalc.com/sparc/). All versions were purchased in January 2015, and used by April 2015. Compounds initially classified as "neutral" were re-classified as "ionizable" and acidic if only "acidic" pK_a values were determined (by all packages) and the lowest pK_a was < 12; or they were re-classified as "ionizable" and "basic" if only basic pK_a values were determined (by all packages) and the highest pK_a was > 2. For clarity, throughout this paper the pK_a for organic bases refers to the conjugated acid structure (i.e. pK_{BH+} values for BH+ moieties, such as in the protonated amine R-NH_3+). Compounds initially classified as "single cation" were re-classified as ionizable and basic if they had a basic pK_a > 2. Similarly, compounds classified as "single anion" were re-classified as ionizable and acidic if they had pK_a < 12. Otherwise, the classification of "neutral", “single anion" or "single cation" was retained. Compounds were considered amphoteric if both an acidic pK_a < 12 and basic pK_a > 2 was predicted by any one or a combination of the above software packages. Substances classified as "multiple cation", "multiple anion" and "zwitterion" were not reclassified, though it was noted if they were predicted to behave as acidic, basic or amphoteric compounds within the pH range from 4 to 10. Note that we did not use

Please do not adjust margins

Please do not adjust margins

experimental pK_a values for this classification, as the databases available during the time of this study (see below) did not consistently identify if the substances were acidic, basic or amphiprotic. Substances where no pK_a was available were also not re-classified.

2.5 Persistency Criteria

In Annex XIII of REACH, a substance is considered to be persistent in fresh or estuarine water if its degradation half-life is > 40 days and very persistent if it is > 60 days. Note that this half-life should ideally refer to 12 °C based on the new PBT guidance in REACH.[19] If half-lives at this temperature were not available, data and models for 20 – 25 °C were used and not corrected further. As persistency estimates can be uncertain, we also considered half-lives of > 20 days to be "potentially persistent". Results from biodegradation screening tests from Organization for Economic Co-operation and Development technical guidance (OECD TG) 301 A-F, OECD 302 B-C and OECD 310, were also taken into consideration, where a result of "readily biodegradable" was considered not-persistent (results of "inherently biodegradable" were not considered to err on the side of caution). Four persistency scores (P-scores) were chosen: P1 (freshwater half-life < 20 days or at least one OECD TG result of "readily biodegradable"), P2 (20 d < freshwater half-life < 40d), P3 (40 d < freshwater half-life < 60 d) and P4 (60 d < freshwater half-life).

There are many different pathways that can influence persistency in surface freshwater. Of these, only four were considered due to data availability: aerobic biotransformation in water, hydrolysis, phototransformation, and volatilization from surface water under still conditions. For the first three of these processes, experimental or estimated half-lives were collected directly, according to the data prioritization section presented below. Volatilization half-lives, on the other hand, were estimated from the following equation, which applies to completely still conditions:

$$t_{1/2,volatilization} = 0.69/(v_{aw} * h) \quad (1)$$

Where h is the depth of the water (here assumed as 1 m) and v_{aw} is the air-water exchange velocity.[20] As explained in the Electronic Supporting Information (ESI)-Section S1, v_{aw} can be estimated by the Henry's Law constant, K_{aw}, and compound specific diffusivities in air and water. If K_{aw} is not available it can be estimated using $K_{aw}=v.p./(S_{water,L}RT)$, where v.p. is the sub-cooled liquid vapour pressure (Pa), $S_{water,L}$ the subcooled liquid water solubility (mmol/L), R the ideal gas constant and T the temperature.

Because phototransformation and volatilization are only relevant for surface waters, a separate P score was assigned for surface water (PS-score) and ground water (PG-score). For the PG-score, the shortest half-life from aerobic biotransformation and hydrolysis was considered, exclusively. If a PG-score could not be provided (i.e. when predictive Quantitative Structure-Activity Relationships (QSARs) provided only blank output for both biotransformation and hydrolysis), a PS-score was not calculated, as it was considered presumptuous to derive a P-score without this information. If a PG-score was present, the PS-score was based on the shortest half-life or lowest P-score from all four processes. Therefore, surface water half-lives will be equal to or less than ground water half-lives, and therefore PS-scores are equal to or less than PG-scores. Other potential transformation processes for non-aquatic environments (e.g. transformation in soil or sediment) were not considered as part of this assessment. The experimental databases and QSARs used to assess persistency, and how data were prioritized amongst these sources, are described in Section 2.8.

2.6 Mobility Criteria

There is no formal mobility criterion in REACH. As a suggestion, a recent guidance document from the German Federal Environment Agency[11] favoured use of K_{OC} as the best parameter to describe mobility, as it was found to be the most sensitive parameter to describe breakthrough of a neutral substance through a wastewater treatment plant (WWTP); such a breakthrough event can be considered an indicator of mobility in drinking water cycles. This report suggested a threshold log K_{oc} of 4.5 as the mobility criterion, or alternatively an S_{water} of 0.15 mg/L. Though it can be argued that these are very conservative thresholds for mobility (0.15 mg/L is a solubility that is hard to measure), it was recommended as the baseline threshold to account for extremely persistent or non-degradable substances eventually reaching drinking water sources over long time frames. Favouring the use of these criterion, both S_{water} and K_{oc} are required during REACH registration. S_{water} has to be reported based on Annex VII for most substances manufactured or imported in quantities greater than 1 to 10 tonnes/y. Information on log K_{oc} can be obtained by information mandated in Annex VIII and IX, for certain types of substances manufactured or imported in quantities greater than 10 to 100 tonnes/y.

This cutoffof log K_{oc} < 4.5 only applies to compounds whose ionization state does not change with pH. For ionizable compounds and ionic compounds, the pH dependent D_{oc} needs to be considered, which is dependent on the substance pK_a, i.e..

$$D_{oc} = (1/(1+10^{(pH - pK_a)}))K_{oc} \quad \text{(monoprotic acids)} \quad (2)$$
$$D_{oc} = (1 - 1/(1+10^{(pH - pK_a)}))K_{oc} \quad \text{(monoprotic bases)} \quad (3)$$

Similarly, the pH dependency of $S_{water,L}$ (mmol/L) can also be related to pH:

$$\log S_{water,L} = \log S_{water,L}(neutral) + \log(1 + 10^{pH-pKa}) \quad \text{(monoprotic acids)} \quad (4)$$
$$\log S_{water,L} = \log S_{water,L}(neutral) + \log(1 + 10^{pKa-pH}) \quad \text{(monoprotic bases)} \quad (5)$$

The mobility cut-off values ionizable compounds were applied between the pH range of 4 to 10, meaning that either the lowest

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

log D_{OC} value or maximum S_{water} (=$S_{water,L}$ x M.W (g/mol)) within this pH range was used as a conservative assumption. It follows that i) monoprotic compounds with a pK_a > 12 (or pK_{BH+} < 2) would not have pH dependent D_{oc} or S_{water} values in this range; ii) for monoprotic acids the D_{oc} or S_{water} values at pH 10 were used, and iii) for monoprotic bases the D_{oc} or S_{water} at pH 4 were used. For complex amphiprotic and polyprotic compounds, the QSAR predicted minimum D_{OC} or maximum S_{water} at pH 4, 5, 6, 7, 8, 9 and 10 was used (as described in more detail section 2.8.4), unless pH dependant experimental data was available.

For charged ionic and ionizable molecules there are other interactions that can decrease mobility in the environment that are not accounted for with D_{OC} or S_{water} alone, such as cationic or anionic exchange interactions to minerals and other surfaces[21-23] or precipitation with counter-ions. Thus, in the absence of accounting for such additional interactions, basing mobility on pH dependent D_{oc} or S_{water} represents a maximum assumption for mobility. Further, for neutral compounds that are solid at room temperature, data sources do not always specify if S_{water} is referring to the solubility of solid crystals $S_{water,S}$, or the liquid state $S_{water,L}$. $S_{water,L}$ is more relevant for environmental fate models, and can be derived from $S_{water,S}$ by accounting for the enthalpy of fusion[20] or the fugacity ratio.[24] As $S_{water,L}$ is larger than $S_{water,S}$, it can be considered a maximum and more environmentally-relevant solubility; hence it was used preferentially here, when the choice was provided.

If no experimental or estimated log K_{oc} or S_{water} data were available, the Karickhoff single-parameter linear-free energy relationships was used to give an estimation of K_{oc}, based on the octanol-water partition coefficient, K_{ow}:[25]

$$\log K_{oc} = \log K_{ow} - 0.21 \qquad (6)$$

It is noted that estimated S_{water} values were here always given priority over experimental and estimated K_{ow} values extrapolated to K_{oc}, as equation 6 is known to be weakly performing for highly polar (mobile) molecules,[20, 26-28] and is irrelevant for ionic substances.[21-23] On the other hand, K_{ow} would be more appropriate than S_{water} for large neutral molecules, particularly because these are generally in the solid state at room temperature, and S_{water} data may be based on $S_{water,S}$ whereas K_{ow} by definition refers to the subcooled liquid state.[20] Further theoretical corrections from the solid to the subcooled liquid state become more inaccurate the higher the melting point.[24] Thus, to account for these instances when K_{ow} values extrapolated to K_{oc} are more appropriate than S_{water}, a secondary mobility assessment was derived based on this consideration as part of the sensitivity analysis (described below).

The following classification for the mobility score (M-score) was used, spanning from M1 (the lowest) to M5 (the highest). It is based on the lowest log K_{OC} (or D_{OC} value if ionizable), or maximum S_{water}, between a pH of 4-10 and temperature between 12 and 25°C: M1 (log K_{oc} ≥ 4.5 or S_{water} ≤ 150 µg/L), M2 (3 ≤ log K_{oc} < 4.5 or 50 mg/L ≥ S_{water} >150 µg/L), M3 (2 ≤ log K_{oc} < 3 or 1 g/L ≥ S_{water} > 50 mg/L), M4 (1 ≤ log K_{oc} < 2 or 10 g/L ≥ S_{water} > 1 g/L) and M5 (log K_{oc} < 1 or S_{water} > 10 g/L). In the case of equivalent data quality (see below), log K_{OC} was favoured over S_{water}, and data closer to 12°C was favoured over data closer 25 °C (to reflect REACH guidance of being representative of average environmental conditions).[19] The experimental databases and QSARs used to obtain mobility data, and how data were prioritized amongst these sources, are described in Section 2.8.

2.7. PMOC Scoring

An integrated PMOC scoring system was implemented to combine the P-score and M-score, as shown in Figure 1. The basis of the PMOC score is that compounds with the lowest mobility and persistency categories (M1 and P1) have the lowest probability of being a PMOC, while scores of M5 and P5 are for those compounds that have the highest chance of being a PMOC and permeating a drinking water resource if emitted into the environment. Compounds which have either P1, M1, or a combination of P2 and M2 are considered non-PMOC structures, belonging to one of three categories: immobile persistent organic compounds ("immobile POC") that have a P-score of 3 or 4 and an M-score of 1, unstable mobile organic compounds ("unstable MOC") that have an P-score of 1 and an M-score of 5, and "transient" substances for all other combinations. Compounds that are considered PMOCs are ranked with a score from 1 to 5 and assigned a colour code for clarity.

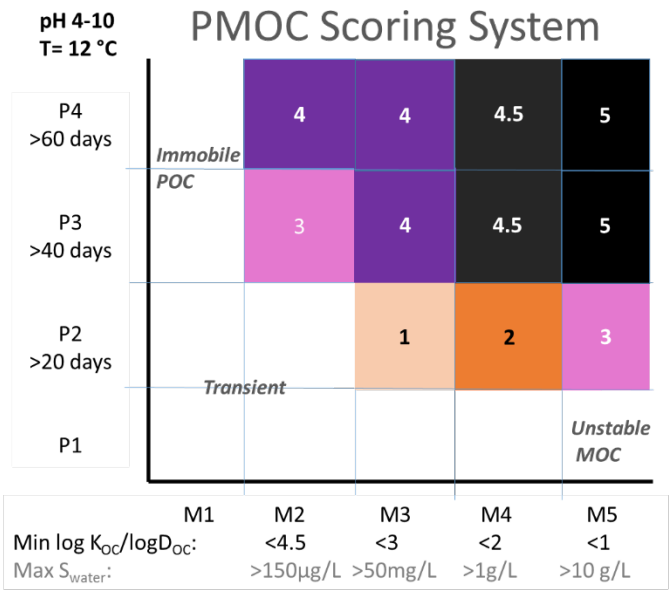


Figure 1. Depiction of the PMOC scoring system considering persistency (P) based on half-lives in freshwater and mobility (M) based on either the minimum log D_{OC}, minimum log K_{OC}, or maximum S_{water} value (where log D_{OC} or log K_{OC}, is prioritized over S_{water}) over a pH range from 4 to 10. Data for 12°C or the closest temperature possible is used.

2.8. Data Prioritization.

Please do not adjust margins

An underlying concern of any classification system based on chemical properties is the availability and the quality of the required chemical property data. Further, as chemical property data of varying quality can originate from many sources, a protocol needs to be in place that prioritizes how to choose amongst these sources. The data source prioritization system used here is presented in Table 1. Only data available from the highest priority were used, without further consideration of data from lower priority sources. If multiple data occurred at the same priority level, these data were typically averaged. A description of each of these data sources is described in the following subsections.

Table 1. Data source prioritization for P and M scoring

Priority	Source
1st	REACH dossier *experimental* data
2nd	Peer-reviewed *experimental* databases and PP-LFERs (using experimental input data)
3rd	*EPI Suite* experimental database
4th	*Estimated* from available QSARs: P – EPISuite (Biowin, Hydrowin), QSARToolbox M - SPARC, EPISuite, Chemaxon, Insight for Excel, ADMET
5th	IFS PMOC QSAR

2.8.1. REACH dossier experimental data.

Experimental data in REACH dossiers that were reported as being of high quality was prioritized above other data, to address the subgoal of this PMOC screening study to make it as consistent with the REACH registration process as possible. It is noted, however, that reporting in REACH does not consistently require peer-review, but it does reflect how REACH registrants/industry themselves have characterized the substances they registered; therefore, prioritizing these data was chosen more out of practical and applied reasons than that of scientific rigour. It should be noted that a recent study has found that REACH dossiers are often lacking in experimental data, and rely on estimation methods.[29] To access the dossiers in a practical manner, the eChemPortal database available from ECHA and OECD (ww.echemportal.org, last accessed for this study in March 2015) was used. The eChemPortal allows for users to enter search criteria for a given chemical property from a variety of databases, including REACH dossiers, and provides an output as CSV or Microsoft Excel tables. Chemical property data from the eChemPortal utilized here include aerobic biodegradation test results (301 A-F, OECD 302 B-C and OECD 310, in addition to half-life data), hydrolysis half-lives, phototransformation rates in water, K_{aw}, S_{water}, vapour pressure, pK_a, K_{ow}, and K_{oc}, with the latter typically measured using adsorption studies (EC C18; OECD 106, 2000a) or HPLC studies (EC C19; OECD 121, 2001a).[11]

When accessing eChemPortal, the search filters were set to experimental data with a reliability score of 1 (reliable without restrictions) or 2 (reliable with restrictions), and accepted without further scrutiny. Half-life data given with the operators ">" and "<" were only used if there was no ambiguity in relation to the P-Score of 3, meaning that only half-lives given as < 40 days or > 40 days could be used, but not e.g. < 50 days or > 30 days. For mobility, all operators were interpreted as "=", due to the comparative rarity in which data was presented as ">" and "<" and because approximate data would be less likely to influence the M-score, which are based on differences of a factor 10. Cases of high standard deviations were flagged for manual follow-up, and data suspected as being erroneous were either deleted or corrected on a case-by-case basis (e.g. by log normalizing). An identified shortcoming of using the eChemPortal database to export REACH dossier data, particularly for pK_a, was that data were not consistently log normalized and empty data cells in the exported CSV files generally meant the experimental data were in the "comments" section of online dossiers (these data were manually transferred when spotted). There were also instances where data in the online dossiers were not present in eChemPortal at the time of data extraction.

2.8.2. Peer-reviewed experimental data and PP-LFERs.

Peer-reviewed experimental databases and compilations were taken as the next level of priority. This literature search focused on databases, rather than on reports for individual structures (due to time limitations). Parameters for which peer-reviewed databases could be obtained include pK_a,[30, 31] vapour pressure,[31, 32] K_{aw},[33-35] and K_{ow}.[36-40]

Additionally, at this level of priority, poly-parameter linear-free energy relationships (PP-LFERs) were used for K_{oc}[28] and K_{aw}[41], as defined in the following equations:

$$\log K_{oc} = 0.02 + 1.20V - 0.98S - 0.42A – 3.34B + 0.54L \quad (7)$$
$$\log K_{aw} = -1.27 + 0.82E + 2.74S + 3.90A + 4.81B - 0.21L \quad (8)$$

Where V the McGowan molecular volume, S is the polarizability/dipolarizability descriptor, *A* is the H-bond basicity descriptor, *B* the H-Bond acidity descriptor, L is the hexadecane-water partition coefficient, and E is the excess molar refraction. Note that *L*, *V* and *E* are proxies for non-specific interactions (London dispersion, cavity formation), and *S*, *A* and *B* for specific/polar interactions. It is important to emphasize that these PP-LFER descriptors should all be determined experimentally, as estimation methods are considered dubious, particularly for very-polar compounds,[27] with the exception of the *L* parameter.[42] PP-LFER descriptors were compiled from the Helmholtz Centre for Environmental Research - Linear Solvation Energy Relationship (UFZ LSER) database during March 2015.[43]

No peer-reviewed data-bases for biodegradation, hydrolysis or phototransformation could be found for this work.

2.8.3. EPI Suite experimental database

The data source considered as the third priority was the experimental database published by the U.S. EPA's EPISuite[44, 45]

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

(Estimations Programs Interface), which contained experimental K_{aw}, S_{water} and K_{ow} data.

2.8.4. QSAR property data

Finally, if no experimental data were available, it was necessary to use QSARs. The eChemPortal database and REACH dossiers provide QSAR output of properties related to persistency and mobility. However, in this study, we ignored these data and conducted original QSAR analysis. The reasons for doing this were that a) QSARs generally only require SMILES structure as input, and can be done in batch mode for a large set of chemicals; b) the QSAR data presented in the REACH dossiers are from highly-variable sources, so accounting for accuracy and consistency across substances is difficult; and c) for half-life data there were very few QSAR predictions available through eChemPortal (e.g. for aerobic biodegradation only half-lifes for 21 compounds were predicted using QSARs with high reliability scores).

Regarding persistency, Q*SARToolbox* (v 3.3, available from http://www.qsartoolbox.org/) was used to run the EPISuite's BIOWIN (output from BIOWIN 1 through 6), EPISuite's HYDROWIN, and the LMC hydrolysis model. Biodegradation half-lives were estimated from BIOWIN output using the method presented in Arnot et al. (2005).[46] This method presents several alternative models to derive half-lives, here the geometric average of these models plus one geometric standard deviation was used, to err on the side of being conservative. Estimated hydrolysis P-scores were derived for both Hydrowin and LMC estimates. The Hydrowin P-score was based on a combination of Hydrowin half-life categories (e.g. 0 to 1 day, 1 to 10 days, >100 days) and half-lives under basic and acidic conditions from pH 6.5 to 7.4, without further scaling to account for a pH range of 4 to 10, due to the perceived uncertainty of the method. The LMC hydrolysis model output of categories "very slow", "slow" and "moderate" were given a score of P4, P3 and P2, respectively. If both Hydrowin and LMC gave two different P-scores, the lowest of the two was used. For volatilization rates, vapour pressure and K_{aw} data were obtained by EPISuite (MPBPWIN, HENRYWIN Bond Method and Group Method) (at STP) as well as SPARC (at 12 °C). No suitable QSAR for phototransformation rates could be identified at the time of the study.

For mobility, ChemAxon, Insights for Excel, ADMET and SPARC were used to predicted pK_a (as mentioned above), as well as pH dependent S_{water} and K_{ow}/D_{ow} values. SPARC was the only one of these for which S_{water} and D_{ow} could be predicted at 12 and 25 °C. In addition, EPISuite[44, 45] (via QSARToolbox) was also used to predict the S_{water} and K_{ow} for neutral compounds (EPIsuite was not used for ionic compounds, as it appeared to automatically convert charged atoms to neutral, simply by deleting the charge, resulting in unreasonable predictions).

2.8.5. IFS QSAR estimations

It was anticipated at the beginning of this study that there would be some substances for which no experimental data exist and for which QSARs would not be able to predict the needed parameters for the P-score and M-score. Therefore, in order to include all REACH OC structures, original group contribution QSARs were designed to estimate approximate rankings for persistence and mobility.

This was done using experimentally based M-scores (n = 1320) following the Iterative Fragment Selection (IFS) method,[47] which automatically generates and selects fragments (functional groups) that are used in a multiple linear regression (MLR) model. Calibration (n=663) and validation (n=657) datasets were automatically selected and the prediction accuracy was quantified. For the PG- and PS-scores the IFS results were poor, and a custom method was designed. In brief, for compounds with experimentally based PG- and PS-scores (n = 834 and n = 824, respectively), fragments corresponding to all atoms and all bonded atom pairs were defined. To this pool of fragments was added more complex functional groups known to be important for persistency. Then the fragments were all added to an MLR model, and finally the fragments with the most uncertainty in their regression coefficients were iteratively removed until all remaining fragments had acceptable uncertainty. In both cases a subset of molecules were used for the calibration of the QSARs (PG: n=396; PS: n=390), while the remainder were used for validation (PG: n=438; PS: n=434). The resulting group contribution QSARs were then compared with the validation set, and the resulting accuracy in prediction was quantified. These three QSARs for M-, PG- and PS-scores are hereafter referred to as the IFS QSAR.

For the M score, the validation check of the IFS QSAR gave a moderate Pearson correlation coefficient (r^2) of 0.4; but there was an apparent separation of the M1 and M4-5 predicted values. Therefore, the IFS QSAR was used to predict if the M score was low (M1), medium (M2-3) or high (M4-5). The final model predicted these scores correctly 78.6% in the training set (n=663 compounds) and 71.7% in the validation set (n=657 compounds).

For the P score, the IFS approach did not work well, and gave weak r^2 of 0.05, with no good separation between P1 and P4 compounds. There are likely many reasons why the IFS approach did not work as well for P as for M, with main ones being that processes underlying the P score are quite heterogeneous (hydrolysis, biotransformation, photransformation, etc.), whereas the underlying data for the M score (K_{OW}, K_{OC}, S_{water}) are correlated. A second reason is the general availability of experimental M data compared to P data. Thus, instead P-scores were divided into two groups: low (P score 1-2) and high (P score 3-4). For groundwater, the final model predicted these scores at 78.3% in the training set (n = 396) and 69.2% in the validation set (n = 438). For surface water,

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

Table 2. Number and distribution of REACH registered organic, organoborane, organometallic, organosilane and pseudoorganic structures (as of December 2014), as well as predicted hydrolysis products, in terms of their charge and ionizability categories.

REACH OC with CAS	Substance entries	Unique structures	Not pH dependent	Acids	Bases	Amphiprotic	Including hydrolysis products
n (and %) charge type							
neutral (pH 4-10)	2673 (48.3 %)	2601 (50.5 %)	2601 (50.5 %)	-	-	-	4158 (40.8 %)
ionizable	2283 (41.3 %)	2119 (41.1 %)	-	760 (14.7 %)	742 (14.4 %)	599 (11.6 %)	5559 (54.5 %)
ionic	574 (10.4 %)	435 (8.4 %)	111 (2.2 %)	33 (0.6 %)	17 (0.3 %)	265 (5.1 %)	481 (4.7 %)
single anions	185 (3.3 %)	145 (2.8 %)	44 (0.9 %)	0 (0.0 %)	4 (0.1 %)	94 (1.8 %)	145 (1.4 %)
multiple anions	220 (4.0 %)	145 (2.8 %)	5 (0.1 %)	28 (0.5 %)	3 (0.1 %)	108 (2.1 %)	145 (1.4 %)
single cations	106 (1.9 %)	85 (1.6 %)	59 (1.1 %)	3 (0.1 %)	0 (0.0 %)	18 (0.3 %)	105 (1.0 %)
multiple cations	22 (0.4 %)	19 (0.4 %)	3 (0.1 %)	2 (0.0 %)	10 (0.2 %)	4 (0.1 %)	24 (0.2 %)
zwitterions	41 (0.7 %)	41 (0.8 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	41 (0.8 %)	62 (0.6 %)
n (and %) organic type							
organic compounds	5175 (93.6 %)	4850 (94.1 %)	2491 (48.3 %)	777 (15.1 %)	716 (13.9 %)	839 (16.3 %)	9852 (96.6 %)
organoborates	17 (0.3 %)	16 (0.3 %)	10 (0.2 %)	3 (0.1 %)	2 (0.0 %)	1 (0.0 %)	16 (0.2 %)
organometallics	97 (1.8 %)	97 (1.9 %)	71 (1.4 %)	0 (0.0 %)	8 (0.2 %)	18 (0.3 %)	97 (1.0 %)
organosilanes	160 (2.9 %)	160 (3.1 %)	126 (2.4 %)	5 (0.1 %)	28 (0.5 %)	1 (0.0 %)	172 (1.7 %)
pseudoorganics	81 (1.5 %)	32 (0.6 %)	14 (0.3 %)	8 (0.2 %)	5 (0.1 %)	5 (0.1 %)	61 (0.6 %)
n (and %) total	5530 (100.0 %)	5155 (100.0 %)	2712 (100.0 %)	793 (100.0 %)	759 (100.0 %)	864 (100.0 %)	10198 (100.0 %)

the final model predicted these scores at 78.2% in the training set (n = 390) and 69.4% in the validation set (n = 434).

More details about the IFS QSAR calibration and validation is presented in the ESI-Section S2.

2.9. Hydrolysis products

As compounds tend to get more mobile following oxidative transformation reactions, like aerobic biotransformation and hydrolysis, it was also of relevance to consider such transformation products as part of this study. Here the LMC hydrolysis model in the QSAR toolbox was used to predict the hydrolysis structures of the reaction products. P-scores and M-scores for each of these reaction products were derived as above, which generally implied using the available QSARs or IFS QSAR, except for cases when a reaction product happened to be the same as a parent REACH OC with available experimental data.

2.10. Sensitivity analysis

The P-score and M-score are dependent on many variables and assumptions, including a) the general definition, parameters and cut-off values of the P-score and M-score, b) the prioritization of data sources and c) the accuracy of the underlying data in the prioritized data sources. Regarding a), half-life cut-off values of the P-score were based on REACH definitions, so it was not considered necessary to test the role of this cut-off. However, it was considered important to compare PMOC scores derived with PS and PG values. Therefore, PMOC-scores derived with PG-scores will primarily be presented, and compared with those derived with PS-scores as part of the sensitivity analysis. The influence on the M-score when prioritizing K_{ow} derived K_{oc} values using eq 6, instead of S_{water} when both data were available, was also investigated in the sensitivity analysis. Regarding b) we did not investigate changing the priority of the data prioritization sources (Table 1), as these were considered appropriate for making an assessment tool compatible with the REACH registration, and further the goals of this study were not primarily to validate if the peer-review of literature corresponds to those in non-peer reviewed databases (though this would be an interesting follow-up study). Regarding c) the accuracy of using QSARs was investigated, by seeing how much they deviated from experimental values from REACH registration dossiers and the peer-reviewed literature.

The endpoint parameters used in the sensitivity analysis was the number of structures obtaining a PMOC scores of 4.5 to 5 (the highest ranked PMOCs, see Figure 1), and the number of compounds that are not considered PMOCs.

3. Results and Discussion

Information about the 5155 unique REACH OC structures and their predicted hydrolysis structures, including CAS, Name, Molecular Weight, SMILES code, charge, ionization state, pK_a, substance property data and all other key information for conducting the PMOC scoring is present in the ESI-Part S2 as a Microsoft Excel file. Identities of specific substances are only provided in this text when needed for clarity.

3.1. Classification of Organic Structures in REACH

The distribution of the 5155 REACH OCs into different compound classes (organic, organoborane, organometallic, organosilane, pseudo-organic), charge categories (neutral, ionizable, cationic, anionic and zwitterionic) and ionizability

 This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Table 3. The number of REACH OC structures for which experimental and QSAR data was used for conducting the PMOC scoring. The distribution of available data across neutral, ionizable and ionic substances is also presented.

Source	Priority 1 eChemPortal experimental data (n)	Priority 2&3 Other experimental data not in eChemPortal (n)	Total Experimental (n)	Priority 4 QSAR data (n)	All Data (n)
Mobility					
pK_a	457	141	598	1198	1796
K_{oc} / D_{oc}	1015	311	1326	0[a)]	1326
K_{ow}	841	281	1122	4020	5142
$S_{water,L}$	864	657	1521	3614	5135
Persistence					
K_{aw}	464	512	976	3839	4815
v.p.	1201	503	1704	3164	4868
hydrolysis	612	0	612	1331	1943
phototransformation	85	0	85	0	85
biodegradation	888	0	888	3772	4660
Distribution (neutral/ionizable/ionic)	**(% / % / %)**	**(% / % / %)**	**(% / % / %)**	**(% / % / %)**	**(% / % / %)**
Mobility					
pK_a	16 / 71 / 13	15 / 82 / 3	16 / 73 / 11	0 / 96 / 4	5 / 89 / 6
K_{oc} / D_{oc}	60 / 35 / 5	76 / 24 / 0	63 / 32 / 4	49 / 42 / 9	52 / 39 / 8
K_{ow}	52 / 42 / 5	69 / 31 / 0	57 / 39 / 4	49 / 42 / 9	51 / 41 / 8
$S_{water,L}$	66 / 33 / 1	58 / 42 / 0	63 / 37 / 1	46 / 43 / 11	51 / 41 / 8
Persistence					
K_{aw}	67 / 32 / 0	74 / 26 / 0	71 / 29 / 0	49 / 46 / 6	53 / 42 / 5
v.p.	64 / 36 / 0	79 / 21 / 0	68 / 31 / 0	45 / 48 / 7	53 / 42 / 4
hydrolysis	65 / 25 / 10		65 / 25 / 10	67 / 33 / 1	66 / 30 / 4
phototransformation	40 / 40 / 20		40 / 40 / 20		40 / 40 / 20
biodegradation	63 / 31 / 6		63 / 31 / 6	54 / 45 / 1	56 / 42 / 2

a) QSARs for K_{oc} not considered, as these were generally based on K_{ow} and converted to K_{oc} based on eq 6.

categories (acids, bases and amphiprotic) is presented in Table 2.

Around half (50.5 %) of the unique structures were neutral organic compounds, whereas 41.1 % were ionizable and the remaining 8.4 % classified as ionic. Describing these compounds based on pH dependency (pH 4-10), showed that 52.7% were not pH dependant, 15.3% were acidic, 14.8 % were basic, and 16.7% were amphiprotic.

Franco et al. (2010)[48] performed a similar analysis, using a different methodology, on a random sample of 1510 compounds of the pre-registered REACH list in 2010. That study found a similar distribution of neutral compounds (51%) and bases (14%), but disproportionately more acids (27%) compared to the amphiprotics (8%). The different distribution of acids and amphiprotics is likely related to how the list was established, and the methodology used.

Regarding types of organic compounds, a total of 5.9% of the unique substances were not "pure" organic compounds, but consisted of organoboranes (16 structures), organometallics (97 structures), organosilanes (160 structures) and pseudo-organics (81 substances, but just 32 unique structures, mainly due to the dominance of carbonate in 34 compounds, and cyanide in 7 compounds, typically as alkali or metal salts).

3.2. Availability of P and M data

The number of unique REACH OC structures (out of 5155) for which experimental REACH dossier data (via eChemPortal) could be found to make the P- and M-scores is presented in Table 3. This only covered roughly 20% of the substances, with 1015 substance-specific K_{oc}/D_{oc} values, 457 pK_a values, 612 hydrolysis half-life values and 888 substances with biodegradability test or half-life data. This indicates that experimental data in the REACH dossiers themselves (with reliability score 1 and 2) are sufficient for conducting the proposed PMOC assessment on only a minority of REACH OC structures. Table 3 also presents available data from other experimental databases in cases experimental data could not be found in the REACH dossier data, according to the 2nd and 3rd priority of source data in Table 1. This includes 311 substance-specific K_{oc}/D_{oc} values (mostly from PP-LFER predictions), 656 S_{water} data (mostly from the database in EPI suite) and 141 pK_a values.[30, 31]

Regarding persistency parameters, some experimental data could be found for the parameters used to assess volatilization rates: K_{aw} (for 976 structures) and vapour pressure (for 1704 structures). Regarding other persistency parameters for hydrolysis, phototransformation and biodegradation, only data from the REACH dossiers were available at the time of the study, as no tabulated peer-review of half-lives could be found.

The availability of the experimental data in the REACH dossiers was related to the REACH registration requirements. For instance, substances classified as intermediates or that have

Please do not adjust margins

volumes less than 10 ton/year have reduced reporting requirements.[19] Further, QSARs can be used in specified cases during REACH registration, meaning experimental data reporting is not always a requirement.[49]

In this study, the selected P and M QSARs were able to give predictions for the majority of structures where no experimental data were found. Regarding mobility, experimental data and QSARs combined could provide a basis to evaluate mobility for all but 12 out of the 5155 unique REACH OCs. These 12 structures were all organometallics and organoboranes.

The selected QSARs could also provide a way to estimate a P-score for most of the compounds where no experimental data were available. This was particularly the case for volatilization (K_{aw} could be predicted for 3839 structures for which no experimental data were available), and biodegradation (for 3772 compounds for which no experimental data existed), followed by hydrolysis (1331 compounds).

Unlike the M-score, the P-score could not be derived for a substantial amount of substances (i.e. 420), due to lack of experimental or estimated data of both biodegradation and hydrolysis half-lives. The majority of these were ionic compounds (280 structures), and the remainder were ionizable (139 structures) or pseudoorganic (1 structure, carbon monoxide).

When looking at the distribution of experimental parameters for mobility between neutral, ionizable and ionic substances, it is also apparent from Table 3 that most data were found for neutral compounds, followed by ionizable and ionic. As a starting point to this discussion, it is important to recall the distribution of these three structure categories is 50.5%, 41.1% and 8.4%, (see Table 2). Table 3 shows that for K_{oc}/D_{oc} 63% of experimental data were for neutral compounds, 32% for ionizable compounds and 4% for ionic compounds; clearly, neutral compounds are more likely to have experimental K_{oc}/ D_{oc} data than ionic compounds. It is noted that all of the data for the ionic compounds came from REACH dossier sources. In this case K_{oc}/ D_{oc} largely originated from studies using OECD test guideline 106. Looking at other parameters, only 1% of the experimental data for S_{water} and 0% for K_{aw} were for ionic substances (the latter being less surprising as ionic substances do not volatilize from water in an ionic state).

QSARs helped provide data for many of the ionic substances. However, QSAR predictions for such ionic compounds have to be taken with some scepticism, as the low availability in general of experimental data we could obtain is indicative that they are generally not abundant in QSAR calibration data sets. It could also not be found how accurately ion-solvation interactions and ionic precipitation reactions are accounted for in the selected QSARs. An initial cause for concern was that 6% and 7% of the QSAR predictions for K_{aw} and v.p., respectively, were for ionic substances. At first glance this is surprising as ionic molecules should not volatilize from water. However, a closer look at this data shows that 97% of the K_{aw} values for ionic compounds are $< 10^{-10}$, which for practical purposes is equivalent to negligible volatilization. A partial explanation is that the QSARs may account for some of the ionic substances becoming neutral at extreme pH. It was observed that Insights and ADMET almost always gave S_{water} output for ionic compounds (>99% of them), ChemAxon often (66%), SPARC occasionally (2%). The general relative standard deviations of maximum S_{water} (pH 4 – 10) across QSARs for a given ionic substance ranged between 62% – 300% (ESI-Part S2), showing reproducibility within a factor 3 from each other, which is surprisingly consistent. The majority of these ionic substances (73%) had an average maximum S_{water} corresponding with an M-score 5, as may be expected due to the general high solubility of ionic substances. For log D_{ow}, ADMET and Insights gave predictions for over 99% of ionic compounds, ChemAxon 97%, and SPARC just 48%. The agreement of compound-specific D_{ow} values, however, was not as strong as for S_{water}, with the standard deviations ranging from 0 to 17 orders of magnitude across QSARs (for methyl sulphate and tripotassium propylsilanetriolate, respectively) with an average of 2.5 orders of magnitude; indicating that the QSARs differ more in how they account for ionic interactions with octanol than water. Most of the predicted minimum log D_{ow} (between pH 4-10) for ionic compounds corresponded with an M-score of 5 (86%). Hence, despite the lack of consistency across QSARs, they in aggregate agree that ionic substances are mobile.

For assessing the hydrolysis and biodegradation of ionic substances, a more representative portion of experimental data were available: 10% of experimental hydrolysis half-lives were for ionic compounds, and 6 % of biodegradation data were for ionic compounds. Yet, in contrast, the QSAR models used for persistency (BIOWIN and LMC) generally did not offer output for such substances (i.e. only 1% of QSAR-derived hydrolysis half-lives and biodegradation half-lives were for ionic structures). For these QSARs, most ionic substances were not included within their chemical applicability domain.

A comparison of QSAR and experimental data is presented in the next section. Compounds for which no QSAR data were available were evaluated with the original IFS QSAR for P and M-scores, as presented in section 3.4.

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

Table 4. The performance of the QSAR models used in this study compared to the obtained experimental data for mobility and volatilization parameters. Values in bold represent the best performing non-PP-LFER model

Parameter (# consistent outliers)[a]	**QSAR deviation = log (experimental value) – log (estimated value)**	**SPARC**	**ADMET**	**Chemaxon**	**Insights**	**EPISuite (Bond method)**	**PP-LFER**
pK_a	filtered average ± s.d.	**-0.1 ± 1.1**	0.1 ± 1.2	-0.1 ± 1.3	-0.1 ± 1.4		
(33)	raw average ± s.d.	**0.0 ± 1.9**	0.1 ± 1.9	0.0 ± 2.1	-0.1 ± 1.9		
	n outliers of						
	log >2 / >4 / >6	**38/17/8**	52/22/11	52/25/11	63/22/7		
	n	**318**	380	370	322		
log K_{ow} (-)	filtered average ± s.d.	0.0 ± 1.0	**0.0 ± 0.9**	0.2 ± 1.3	0.1 ± 1.2	-0.1 ± 1.2	*0.0 ± 0.3*
(41)	raw average ± s.d.	-0.2 ± 1.6	**-0.2 ± 1.3**	0.0 ± 1.7	-0.1 ± 1.5	-0.2 ± 1.6	0.0 ± 0.5
	n outliers of						
	log >2 / >4 / >6	69 / 29 / 13	**61 / 28 / 4**	87 / 30 / 13	69 / 25 / 7	82 / 34 / 10	2 / 1 / 0
	n	705	**745**	745	745	744	204
log S_{water} (mg/L)	filtered average ± s.d.	0.0 ± 1.2	**0.1 ± 0.7**	0.0 ± 0.9	0.2 ± 1.1	0.0 ± 1.0	
(24)	raw average ± s.d.	0.1 ± 1.4	**0.2 ± 0.9**	0.1 ± 1.3	0.3 ± 1.4	0.1 ± 1.3	
	n outliers of						
	log >2 / >4 / >6	61 / 19 / 8	**39 / 10 / 1**	61 / 20 / 7	91 / 21 / 8	77 / 23 / 10	
	n	905	**949**	949	949	949	
log K_{aw} (-)	filtered average ± s.d.	0.5 ± 2.9				0.3 ± 2.5	*0.5 ± 0.9*
(17)	raw average ± s.d.	0.5 ± 2.9				0.3 ± 2.6	0.5 ± 0.9
	n outliers of						
	log >2 / >4 / >6	115 / 45 / 28				134 / 52 / 29	18 / 5 / 0
	n	876				907	370
log v.p. (Pa)	filtered average ± s.d.	0.1 ± 0.9				0.1 ± 1.7	
(133)	raw average ± s.d.	0.6 ± 2.6				0.4 ± 2.6	
	n outliers of						
	log >2 / >4 / >6	197 / 94 / 57				217 / 99 / 63	
	n	1508				1598	

a) Consistent outliers were defined as those in which all tested QSAR predictions were off by two-orders of magnitude (or 3 out of 4 predictive QSARs in the case of pKa, and 6 out of 7 predictive QSARs in the case of water solubility, to account for SPARC not providing data for as many compound classes as the other QSARSs at the time of running the models).

3.3. Performance of utilized QSARs

An overview of the comparison between obtained experimental data and the QSAR models used in this study is presented in Table 4. The data are presented ias the deviation in log normalized values of the experimental and estimated value:

QSAR deviation = log (experimental value) – log (QSAR value) (9)

Average QSAR deviations were compared before and after removing "consistent outliers", these are referred to as the "raw average" and "filtered average", respectively. Consistent outliers were arbitrarily defined in this study as those in which *all* tested QSAR predictions provided data that deviated from the experimental value by two-orders of magnitude (with the exception of 3 out of 4 predictive QSARs being sufficient in the case of pK_a, and 6 out of 7 predictive QSARs in the case of S_{water}, to account for SPARC not providing data for as many compounds as the other QSARs at the time of running the models). Also shown in Table 4 is the number of compounds for which both experimental and QSAR data were available, and the number of compounds for which deviations were more than 2, 4 and 6 orders of magnitude.

For pK_a, all models gave an average QSAR deviation near 0.0, though with large standard deviations of up to two orders of magnitude. When the consistent outliers were removed (33 structures), the standard deviations were reduced to an order of magnitude. SPARC and ADMET were the best performing models, with filtered average QSAR deviations of -0.1 ± 1.1 and 0.1 ± 1.2, respectively. Based on this, pK_a values from SPARC were used when experimental data were not available, and average pK_a values from the other models were used when SPARC data were not available. These results can be compared with Liao et al. (2009),[30] which compared 9 QSARs for their ability to predict pK_a for 197 pharmaceutical substances. This study included earlier versions of SPARC, ADMET, ChemAxon (based on the Marvin program), and Insights for Excel (based on

Please do not adjust margins

Please do not adjust margins

Table 5. The performance of the QSAR models used in this study compared to the obtained experimental data for mobility and volitilization parameters.

Hydrolysis	**Hydrowin deviation**		
	Average Δ log($t_{1/2}$) ± s.d.	-0.9 ± 2.0	n = 253
	n outliers of log >2 / >4 / >6	42 / 16 / 6	
Biodegradaton	**BIOWIN & Arnot et al. (2005) deviation**		
	Average Δ log($t_{1/2}$) ± s.d.	-0.2 ± 1.1	n = 29
	n outliers of log >2 / >4 / >6	2 / 0 / 0	
	Comparison with OECD biodegradation tests		
	Correctly predicted "readily biodegradable" as < 20 days	81%	n = 776
	Correctly predicted "not readily biodegradable" as > 20 days half-life	64%	n = 938
	Total correct (overall efficiency)	72%	n = 1714

Pipeline pilot protocols), and reported mean absolute deviations of 0.65, 0.66, 0.87 and 0.77, respectively. These were substantially smaller than our mean absolute deviations for a larger, non-pharmaceutical data set of >300 compounds of 1.64, 1.64, 1.72 and 1.55, respectively. There was also differences in the number of compounds in which QSAR deviations were more than a factor 2, which in Liao et al.'s case were 12, 18, 22 and 24, respectively, and in our case 38, 52, 52 and 63, respectively. In other words, all tested QSARs seem to predict pK_a values of most substances well, though a minority of substances are poorly predicted by a substantial margin. A deeper investigation into the reasons for these deviations would make for an interesting follow-up study, but is outside the scope of the current manuscript.

K_{ow} and S_{water} predictions were in general better than for pK_a, with average deviations being close to 0 for all models (within 0.2 log units), and standard deviations ranging from ±0.9 to ±1.6 across the different QSAR models without removing consistent outliers, and from ±0.7 and ±1.2 when removing consistent outliers. ADMET was the best performing of all QSARs, having the fewest number of outliers greater than two orders of magnitude (61 out of 745 for K_{ow}, and 39 out of 949 for S_{water}), and the smallest standard deviations both before and after removing consistent outliers (with the former being -0.2 ± 1.3 for K_{ow} and 0.2 ± 0.9 for S_{water}). As a comparison, the PP-LFER approach using experimental substance descriptors data were much better performing than any of the QSARs, predicting log K_{ow} at 0.0 ± 0.5 when not removing any consistent outliers. This validates the higher prioritization of this PP-LFER prediction over QSAR estimates (Table 1).

The implications of this is that QSAR data for the mobility parameters are in general only accurate within an order of magnitude for neutral compounds. However, in some cases, and in particular for ionizable compounds with predicted pK_a values, derived D_{OC} and S_{water} can be off by two orders of magnitude or more. Therefore, considering the M-score is based largely on factor 10 increments, compounds with predicted mobility descriptors have an M-score accuracy of approximately plus/minus 1 to 2 (unless the D_{oc} or S_{water} are substantially below or above the range considered in the scoring system). This is explored in more detail as part of the sensitivity analysis (section 3.8).

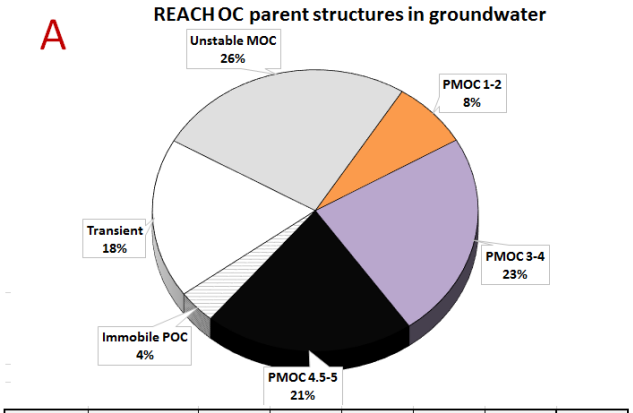
Regarding volatilization, only two QSARs from the ones selected could predict K_{aw} and vapour pressure: SPARC and EPISuite. Both performed relatively similarly for the REACH OC data, with SPARC and EPISuite predicting experimental log K_{aw} values with deviations 0.5 ± 2.9 and 0.3 ± 2.6, respectively (for circa 900 compounds). There was a large number of extreme outliers that influenced the statistics, with 45 and 52 compounds deviating by more than 4 orders of magnitude for SPARC and EPISuite, respectively.

Performance for the biodegradation and hydrolysis QSARs is presented in Table 5. The performance of hydrolysis half-life predictions using Hydrowin was poor, with average estimations of log half-lives deviating by -0.9 ± 2.0, implying that on average Hydrowin overestimated the persistency by nearly a factor 10, though with uncertainties of a factor 100 (for the 253 compounds in the validation data set). To some extent, this is due to the pH dependence of hydrolysis, which is very difficult to adequately account for in a hydrolysis QSAR model (see e.g. the help files for EPISuite HYDROWIN). For biodegradation, the QSAR approach performed well, with average estimations of log half-lives deviating by -0.2 ± 1.1 (for the only 29 compounds in the validation data set). In addition, it was investigated if the biodegradation QSARs could predict the OECD biodegradation test (301 A-F, 302 B-C and 310) results of "readily biodegradable", by comparing this experimental outcome with QSAR predicted half-lives of < 20 days. Note that when multiple OECD tests were performed and only one reported "readily biodegradable", this was considered as a "readily biodegradable". As presented in Table 5, the occurrence of an OECD test result of "readily biodegradable" was correctly predicted 80% of the time, and not "readily biodegradable" was correctly predicted 64% of the time, giving an overall efficiency of 72% for 1714 substances. When a predicted half-life of 40 days is chosen, then the agreement between OECD "readily biodegradable" increases to 94%, but predictions for not "readily biodegradable" decrease to 40%, giving an overall efficiency of 64%. The longer the half-life threshold for persistence, the more OECD "readily biodegradable" results will be predicted correctly, but the overall efficiency will decrease. In conclusion, the QSAR approach used for biodegradation is

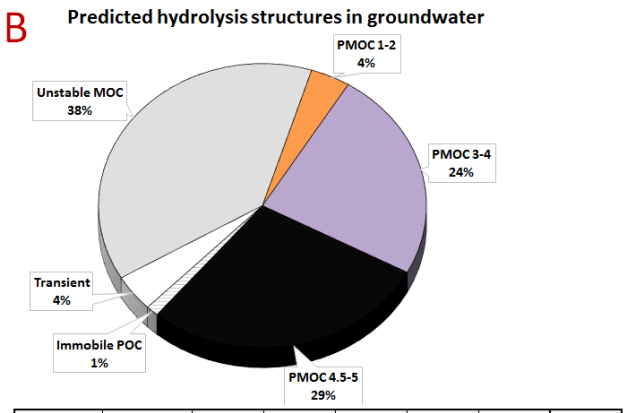
This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins



Sum	0%	9%	16%	16%	15%	45%	%
4	0%	3%	6%	5%	4%	8%	27%
3	0%	0%	1%	1%	1%	3%	7%
2	0%	1%	2%	3%	3%	8%	17%
1	0%	3%	6%	6%	6%	19%	41%
no P data	0%	0%	1%	1%	1%	6%	8%
P vs M	no M data	1	2	3	4	5	Sum



Sum	0%	1%	5%	6%	9%	76%	%
4	0%	1%	3%	2%	3%	19%	28%
3	0%	0%	0%	1%	1%	5%	7%
2	0%	0%	1%	1%	2%	17%	22%
1	0%	0%	1%	2%	3%	35%	41%
no P data	0%	0%	0%	0%	0%	0%	1%
P vs M	no M data	1	2	3	4	5	Sum

Figure 2. Distribution of PMOC and non-PMOC categories for all structures considered in this study as pie charts, as well as the distribution of P vs M-scores following the PMOC scoring chart as presented in Figure 1 for a) the 5515 unique REACH OC structures considered and b) the 5043 unique hydrolysis structures.

more successful than that for hydrolysis. However, more experimental data and further work on the development of QSARs for both processes is needed for better accuracy.

Owing to the fact that much of the data came from the REACH dossiers, it cannot be ruled out that many of the extreme outliers are due to low quality experimental data or bad reporting of experimental data in the dossiers (e.g. unit reporting errors), as presumably the scrutiny of reporting is not as stringent in the REACH dossiers as in the peer-reviewed literature. This is especially the case for persistency data, as half-lives are not as straightforward to measure or report as mobility parameters. A follow-up study investigating the cause of outliers would be of interest, as these could provide information on the quality of the data in the REACH dossiers. In cases where QSARs make consistently poor predictions of high quality data for a particular class of chemical, this could introduce opportunities to improve these QSARs. As an example of this, consider perfluorinated polar compounds, many of which would generally qualify as PMOCs. Popular QSARs originally deviated in their predictions for their mobility parameters by several orders of magnitude.[50, 51] Once the cause of this was identified,[50] modified QSARs were suggested to better account for the mobility of perfluorinated substances. [50, 51]

3.4. IFS QSAR

Of the 5155 REACH OC structures identified, there were approximately 432 compounds for which either no P-score (429 structures), M-score (12 structures) or both (9 structures) could be derived with experimental data or QSAR output. The majority (289 structures) were considered ionic. For the structures for which P could not be predicted a majority of them had an M-score of 5 (313 structures), indicating they are generally quite mobile. The IFS QSAR approach was used to estimate ranges P-scores (P1-2 and P3-4) and M-scores (M1, M2-3 and M4-5). These results were combined to make PMOC score ranges of 4.5–5 (corresponding to P3-4 and M4-5), 3-4 (corresponding to P3-4 and M2-3), and 1-2 (all other combinations) all other compounds a PMOC score of 1-2. None of the compounds were considered non-PMOCs to err on the side of caution.

3.5. Hydrolysis products

The LMC hydrolysis model generated in total 5527 unique hydrolysis structures, of which 484 were identical to parent REACH OC structures, and 5043 were not identical to other compounds on the REACH list (Table 2). The majority of the unique hydrolysis products were ionizable (3440 structures, or 68.2%), a marked increase from the parent REACH OC data set (37.3%). Very few of the new hydrolysis products were identified as ionic (46 structures, or 0.9%), a marked decrease from the distribution of ionic compounds in the parent data set (8.4 %). None of these ionic hydrolysis products were anionic (over the pH range of 4 – 10), but were either cationic or zwitterionic. This new distribution is to some extent based on what types of structures the LMC model makes predictions for, and its algorithms for making predictions; however, one can interpret this as being due to hydrolysis reactions tending to produce ionizable or polar functional groups (typically -COOH or –OH groups) rather than permanent ions.

For the majority of 5043 unique hydrolysis structures, their P and M score could be predicted using the available QSARs. The IFS QSAR was only needed for 26 unique hydrolysis structures, as few of the predicted hydrolysis products were ionic.

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

Table 6. Distribution of groundwater PMOC scores and PMOC categories across different types of neutral, ionizable and ionic REACH registered organic compounds, as well as predicted hydrolysis products. The IFS QSAR output is integrated in the PMOC scoring.

PMOC-score	All	Neutral	Ionizable	Ionic	Single anion	Single cation	Multiple anion	Multiple cation	Zwitterion	Hydrolysis Products
Immobile POC	197	104	92	1	0	0	1	0	0	62
Transient	944	832	99	13	3	5	0	0	5	211
Unstable MOC	1325	666	594	65	23	10	14	2	16	1933
PMOC 1-2	397	214	117	66	9	41	7	7	2	197
PMOC 3-4	1216	554	577	85	35	12	23	5	10	1211
PMOC 4.5-5	1076	231	640	205	75	17	100	5	8	1429
Sum	5155	2601	2119	435	145	85	145	19	41	5043
Immobile POC	4%	4%	4%	0%	0%	0%	1%	0%	0%	1%
Transient	18%	32%	5%	3%	2%	6%	0%	0%	12%	4%
Unstable MOC	26%	26%	28%	15%	16%	12%	10%	11%	39%	38%
PMOC 1-2	8%	8%	6%	15%	6%	48%	5%	37%	5%	4%
PMOC 3-4	24%	21%	27%	20%	24%	14%	16%	26%	24%	24%
PMOC 4.5-5	21%	9%	30%	47%	52%	20%	69%	26%	20%	28%
Sum PMOCs	52%	38%	63%	82%	82%	82%	90%	89%	49%	56%

3.6. The distribution of PMOC scores

The distribution of M-scores, groundwater P-scores and groundwater PMOC-scores, for the parents and hydrolysis products, is presented in Figure 2 and Table 6. A corresponding figure for surface water is presented in the ESI-Section S3. In summary, the percent distribution of groundwater PG scores are 41% (P1), 17% (P2), 7% (P3) and 27% (P4), and 8% unknown (requiring IFS QSAR estimations). The distribution of groundwater PG scores for the hydrolysis products were nearly identical, at 41% (P1), 22% (P2), 7% (P3) and 28% (P4), with 1% unknown. This provides some indication that hydrolysis products have an equal distribution of persistency as the parent REACH OC. For the surface water PS scores, the distribution is 55% (P1), 12% (P2), 5% (P3) and 21% (P4); reflecting the general expectation that compounds are less persistent in surface water by accounting for volatilization and phototransformation. If there was more phototransformation data, or a phototransformation QSAR, the distribution towards low PS-scores would be even greater.

Regarding mobility, the distribution was 9% (M1), 16% (M2), 16% (M3), 15% (M4) and 45% (M5), implying that almost half of the compounds have the highest chosen mobility ranking. Looking at the hydrolysis products, the chemicals have become considerably more mobile, at 1% (M1), 5% (M2), 6% (M3), 9% (M4) and 76% (M5). This is as expected as hydrolysis (and other forms of transformation) tend to lead to oxidative reactions that add polar and ionizable functional groups, which generally increase the solubility of a substance.

As evident from Table 6, 52% of the parent REACH OC structures were considered PMOCs, with their score distribution being 8% (PMOC-score 1-2), 24% (PMOC-score 3-4) and 21% (PMOC-score 4.5-5). The remaining compounds were not considered PMOCs, but rather immobile POCs (4%), unstable MOCs (26%) or transient (18%). Looking at the predicted hydrolysis products, there is an increase in PMOCs to 56%, and in particularly for the highest ranked PMOCs to 28% (PMOC-score 4.5-5), as well as an increase in unstable MOCs (38%). This change can be accounted for mainly by the hydrolysis products being more mobile than parent products, as just explained.

Ionic substances tend to have higher PMOC scores than ionizable and neutral substances. Only 9% of neutral compounds received PMOC-score 4.5-5, compared to 30% of the ionizable compounds and 47% of the ionic ones. The opposite trend can be seen for compounds that are not PMOCs, which covers 62% of the neutral compounds, 37% of the ionizable compounds, and only 18% of the ionic ones. Therefore, ionic compounds have the largest chance of being PMOCs; though it should be recalled that their parameters were derived with QSAR data for mobility (which ignore ion-interactions with soil and mineral surfaces) or commonly the IFS QSAR for persistence, so their PMOC scores are the most uncertain.

3.7. Sensitivity analysis

The purpose of the sensitivity analysis was to test the role of the assumptions made in the PMOC scoring system. A major assumption is that the QSARs give accurate predictions. Based on the comparison of QSAR estimations and experimental

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

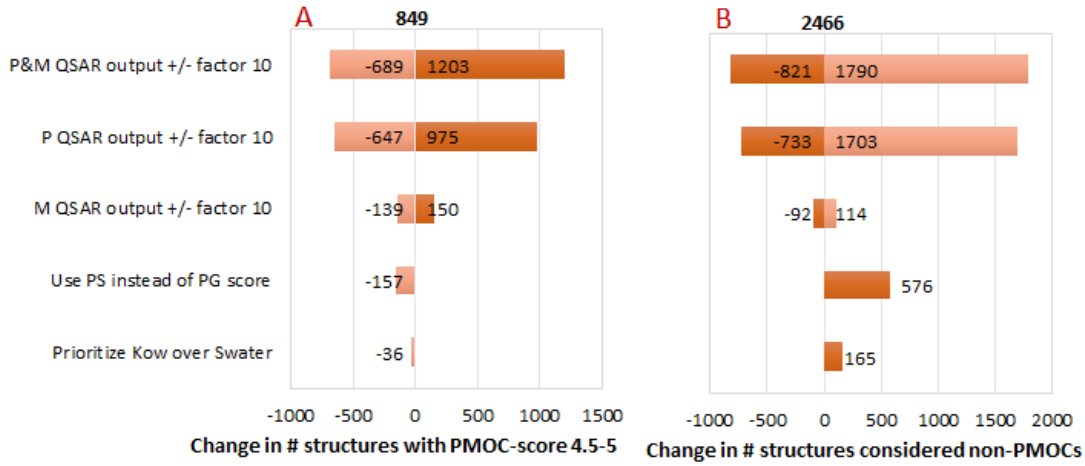


Figure 3. Sensitivity of the output parameters "# of REACH OC structures with PMOC score 4-5 -5" (Panel A) and "# of REACH OC structures not considered PMOCs" (Panel B), when deviating QSAR predictions for persistency half-lives (P-QSAR) and mobility parameters (M-QSAR) by an order of magnitude in either direction, as well as using the PS-score rather than the PG-score for persistence, or prioritizing K_{ow} over S_{water} data to derive the mobility-score.

results, we can generally conclude that QSARs are on average within a factor 10 of reported experimental data most of the time (with the exception of the hydrolysis half-life QSAR, which is even more inaccurate). Therefore, the change in PMOC-scores when increasing and decreasing QSAR predicted mobility parameters, S_{water} and K_{oc}/D_{oc}, and QSAR predicted half-lives, for biodegradation and hydrolysis, by a factor 10 was carried out. Other assumptions tested in the sensitivity analysis were a) comparing what happens when the P score is based on PS (surface water) rather than PG (ground water), and b) investigating what happens when K_{oc} values estimated from K_{ow} data were prioritized over S_{water} data (see eq 6 and related discussion) for the M-score. The endpoint of this sensitivity analysis is presented in Figure 3, showing the change in the number of substances receiving a PMOC-score 4.5-5 (panel A) or non-PMOCs (panel B).

As is evident, of all the tested considerations, the most sensitive parameter was the use of QSAR estimations for persistency half-lives. Decreasing the predicted half-lives by an order of magnitude decreased the number of PMOC-score 4.5-5 compounds from 849 to 202, while increasing the predicted half-lives, increased the number of PMOC-score 4.5-5 compounds from 849 to 1824. By comparison, changing the mobility QSAR estimations to be a factor 10 less mobile only reduced the number of PMOC-score 4.5-5 compounds from 849 to 710, while increasing by a factor 10 increased the number structures from 849 to 999. A similar trend can be seen in the number of non-PMOC compounds (originally 2466), in that a change in predicted persistency half-lives changed the number of non-PMOCs to 4169 (decreased half-life) or 1733 (increased half-life). The influence of varying mobility QSAR estimations by a factor 10 was comparatively minor, changing the number of non-PMOCs to 2374 (decreased mobility) or 2580 (increased mobility). This is largely related to how the score itself was constructed (Figure 1). The difference between P4 and P2 cut-offs are only a factor 3, whereas the cut-offs for the M-score span nearly 5 orders of magnitude. Therefore, changing a P half-life by an order of magnitude would be expected to have a huge effect, with the exception of compounds with predicted half-lives that are more than a factor 10 from the 40 day cut-off (P-score 3, see Figure 1), i.e. those with predicted half-lives > 400 days or < 4 days.

Comparing the number of PMOC-score 4.5-5 structures when using PS scores instead of PG scores, the number of compounds with PMOC-score 4.5-5 dropped, because volatilization and photo-transformation is included. However, this only had a minor influence compared to the accuracy of the P-score, causing a mere decrease by 157 structures with P-score 4.5-5, and an increase of 576 non-PMOC structures.

Finally, prioritizing K_{ow}/D_{ow} QSAR predictions over S_{water} predictions only had a minor impact on the number of structures, with 36 less compounds with P-score 4.5-5 occurring when prioritizing K_{ow}/D_{ow} (and 165 more non-PMOCs). However, this change resulted in re-scoring of several PMOCs. Therefore, in the ESI, we present two data sets of PMOC scores of REACH OC structures, those when S_{water} is prioritized over K_{ow}/D_{ow}, and those when K_{ow}/D_{ow} is prioritized over S_{water}. The small difference in the sensitivity analysis is accounted for by the pH dependent S_{water} being correlated with K_{ow}/D_{ow} to some extent. When applying this PMOC scoring system in future for other data sets, we recommend to choose either S_{water} or K_{ow}/D_{ow} based on which of these two parameters has the best data quality for a specific substance.

Please do not adjust margins

An earlier screening study to identify candidates for PBT substances from a large data set of substances reported that "uncertainty in persistence data contributes most to the uncertainty in the number of potential PBT chemicals". [8] This is indeed the case for PMOCs as well. The lack of good persistency data and estimation models has also been brought up in other contexts, such as the fate of ionizable substances in diverse water matrices[52] and "benign by design" approaches in green chemistry.[53] This study therefore joins these research efforts in underscoring the need for more research, experimental data and better tools for estimating the aquatic persistency of existing and proposed substances.

3.8. PMOCs that fall into other hazardous chemical categories

The ability of a substance to be both mobile in the aquatic environment and bioaccumulative in the food chain is not mutually exclusive. There are well-known examples of substances that are highly mobile as well as bioaccumulative (e.g. perfluorinated acids[54]). A screening study for PBT substances across diverse lists (including the European Inventory of Existing Commercial Chemical Substances, SMILES CAS database and European List of Notified Chemical Substances) identified 29 registered REACH OC and 1180 pre-registered substances (as of May 2013) that met the PBT criteria. 33 of these identified PBT substances appeared on the REACH OC data set used here (from December 2014). Nine of these PBT received a PMOC score from 4 to 5 (ESI-Part 2).

Some of the identified PMOC substances are already considered by the European Chemicals Agency (ECHA) as Substances of Very High Concern (SVHC). For instance, at the time of writing (as of March 12, 2017), in accordance with Article 59(10) of the REACH Regulation, ECHA has listed 189 SVHC on the Candidate List for Authorization (https://echa.europa.eu/candidate-list-table). Of these, 62 substances were on the REACH OC structure list used here, and 23 of these were considered PMOCs, with 17 receiving a PMOC-score from 4 to 5 (only one of which because it was a PBT, anthracene, the remainder were because they were carcinogens or toxic for reproduction) (ESI-part 2).

Zarfl and Matthies (2013)[55] carried out a screening of the registered REACH list from 2012 to see which substances satisfied proposed criteria for long-range transport potential (LRTP) (e.g. half-life in air > 2 days, or other LRTP criteria from previous studies[56, 57]) and in addition satisfied the criteria for being PB substances according to REACH. From this list, 289 substances were identified as LRTP and PB, and 268 of them are still on the registered REACH list of December 2014. Of these 268 LRTP and PB substances, 104 are considered PMOCs and 57 have PMOC-score from 4 to 5 (ESI-part 2).

Finally, it was also investigated which of the identified PMOCs were listed as drinking water contaminants. The US EPA's 2012 edition of their drinking water standards[58] lists 173 organic compounds (with a CAS number) as drinking water contaminants. 71 of them are considered in this study as REACH OC. Of these, 25 were considered non-PMOCs (23 because of short half-lives, e.g. hexane, and 2 because of low mobility, (bis(2-ethylhexyl) adipate and bis(2-ethylhexyl) phthalate)). These non-PMOCs are expected to only pose a threat to drinking water sources if the emission event is close by, due to their short half-lives or reduced mobility. The remaining 46 were considered PMOCs, with 27 receiving a PMOC score of 4-5. The organic compounds mentioned as contaminants in the EU drinking water directive[1] were all considered PMOCs (i.e. tetrachloroethene with a PMOC score 4, trichloroethene with a PMOC score 4, vinyl chloride PMOC score 3, and chloroform with a PMOC score 4.5).

In summary, many PMOCs can also meet PBT and LRTP criteria. Substances classified as all three deserve special attention, as environments and humans can be exposed to these substances through multiple, long-distance exposure routes.

4. Environmental Implications

From this screening procedure to identify and rank REACH OCs for being PMOCs, several issues related to data quality were raised. The most central one is the need for high quality, experimental persistence data and better estimation models. Therefore, all estimations of persistency presented here should be treated with some degree of scepticism. Further, mobility estimations for ionic substances are highly uncertain, due to the lack of experimental data and explicit modelling approaches that account for ionic exchange, complexation and precipitation reactions in the aquatic environment and sub-surface. We therefore encourage future research in this direction, not only to improve the identification of PMOCs, but as previously mentioned to improve similar screening procedures for PBT substances, LRTP substances and "green" chemical alternatives.[53, 59]

It could be argued that the inclusion of ionic substances in this PMOC screening was premature, and that only neutral compounds should have been considered, until more experimental data and better quality QSARs are available. However, the result of the screening show that only a minority (9%) of neutral compounds received the highest rank (PMOC-score 4.5-5), whereas a large portion of ionic compounds received the highest PMOC score (47%). This implies that not including them would be tantamount to ignoring a substantial number of PMOCs. This also implies that more experimental data and methods are needed to both measure ionic substances in the environment, and characterize their mobility and persistency.

It is hoped that models to predict persistency and mobility will improve over time. Regarding persistency, unfortunately, examples of substantial improvement in the recent literature could not be found. For mobility, however, there are more signs of progress. More experimental data and models for the mobility of ionic compounds are emerging.[21-23] Further, between the time we completed the PMOC scoring and wrote

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

this manuscript, the UFZ LSER database has undergone a substantial upgrade with new experimental data being added, along with a SMILES based K_{oc} predictor.[43] In addition to this, more substances have been registered in REACH, and more experimental data for these substances were reported. Thus, a repeat of this screening procedure in future is warranted, to account for the new substances, new data and updated QSAR/PP-LFER mobility models.

The NORMAN list of emerging substances and pollutants in the environment (downloaded from http://www.norman-network.net/ on June 18, 2015) contains 969 substances of which 213 were on the REACH OC list, 104 were considered PMOCs and 66 of these had PMOC-score from 4 to 5 (ESI-part 2). Therefore, many of these PMOC substances are known to be in the environment already (where they could be potentially impacting drinking water resources). It would be interesting to explore the reasons why some of the other PMOC substances, particularly those with high PMOC scores, are not on the NORMAN list.

One of these reasons is that it is currently difficult to screen for highly mobile compounds in the aquatic environment, particularly highly mobile ionic compounds, due to the lack of analytical techniques.[7] The list of highest ranked PMOCs amongst the REACH OCs could be useful in identifying unknown contaminants appearing in drinking water, either as part of targeted or non-targeted sampling campaigns. As an example, recently the REACH OC substance trifluoromethane sulfonic acid was identified in drinking water sources for the first time,[60] this substance was included in this study and received a PMOC score of 5. Regarding the likelihood of the occurrence of other highly ranked PMOCs from the REACH OC list of substances, and their hydrolysis products, appearing in raw water sources, it should be emphasized that the PMOC score itself just presents the hazard or potential of a substance to be in raw water. However, risk equals hazard times exposure, whereby exposure considerations would be the amount of substance used, what it is used for and its environmental emissions. In a follow up study to this one, we will propose a way forward to address exposure considerations. Future research efforts should also include toxicity, in order to develop PMT screening approaches.[11, 61]

This is the first general screening approach to identify PMOC substances that may appear in drinking water for a large data set of existing substances. The results of this study, or the approach used to rank PMOCs presented here, could be used for other chemical inventories, or proposed substances, as part of efforts to better anticipate or identify drinking water contaminants, and protect our drinking water sources.

Acknowledgements

This work is part of the PROMOTE project, as part of the European Union Joint Programming Initiative 'Water Challenges for a Changing World' (Water JPI) with financial support by the Research Council of Norway (project no. 241358/E50). The design of the PMOC screening approach was assisted by critical discussions with Michael Neumann, Daniel Sättler, Lena Vierke (Umweltbundesamt) & Stefanie Schulze, Thorsten Reemtsma (UFZ Leipzig).

Notes and references

1. EC, *Council Directive 98/83/EC of 3 November 1998 on the quality of water intended for human consumption, with its latest amendments including Commission Directive (EU) 2015/1787*, 2015.
2. WHO, *Guidelines for drinking-water quality, 4'th edition*, WHO, 2011.
3. UN, *RES/64/292. The human right to water and sanitation*, 2010.
4. D. Stepien, J. Regnery, C. Merz and W. Püttmann, *Sci. Total Environ.*, 2013, **458**, 150-159.
5. T. Reemtsma and M. Jekel, *Organic pollutants in the water cycle: properties, occurrence, analysis and environmental relevance of polar compounds*, John Wiley & Sons, 2006.
6. T. Reemtsma, S. Weiss, J. Mueller, M. Petrovic, S. González, D. Barcelo, F. Ventura and T. P. Knepper, *Environ. Sci. Technol.*, 2006, **40**, 5451-5458.
7. T. Reemtsma, U. Berger, H. P. H. Arp, H. Gallard, T. P. Knepper, M. Neumann, J. B. Quintana and P. d. Voogt, *Environ. Sci. Technol.*, 2016, **50**, 10308-10315.
8. S. Strempel, M. Scheringer, C. A. Ng and K. Hungerbühler, *Environ. Sci. Technol.*, 2012, **46**, 5680-5687.
9. S. Jensen, *Ambio*, 1972, 123-131.
10. R. Carson, *Silent spring*, Houghton Mifflin Harcourt, 2002.
11. F. Kalberlah, J. Oltmanns, M. Schwarz, J. Baumeister and A. Striffler, *Guidance for the precautionary protection of raw water destined for drinking water extraction from contaminants regulated under REACH. UFOPLAN Project FKZ 371265416.*, German Federal Environmental Agency, 2015.
12. P. Gramatica, S. Cassani and A. Sangion, *Environ. Int.*, 2015, **77**, 25-34.
13. F. Pizzo, A. Lombardo, A. Manganaro, C. I. Cappelli, M. I. Petoumenou, F. Albanese, A. Roncaglioni, M. Brandt and E. Benfenati, *Environmental Research*, 2016, **151**, 478-492.
14. E. Undeman, T. N. Brown, F. Wania and M. S. McLachlan, *Environ. Sci. Technol.*, 2010, **44**, 6249-6255.
15. J. A. Arnot, D. MacKay, E. Webster and J. M. Southwood, *Environ. Sci. Technol.*, 2006, **40**, 2316-2323.
16. J. A. Arnot, T. N. Brown, F. Wania, K. Breivik and M. S. McLachlan, *Environmental health perspectives*, 2012, **120**, 1565.
17. D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31-36.
18. D. Weininger, A. Weininger and J. L. Weininger, *Journal of chemical information and computer sciences*, 1989, **29**, 97-101.

Please do not adjust margins

Please do not adjust margins

19. ECHA, *Guidance on Information Requirements and Chemical Safety Assessment Chapter R.11: PBT/vPvB assessment. ECHA-14-G-07-EN*, 2014.
20. R. P. Schwarzenbach, P. M. Gschwend and D. M. Imboden, *Environmental Organic Chemistry*, 2. edn., John Wiley & Sons, Hoboken, 2003.
21. S. Droge and K.-U. Goss, *Environ. Sci. Technol.*, 2012, **46**, 5894-5901.
22. S. T. Droge and K.-U. Goss, *Environ. Sci. Technol.*, 2012, **47**, 798-806.
23. S. T. J. Droge and K.-U. Goss, *Environ. Sci. Technol.*, 2013, **47**, 14233-14241.
24. P. C. M. van Noort, *Chemosphere*, 2004, **56**, 7-12.
25. S. W. Karickhoff, D. S. Brown and T. A. Scott, *Water Res.*, 1979, **13**, 241-248.
26. K.-U. Goss and R. P. Schwarzenbach, *Environ. Sci. Technol.*, 2001, **35**, 1-9.
27. K. U. Goss, H. P. H. Arp, G. Bronner and C. Niederer, *Environ. Toxicol. Chem.*, 2009, **28**, 52-60.
28. G. Bronner and K.-U. Goss, *Environ. Sci. Technol.*, 2010, **45**, 1313-1319.
29. A. Springer, H. Herrmann, D. Sittner, U. Herbst and A. Schulte, *REACH Compliance: Data Availibility of REACH Registration Part 1: Screening of chemicals > 1000 tpa. UBA Report. TEXTE 43/2015*, German Federal Environmental Agency, 2015.
30. C. Liao and M. C. Nicklaus, *J. Chem Inf. Model.*, 2009, **49**, 2801-2812.
31. D. R. Lide, ed., *CRC Handbook of Chemistry and Physics.*, 85 edn., CRC Press, Boca Raton, 2004.
32. T. E. Daubert and R. P. Danner, *Physical and Thermodynamic Properties of Pure Chemicals*, Taylor and Francis, 1997.
33. M. H. Abraham, J. Andonian-Haftvan, G. S. Whiting, A. Leo and R. S. Taft, *J. Chem. Soc. Perkin Trans.*, 1994, **2**, 1777-1791.
34. E. J. Houser, T. E. Mlsna, V. K. Nguyen, R. Chung, R. L. Mowery and R. A. McGill, *Talanta*, 2001, **54**, 469-485.
35. M. H. Abraham and W. E. Acree, *New Journal Of Chemistry*, 2004, **28**, 1538-1543.
36. M. J. Kamlet, R. M. Doherty, M. H. Abraham, V. Marcus and R. W. Taft, *J. Phys. Chem.*, 1988, **92**, 5244-5255.
37. M. D. Borisover and E. R. Graber, *J. Environ. Qual.*, 1998, **27**, 312-317.
38. S. K. Poole and C. F. Poole, *J. Chromatogr. A*, 1999, **845**, 381-400.
39. J. R. Torres-Lapasio, M. C. Garcia-Alvarez-Coque, M. Roses, E. Bosch, A. M. Zissimos and M. H. Abraham, *Anal. Chim. Acta*, 2004, **515**, 209-227.
40. M. H. Abraham, H. S. Chadha, G. S. Whiting and R. C. Mitchell, *J. Pharm. Sci.*, 1994, **83**, 1085-1100.
41. C. Mintz, M. Clark, W. E. Acree and M. H. Abraham, *J. Chem Inf. Model.*, 2007, **47**, 115-121.
42. T. N. Brown, *SAR and QSAR in Environmental Research*, 2013, **25**, 51-71.
43. S. Endo, N. Watanabe, N. Ulrich, G. Bronner and K.-U. Goss, *UFZ-LSER database v 2.1 [Internet], Leipzig, Germany, Helmholtz Centre for Environmental Research-UFZ. [accessed on 02.03.2015]. Available from https://www.ufz.de/index.php?en=31698&contentonly=1&lserd_data[mvc]=Public/start*, 2015.
44. USEPA, *Estimation Programs Interface (EPI) Suite™ v4.0*, 2008.
45. M. L. Card, V. Gomez-Alvarez, W.-H. Lee, D. G. Lynch, N. S. Orentas, M. T. Lee, E. M. Wong and R. S. Boethling, *Environmental Science: Processes & Impacts*, 2017, **19**, 203-212.
46. J. Arnot, T. Gouin and D. Mackay, *Practical Methods for Estimating Environmental Biodegradation Rates, Canadian Environmental Modelling Network, Trent University, Peterborough, Ontario, Canada*, 2005.
47. T. N. Brown, J. A. Arnot and F. Wania, *Environ. Sci. Technol.*, 2012, **46**, 8253-8260.
48. A. Franco, A. Ferranti, C. Davidsen and S. Trapp, *The International Journal of Life Cycle Assessment*, 2010, **15**, 321-325.
49. ECHA, *Guidance on Information Requirements and Chemical Safety Assessment Chapter RChapter R.6: QSARs and grouping of chemicals*, 2008.
50. K. U. Goss and G. Bronner, *J. Phys. Chem. A*, 2006, **110**, 9518-9522.
51. H. P. H. Arp, C. Niederer and K. U. Goss, *Environ. Sci. Technol.*, 2006, **40**, 7298-7304.
52. T. M. Nolte and A. M. J. Ragas, *Environmental Science: Processes & Impacts*, 2017, **19**, 221-246.
53. C. Rucker and K. Kummerer, *Green Chemistry*, 2012, **14**, 875-887.
54. L. Vierke, C. Staude, A. Biegel-Engler, W. Drost and C. Schulte, *Environmental Sciences Europe*, 2012, **24**, 16.
55. C. Zarfl and M. Matthies, *Environmental Sciences Europe*, 2013, **25**, 11.
56. T. N. Brown and F. Wania, *Environ. Sci. Technol.*, 2008, **42**, 5202-5209.
57. D. C. G. Muir and P. H. Howard, *Environ. Sci. Technol.*, 2006, **40**, 7157-7166.
58. USEPA, *2012 Edition of the Drinking Water Standards and Health Advisories, EPA 822-S-12-001*, Washington, D.C., 2012.
59. H. P. H. Arp, *Environ. Sci. Technol.*, 2012, **46**, 4259-4260.
60. D. Zahn, T. Frömel and T. P. Knepper, *Water Res.*, 2016, **101**, 292-299.
61. M. Neumann, *Zbl. Geol. Paläont. Teil I*, 2017, **1**, 91-101.

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Electronic Supplementary Information

Part 1 of 2

to the article

Ranking REACH registered neutral, ionizable and ionic organic chemicals based on their aquatic persistency and mobility

by

Hans Peter H. Arp,*[a] Trevor N. Brown,[b] Urs Berger[c]

and Sarah. E. Hale[a]

a. Norwegian Geotechnical Institute, Postboks 3930 Ullevål Stadion, NO-0806 Oslo, Norway.

b. ARC Arnot Research and Consulting Inc., 5536 Sackville St., Halifax, Nova Scotia, Canada

c. Department of Analytical Chemistry, Helmholtz Centre for Environmental Research – UFZ, Permoserstr. 15, DE-04318 Leipzig, Germany.

* Corresponding author: Email: hpa@ngi.no, Tel: + 47 950 20 667

Contents: 13 pages, 12 Tables, 2 Figures

Part 2 of the ESI is a spreadsheet containing substances considered in the PMOC ranking and identification approach, as well as relevant property information and literature data.

Section S1. Volatilization half-life calculations

1) **Volatilization** – Though volatilization half-lives are not commonly reported properties for compounds in the REACH dossier or elsewhere, volatilization half-lives can be estimated based on molecular properties as well as assumptions regarding the air and water flow conditions. The most conservative assumption in terms of persistency is to assume no turbulence in the water phase (including waves) and that atmosphere contains no wind. In this conservative case, volatilization can be estimated using the following model, based on pages 914-916 in Schwarzenbach et al (2013)[1]:

$$t_{1/2,volatilization} = 0.69/(v_{aw} * h) \qquad (V1)$$

Where $t_{1/2,volatilization}$, v_{aw} is the air-water exchange velocity and h the depth of the water. The v_{aw} term is determined by the following equation:

$$1/v_{aw} = 1/v_w + 1/(v_a * K_{aw}) \qquad (V2)$$

Where v_w is the mass transfer velocity of a substance in water, the v_a is the mass transfer velocity of a substance in air and K_{aw} is the dimensionless Henry's Law constant (adjusted for 12 °C, if possible based on data availability). The term v_a at 0 m/s windspeed is calculated as

$$v_a = (D_a/D_{water\ a})^{0.67} + v_{water\ a} \qquad (V3)$$

Where D_a is the diffusion coefficient of the compound in air ($D_a = 0.26 * (MW/18)^{-0.5}$, where MW is the molecular weight), $D_{water\ a}$ is the diffusion coefficient of water vapours in air (0.26 cm^2/s), and $v_{water\ a}$ is the velocity of water vapors in air at 0 m/s wind speed (0.3 cm/s).

The term v_w at 0 m/s windspeed is calculated as

$$v_w = (Sc_w/600)^{0.67} + v_{CO2\ w} \qquad (V4)$$

Where Sc_w is the Schmidt number of the compound in water ($Sc_w = 0.00893/(0.0000192*(MW/18)^{-0.5}$ at 0 m/s wind speed) and $v_{CO2\ w}$ is the mass transfer velocity of CO_2 in water (0.00065 cm/s).

Section S2. IFS QSARs for the P and M scores

The following text is a basic description of the multiple-linear regressions of individual molecular fragments and experimentally determined P and M scores, as used to calibrate the final IFS QSAR model to estimate P and M categories. Substances that were categorized as P and M were given a PMOC score of 4/5 (due to model uncertainty), substances with P and intM or intP and M were given a score of 2/4 (due to model uncertainty), and all other substances were given a score of 1.

Interpretation of the intercepts: For the M Score, the intercept is close to the maximum M Score of 5 (4.41). This means that for very small molecules, with no fragments present in the QSAR, a prediction of "mobile" will be given by default. As more atoms are added the possibility of becoming immobile increases depending on the functional groups added. For both of the P Scores the intercept is close to the minimum score of 1 (1.20 and 1.16). This means that very small molecules with no fragments in the QSAR would be non-persistent by default, and as functional groups are added the possibility of becoming persistent increases. These results are intuitive and are consistent with other QSARs previously developed.[2,3]

Interpretation of the fragments: As discussed in previous papers, interpretation of the fragments is not always straight forward.[2,4] This is because fragments are often overlapping and the contributions from each fragment need to be properly summed to compare between different QSARs. For some complex fragments comparison between different QSARs may not be possible. As an example, the effect of an aliphatic substituted chlorine and an aromatic substituted chlorine are compared for the M and P(G) scores.:

<u>M score aliphatic Cl:</u>

fragment #16: -0.14 (any chlorine atom)

total effect from each aliphatic Cl: -0.14

<u>M score aromatic Cl:</u>

fragment #16: -0.14 (any chlorine atom)

fragment #15: +0.11 (aromatic carbon with any functional group attached)

total effect from each aromatic Cl: -0.03

In general chlorine atoms slightly decrease the M score, however, specific substitution patterns on aromatic rings also play a role. Fragment #21 adds an additional -0.24 for a specific chlorine substitution pattern. Other more general fragments (9,10,25,34,36,39) have positive or negative regression coefficients for general aromatic substitution patterns, which could include chlorine atoms. It is also debatable if the effect of fragment #18 should be included in this comparison (any carbon atom, -0.15). The overall effect of a C-Cl group will include this additional -0.15 contribution, but comparing vs. an unsubstituted carbon or comparing aromatic vs. aliphatic substitution it does not make sense to include this contribution as the effect will cancel out.

<u>P(G) score aliphatic Cl:</u>

fragment #3: +2.49 (any aliphatic atom)

fragment #43: -1.79 (any chlorine atom)

total effect from each aliphatic Cl: +0.7

P(G) score aromatic Cl:

fragment #3: +2.49 (any aliphatic atom)

fragment #5: +1.45 (aromatic carbon - chlorine bond)

fragment #43: -1.79 (any chlorine atom)

total effect from each aromatic Cl: +2.15

Accounting for the total effect of a C-Cl group is more complicated than the M score. To include the carbon atom these additional factors need to be included:

P(G) score aliphatic C of C-Cl group:

fragment #3: +2.49 (any aliphatic atom)

fragment #10: +0.65 (any carbon atom)

fragment #38: -1.19 (any bond between atoms)

total effect from each aliphatic Cl: +0.7 +1.95 = 2.65

P(G) score aromatic C of c-Cl group:

fragment #2: +2.51 (any aromatic atom)

fragment #10: +0.65 (any carbon atom)

fragment #38: -1.19 (any bond between atoms)

total effect from each aromatic Cl: +2.15 +1.97 = 4.12

A more relevant comparison may be the effect of replacing one aliphatic or aromatic hydrogen with one chlorine:

P(G) score remove one aliphatic hydrogen:

fragment #29: -(-0.57) (bond between hydrogen and an aliphatic carbon)

fragment #40: -(-1.45) (any hydrogen atom)

total effect from each aliphatic Cl: +0.7 +2.02 = 2.72

P(G) score remove one aromatic hydrogen:

fragment #8: -(+0.68) (bond between hydrogen and an aromatic carbon)

fragment #40: -(-1.45) (any hydrogen atom)

total effect from each aromatic Cl: +2.15 +0.77 = 2.92

This calculation may be even more complex if the type of aliphatic carbon is changed by removing a hydrogen, for example if the carbon becomes a quaternary carbon (fragment #16) an additional +0.27 contribution is gained, or if the carbon was a methyl group (fragment #22) then the contribution of -0.15 is lost.

Interpretation of regression coefficients: In general the regression coefficients of M score seem reasonable. However, the P(G) and P(S) scores have both large positive and large negative regression coefficients that tend to balance out to give a score in the range 1-4. This is a common sign of over-fitting, and the prediction results may be especially unstable as the regression models are extrapolated outside of their training domain. The predictions from these QSARs should be treated with careful skepticism.

S2.1 PMOC QSAR Validation Statistics

Table S1: M Score Training Dataset Results Summary

	Predicted Mobile	Predicted Intermediate	Predicted Not Mobile
Expected Mobile	148	22	0
Expected Intermediate	67	300	33
Expected Not Mobile	0	20	73

Table S2: M Score Validation Dataset Results Summary

	Predicted Mobile	Predicted Intermediate	Predicted Not Mobile
Expected Mobile	109	26	3
Expected Intermediate	90	302	43
Expected Not Mobile	5	19	60

Table S3: M Score Training and Validation Summary Statistics

	Training Dataset	Validation Dataset
% Mobile Predictions Correct	68.8	53.4
% False Positives[a]	31.2	46.6
% Intermediate Predictions Correct	87.7	87
% Intermediate False Negatives[b]	6.4	7.5
% Not Mobile Predictions Correct	68.9	56.6
% Not Mobile False Negatives[c]	0	2.8
% Total Correct	78.6	71.7

[a] Mobile predictions for chemicals expected to be intermediate or not mobile.
[b] Intermediate predictions for chemicals expected to be mobile. Remainder are intermediate predictions for chemicals expected to be not mobile.
[c] Not mobile predictions for chemicals expected to be mobile. Remainder are not mobile predictions for chemicals expected to be intermediate.

Table S4: P(G) Score Training Dataset Results Summary

	Predicted Persistent	Predicted Labile
Expected Persistent	79	34
Expected Labile	52	231

Table S5: P(G) Score Validation Dataset Results Summary

	Predicted Persistent	Predicted Labile
Expected Persistent	67	66
Expected Labile	69	236

Table S6: P(G) Score Training and Validation Summary Statistics

	Training Dataset	Validation Dataset
% Persistent Predictions Correct	60.3	49.3
% False Positives[a]	39.7	50.7
% Labile Predictions Correct	87.2	78.1
% False Negatives[b]	12.8	21.9
% Total Correct	78.3	69.2

[a] Persistent predictions for chemicals expected to be labile.
[b] Labile predictions for chemicals expected to be persistent.

Table S7: P(S) Score Training Dataset Results Summary

	Predicted Persistent	Predicted Labile
Expected Persistent	67	36
Expected Labile	49	238

Table S8: P(S) Score Validation Dataset Results Summary

	Predicted Persistent	Predicted Labile
Expected Persistent	63	63
Expected Labile	70	238

Table S9: P(S) Score Training and Validation Summary Statistics

	Training Dataset	Validation Dataset
% Persistent Predictions Correct	57.8	47.4
% False Positives[a]	42.2	52.6
% Labile Predictions Correct	86.9	79.1
% False Negatives[b]	13.1	20.9
% Total Correct	78.2	69.4

[a] Persistent predictions for chemicals expected to be labile.
[b] Labile predictions for chemicals expected to be persistent.

S2.2 IFS QSAR Regression coefficients for the selected fragments

Table S10: M Score QSAR fragments and coefficients

#	Description	SMARTS code	Regression Coefficient	Std. Err.
1	aromatic nitrogen with hydrogen	[nX3H1+0]	1.04	0.25
2	chloro-aldehyde	[ClX1H0]-[CX3H0]=[OX1H0+0]	0.74	0.25
3	carbon-nitrogen double bond attached to an alkyl chain	[CX4H2]-[CX4H2]-[CX3H0]=[NX2H0+0]	0.63	0.24
4	ethene group, one substituent on either side	[CX3H1]=[CX3H1]	0.60	0.12
5	ethyne group	[CX2H0]#[CX2H0]	0.57	0.29
6	aromatic nitro group, two unsubstituted neighbouring carbons	[cX3H1]:[cX3H1]:[cX3H0;$(*-A)]-[NX3H0+0](=[OX1H0+0])=[OX1H0+0]	0.52	0.20
7	any oxygen atom	[5]	0.51	0.04
8	aromatic methoxy group	[CX4H3]-[OX2H0+0]-[cX3H0;$(*-A)]	0.48	0.12
9	aromatic methyl group beside another aliphatic substituent	[CX4H3]-[cX3H0;$(*-A)]:[cX3H0;$(*-A)]:[cX3H1]	0.45	0.17
10	aromatic secondary amine, para to another aliphatic substituent	[NX3H1+0]-[cX3H0;$(*-A)]:[cX3H1]:[cX3H1]:[cX3H0;$(*-A)]:[cX3H1]	0.36	0.08
11	aliphatic ether with a neighbouring substitution	[CX4H2]-[CX4H1]-[OX2H0+0]	0.26	0.05
12	alcohol group attached to quaternary carbon	[CX4H3]-[CX4H0]-[OX2H1+0]	0.23	0.11
13	aliphatic ether	[CX4H2]-[OX2H0+0]	0.21	0.03
14	any nitrogen atom	[6]	0.11	0.04
15	aromatic carbon with any aliphatic substituent	[cX3H0;$(*-A)]	0.11	0.03
16	any chlorine atom	[ClX1H0]	-0.14	0.06
17	two fused aromatic aromatic carbons with three neighboring unsubsituted positions	[cX3H0;!$(*-a);!$(*~A)]:[cX3H0;!$(*-a);!$(*~A)]:[cX3H1]:[cX3H1]:[cX3H1]	-0.15	0.07
18	any carbon atom	[7]	-0.15	0.01
19	ester group	[OX2H0+0]-[CX3H0]=[OX1H0+0]	-0.22	0.06
20	aromatic tertiary carbon with no ortho, meta or para substituents	[CX4H0]-[cX3H0;$(*-A)]:[cX3H1]:[cX3H1]:[cX3H1]	-0.23	0.13
21	aromatic chlorine with no ortho or meta substituents on one side	[ClX1H0]-[cX3H0;$(*-A)]:[cX3H1]:[cX3H1]	-0.24	0.11
22	isopropyl group	[CX4H3]-[CX4H1]-[CX4H3]	-0.26	0.08
23	any bromine atom	[BrX1H0]	-0.27	0.06
24	aliphatic alcohol group with three neighbouring substituents	[CX4H1]-[CX4H1]-[CX4H1]-[OX2H1+0]	-0.27	0.07
25	aromatic alkyl chain with aliphatic para substituent	[CX4H2]-[cX3H0;$(*-A)]:[cX3H1]:[cX3H1]:[cX3H0;$(*-A)]	-0.27	0.10
26	quaternary cabon with a methyl and an isopropyl attached	[CX4H3]-[CX4H0]-[CX4H1](-[CX4H2])-[CX4H2]	-0.31	0.09
27	n-butyl group (also counted for any longer chains)	[CX4H2]-[CX4H2]-[CX4H2]-[CX4H3]	-0.32	0.07
28	silicon with no hydrogens attached	[SiX4H0]	-0.36	0.10
29	linear alkyl chain with some substitutions	[CX4H1]-[CX4H2]-[CX4H1]-[CX4H1]-[CX4H2]	-0.38	0.17
30	terminal ethene group, aliphatic attachment	[CX4H2]-[CX3H1]=[CX3H2]	-0.38	0.14

31	any double bonded pair of aliphatic atoms in a ring	[A!#1x2+0]=[A!#1x2+0]	-0.42	0.15
32	aromatic carbonyl group with no ortho or para substituents	[OX1H0+0]=[CX3H0]-[cX3H0;$(*-A)](:[cX3H1]:[cX3H1]):[cX3H1]:[cX3H1]	-0.42	0.12
33	any ether	[OX2H0+0]	-0.48	0.06
34	aromatic ether with one ortho substituent and no meta substiuents	[cX3H1]:[cX3H0;$(*-A)]:[cX3H0;$(*-A)](-[OX2H0+0]):[cX3H1]:[cX3H1]	-0.48	0.16
35	ethene group, alkyl chain on one side and two substituents on the other	[CX4H2]-[CX4H2]-[CX3H1]=[CX3H0]	-0.53	0.22
36	aromatic primary amine with a meta substituent	[NX3H2+0]-[cX3H0;$(*-A)]:[cX3H1]:[cX3H0;$(*-A)]	-0.58	0.23
37	trifluoro methyl group	[FX1H0]-[CX4H0](-[FX1H0])-[FX1H0]	-0.66	0.20
38	sulfate group	[OX1H0+0]=[SX4H0]=[OX1H0+0]	-0.80	0.22
39	aromatic nitro group with one ortho substituent	[cX3H1]:[cX3H1]:[cX3H0;$(*-A)]:[cX3H0;$(*-A)](:[cX3H1])-[NX3H0+0](=[OX1H0+0])=[OX1H0+0]	-0.93	0.32
40	any nitro group	[OX1H0+0]=[NX3H0+0]=[OX1H0+0]	-1.51	0.16
	intercept		4.41	0.07

Table S11: P(G) Score QSAR fragments and coefficients

#	Description	SMARTS code	Regression Coefficient	Std. Err.
1	any boron	[8]	3.02	1.63
2	any aromatic atom	[a]	2.51	1.42
3	any aliphatic atom	[A]	2.49	1.30
4	any fused aromatic carbon	[cX3H0;!$(*-a);!$(*~A)]	2.18	0.74
5	aromatic chlorine	c-Cl	1.45	0.47
6	siloxane (Si-O-Si)	[SiX4]-[OX2H0]-[SiX4]	1.34	0.52
7	sulfur with aromatic attachment	S-c	0.75	0.86
8	aromatic carbon - hydrogen bond	c-[9]	0.68	0.73
9	three neighbouring substituted aromatic carbons	[cX3H1]:[cX3H1]:[cX3H1]	0.65	0.41
10	any carbon	[7]	0.65	1.08
11	any aliphatic nitrogen attached to an aromatic carbon	c-N	0.54	0.40
12	tertiary amine with any three carbon attachments	[7]-[NX3](-[7])-[7]	0.50	0.48
13	aromatic ether	c-O	0.39	0.33
14	phosphorus-oxygen single bond	P-O	0.37	0.47
15	aromatic-aliphatic carbon carbon bond	c-C	0.35	0.41
16	quaternary carbon	[CX4H0]	0.27	0.15
17	aliphatic ketone	CC(=O)C	0.22	0.26
18	aliphatic ester	CC(=O)OC	0.15	0.17
19	any aromatic attached group also double bonded to oxygen	[cX3H0;$(*-A)]-*=[OX1H0+0]	0.12	0.24
20	non-terminal propyl chain	[CX4H2]-[CX4H2]-[CX4H2]	0.10	0.08
21	number of rings		0.07	0.06
22	methyl	[CX4H3]	-0.15	0.11
23	non-terminal ethyl chain	[CX4H2]-[CX4H2]	-0.15	0.10
24	anhydrous phthalate group	[OX1H0+0]=[CX3H0]-[OX2H0+0]-[CX3H0]=[OX1H0+0]	-0.32	0.58
25	aliphatic secondary amine	C[NH]C	-0.36	0.41
26	aliphatic carbon-carbon bond	C-C	-0.45	0.25
27	carbon-nitrogen double bond	C=N	-0.48	0.76

28	nitrogen-oxygen double bond	O=N	-0.52	0.50
29	aliphatic carbon - hydrogen bond	C-[9]	-0.57	0.36
30	aliphatic ether	COC	-0.68	0.26
31	aliphatic primary amine	C[NH2]	-0.79	0.59
32	nitrogen-nitrogen single bond	N-N	-0.82	0.67
33	any oxygen	[5]	-0.97	0.91
34	any nitrogen	[6]	-1.02	0.88
35	ortho unsubstituted aromatic carbons	[cX3H1]:[cX3H1]	-1.04	0.77
36	alcohol group	OH	-1.04	0.37
37	peroxy group	O-O	-1.06	0.39
38	any bond	*~*	-1.19	0.57
39	cabon-nitrogen aromatic bond	c:n	-1.24	0.77
40	any hydrogen	[9]	-1.45	1.19
41	carbon-oxygen aromatic bond	c:o	-1.65	0.81
42	any bromine	[10]	-1.66	1.14
43	any chlorine	[11]	-1.79	1.07
44	any fluorine	[12]	-1.94	1.07
45	carbon-carbon double bond	C=C	-1.95	0.79
46	any sulfur	[13]	-1.99	1.02
47	any silicon	[14]	-1.99	0.83
48	carbon-carbon aromatic bond	c:c	-1.99	0.93
49	carbonyl group	C=O	-2.10	0.66
50	cyano group	C#N	-2.43	1.25
51	phosphorus-oxygen double bond	P=O	-2.48	1.18
52	any iodine	[15]	-2.52	2.07
53	carbon-carbon triple bond	C#C	-3.23	1.61
	intercept		1.20	0.19

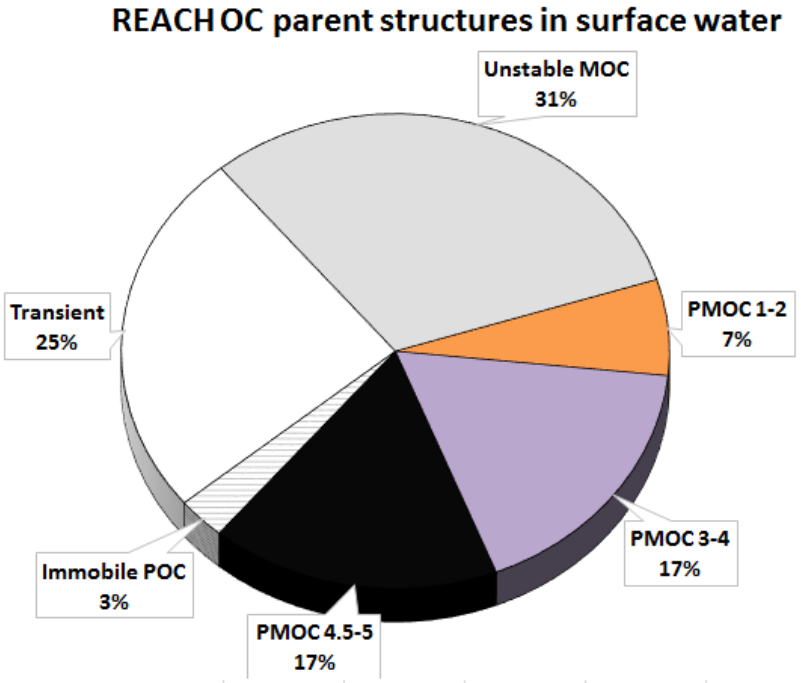
Table S12: P(S) Score QSAR fragments and coefficients

#	Description	SMARTS code	Regression Coefficient	Std. Err.
1	aromatic chlorine	c-Cl	3.93	1.23
2	any fused aromatic carbon	[cX3H0;!$(*-a);!$(*~A)]	3.78	1.22
3	single bond between two aromatic carbons (eg. biphenyl bridge)	c-c	3.40	1.97
4	sulfur with aromatic attachement	S-c	2.70	1.24
5	aromatic carbon - hydrogen bond	c-[9]	2.30	1.02
6	aromatic-aliphatic carbon carbon bond	c-C	2.29	0.97
7	aromatic ether	c-O	1.88	0.93
8	any aliphatic nitrogen attached to an aromatic carbon	c-N	1.71	0.97
9	any aliphatic atom	[A]	1.49	1.02
10	sulfur double bonded to carbon	S=C	1.47	1.05
11	any boron	[8]	1.25	1.40
12	any aromatic atom	[a]	1.09	1.59
13	tertiary amine with any three carbon attachments	[7]-[NX3](-[7])-[7]	0.93	0.56
14	aliphatic chlorine	C-Cl	0.79	0.71
15	aromatic primary or secondary amine	c[NX3;H1,H2]	0.65	0.56
16	three neighbouring unsubstituted aromatic carbons	[cX3H1]:[cX3H1]:[cX3H1]	0.60	0.74
17	any phosphorus	[16]	0.52	1.40

18	four neighbouring unsubstituted aromatic carbons	[cX3H1]:[cX3H1]:[cX3H1]:[cX3H1]	0.45	0.92
19	quaternary carbon	[CX4H0]	0.37	0.28
20	carbon-nirtogen double bond	C=N	0.29	0.39
21	sulfur single bonded to an aliphatic carbon	S-C	0.22	0.44
22	aliphatic ketone	CC(=O)C	0.20	0.23
23	aliphatic ester	CC(=O)OC	0.16	0.14
24	aliphatic alcohol	C[OH]	0.16	0.38
25	silicon with single bond to aliphatic carbon	[Si]-C	0.16	0.37
26	CH2 group	[CX4H2]	0.06	0.09
27	CH1 group	[CX4H1]	0.05	0.18
28	number of rings	rings	0.02	0.06
29	aliphatic tertiary amine	CN(C)C	-0.20	0.56
30	any bond	*~*	-0.20	0.34
31	cyano group	C#N	-0.21	0.42
32	aliphic primary amine	C[NH2]	-0.22	0.45
33	anhydrous phthalate group	[OX1H0+0]=[CX3H0]-[OX2H0+0]-[CX3H0]=[OX1H0+0]	-0.29	0.54
34	nitrogen-oxygen double bond	O=N	-0.34	0.39
35	aliphic ether	COC	-0.37	0.21
36	nitrogen-nitrogen single bond	N-N	-0.68	0.60
37	alcohol	O-[9]	-0.69	0.36
38	carbonyl group	C=O	-0.77	0.27
39	any oxygen	[5]	-0.79	0.87
40	carbon-oxygen aromatic bond	c:o	-0.87	1.23
41	cabon-nitrogen aromatic bond	c:n	-0.94	0.92
42	peroxy group	O-O	-0.99	0.45
43	any bromine	[10]	-1.00	0.99
44	any nitrogen	[6]	-1.04	0.85
45	ortho unsubstituted aromatic carbons	[cX3H1]:[cX3H1]	-1.14	0.83
46	any carbon	[7]	-1.22	0.83
47	any fluorine	[12]	-1.26	0.93
48	any hydrogem	[9]	-1.33	0.89
49	carbon-carbon aromatic bond	c:c	-1.52	1.04
50	any iodine	[15]	-1.70	1.89
51	any silicon	[14]	-1.82	1.17
52	and sulfur	[13]	-1.90	1.02
53	any chlorine	[11]	-2.16	1.15
54	phosphorus-oxygen double bond	P=O	-2.47	1.25
	intercept		1.16	0.19

Section S3. Distribution of PS, M and PMOC scores for surface water

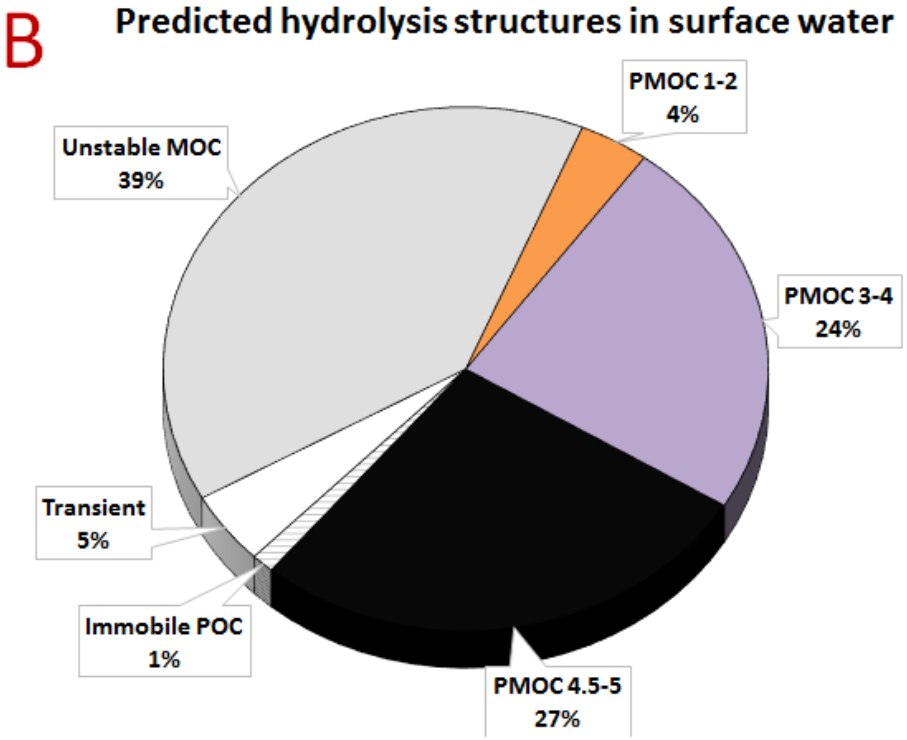
A



Sum		443	823	812	767	2298	*n*
P4	1	138	218	177	**151**	**379**	1064
P3	0	10	38	48	**36**	**126**	258
P2	1	24	60	87	106	315	593
P1	1	259	475	472	440	1164	2811
no P data	9	12	32	28	34	314	429
P vs M	**no M data**	**M1**	**M2**	**M3**	**M4**	**M5**	**Sum**
Sum	0%	9%	16%	16%	15%	45%	%
4	0%	3%	4%	3%	**3%**	**7%**	21%
3	0%	0%	1%	1%	**1%**	**2%**	5%
2	0%	0%	1%	2%	2%	6%	12%
1	0%	5%	9%	9%	9%	23%	55%
no P data	0%	0%	1%	1%	1%	6%	8%
P vs M	**no M data**	**1**	**2**	**3**	**4**	**5**	**Sum**

Figure S3A. Distribution of PMOC and non-PMOC categories in surface water for all structures considered in this study as pie charts, as well as the distribution of P vs M-scores following the PMOC scoring chart as presented in Figure 1 for the 5515 unique REACH OC structures considered.

B



Sum		75	276	325	452	3914	*n*
P4	0	55	143	115	**153**	**921**	1387
P3	0	6	14	38	**50**	**251**	359
P2	0	4	38	73	98	882	1095
P1	0	10	80	95	144	1847	2176
no P data	1	0	1	4	7	13	26
P vs M	**no M data**	**M1**	**M2**	**M3**	**M4**	**M5**	**Sum**
Sum	0%	1%	5%	6%	9%	76%	*%*
4	0%	1%	3%	2%	**3%**	**18%**	27%
3	0%	0%	0%	1%	**1%**	**5%**	7%
2	0%	0%	1%	1%	2%	17%	21%
1	0%	0%	2%	2%	3%	36%	42%
no P data	0%	0%	0%	0%	0%	0%	1%
P vs M	**no M data**	**1**	**2**	**3**	**4**	**5**	**Sum**

Figure S3B. Distribution of PMOC and non-PMOC categories in surface water for all structures considered in this study as pie charts, as well as the distribution of P vs M-scores following the PMOC scoring chart as presented in Figure 1 for the 5043 unique hydrolysis structures.

References

1. Schwarzenbach, R. P.; Gschwend, P. M.; Imboden, D. M., *Environmental Organic Chemistry*. 2. ed.; John Wiley & Sons: Hoboken, 2003.
2. Brown, T. N.; Arnot, J. A.; Wania, F., Iterative fragment selection: A group contribution approach to predicting fish biotransformation half-lives. *Environ. Sci. Technol.* **2012,** *46*, (15), 8253-8260.
3. Arnot, J. A.; Brown, T. N.; Wania, F., Estimating screening-level organic chemical half-lives in humans. *Environ. Sci. Technol.* **2013,** *48*, (1), 723-730.
4. Brown, T. N., Predicting hexadecane–air equilibrium partition coefficients (L) using a group contribution approach constructed from high quality data. *SAR and QSAR in Environmental Research* **2013,** *25*, (1), 51-71.
5. Breedveld, G. D.; Pelletier, E.; St Louis, R.; Cornelissen, G., Sorption characteristics of polycyclic aromatic hydrocarbons in aluminum smelter residues. *Environ. Sci. Technol.* **2007,** *41*, (7), 2542-2547.
6. Cornelissen, G.; Pettersen, A.; Broman, D.; Mayer, P.; Breedveld, G. D., Field testing of equilibrium passive samplers to determine freely dissolved native polycyclic aromatic hydrocarbon concentrations. *Environ. Toxicol. Chem.* **2008,** *27*, (3), 499-508.
7. Cornelissen, G.; Pettersen, A.; Nesse, E.; Eek, E.; Helland, A.; Breedveld, G. D., The contribution of urban runoff to organic contaminant levels in harbour sediments near two Norwegian cities. *Mar. Pollut. Bull.* **2008,** *56*, (3), 565-573.
8. Barthe, M.; Pelletier, E.; Breedveld, G. D.; Cornelissen, G., Passive samplers versus surfactant extraction for the evaluation of PAH availability in sediments with variable levels of contamination. *Chemosphere* **2008,** *71*, (8), 1486-1493.
9. Cornelissen, G.; Wiberg, K.; Broman, D.; Arp, H. P. H.; Persson, Y.; Sundqvist, K.; Jonsson, P., Freely Dissolved Concentrations and Sediment-Water Activity Ratios of PCDD/Fs and PCBs in the Open Baltic Sea. *Environ. Sci. Technol.* **2008,** *42*, (23), 8733-8739.
10. Yeo, H. G.; Choi, M.; Chun, M. Y.; Sunwoo, Y., Gas/particle concentrations and partitioning of PCBs in the atmosphere of Korea. *Atmos. Environ.* **2003,** *37*, (25), 3561-3570.
11. van Noort, P. C. M.; Cornelissen, G.; ten Hulscher, T. E. M.; Vrind, B. A.; Rigterink, H.; Belfroid, A., Slow and very slow desorption of organic compounds from sediment: influence of sorbate planarity. *Water Res.* **2003,** *37*, (10), 2317-2322.
12. Oen, A. M. R.; Breedveld, G. D.; Kalaitzidis, S.; Christanis, K.; Cornelissen, G., How quality and quantity of organic matter affect polycyclic aromatic hydrocarbon desorption from Norwegian harbor sediments. *Environ. Toxicol. Chem.* **2006,** *25*, (5), 1258-1267.
13. Cornelissen, G.; Elmquist, M.; Groth, I.; Gustafsson, O., Effect of sorbate planarity on environmental black carbon sorption. *Environ. Sci. Technol.* **2004,** *38*, (13), 3574-3580.
14. Cornelissen, G.; Breedveld, G. D.; Kalaitzidis, S.; Christanis, K.; Kibsgaard, A.; Oen, A. M. P., Strong sorption of native PAHs to pyrogenic and unburned carbonaceous geosorbents in sediments. *Environ. Sci. Technol.* **2006,** *40*, (4), 1197-1203.
15. Arp, H. P. H.; Schwarzenbach, R. P.; Goss, K. U., Equilibrium sorption of gaseous organic chemicals to fiber filters used for aerosol studies. *Atmos. Environ.* **2007,** *41*, (37), 8241-8252.
16. Schenker, U.; MacLeod, M.; Scheringer, M.; Hungerbuhler, K., Improving data quality for environmental fate models: A least-squares adjustment procedure for harmonizing physicochemical properties of organic compounds. *Environ. Sci. Technol.* **2005,** *39*, (21), 8434-8441.