CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Epsilon Open Archive

CSIRO PUBLISHING

Environ. Chem. **2020**, *17*, 498–508
https://doi.org/10.1071/EN19296

Research Paper

Investigating the OECD database of per- and polyfluoroalkyl substances – chemical variation and applicability of current fate models

Ioana C. Chelcea,[A] *Lutz Ahrens,*[B] *Stefan Örn,*[C] *Daniel Mucs*[D] *and Patrik L. Andersson*[A,E]

[A]Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden.
[B]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences (SLU), Box 7050, SE-750 07 Uppsala, Sweden.
[C]Department of Biomedical Sciences and Veterinary Public Health, Swedish University of Agricultural Sciences (SLU), SE-750 07 Uppsala, Sweden.
[D]RISE SP – Chemical and Pharmaceutical Safety, Forskargatan 20, 151 36 Södertälje, Sweden.
[E]Corresponding author. Email: patrik.andersson@umu.se

Environmental context. A diverse range of materials contain organofluorine chemicals, some of which are hazardous and widely distributed in the environment. We investigated an inventory of over 4700 organofluorine compounds, characterised their chemical diversity and selected representatives for future testing to fill knowledge gaps about their environmental fate and effects. Fate and property models were examined and concluded to be valid for only a fraction of studied organofluorines.

Abstract. Many per- and polyfluoroalkyl substances (PFASs) have been identified in the environment, and some have been shown to be extremely persistent and even toxic, thus raising concerns about their effects on human health and the environment. Despite this, little is known about most PFASs. In this study, the comprehensive database of over 4700 PFAS entries recently compiled by the OECD was curated and the chemical variation was analysed in detail. The analysis revealed 3363 individual PFASs with a huge variation in chemical functionalities and a wide range of mixtures and polymers. A hierarchical clustering methodology was employed on the curated database, which resulted in 12 groups, where only half were populated by well-studied compounds thus indicating the large knowledge gaps. We selected both a theoretical and a procurable training set that covered a substantial part of the chemical domain based on these clusters. Several computational models to predict physicochemical and environmental fate related properties were assessed, which indicated their lack of applicability for PFASs and the urgent need for experimental data for training and validating these models. Our findings indicate reasonable predictions of the octanol-water partition coefficient for a small chemical domain of PFASs but large data gaps and uncertainties for water solubility, bioconcentration factor, and acid dissociation factor predictions. Improved computational tools are necessary for assessing risks of PFASs and for including suggested training set compounds in future testing of both physicochemical and effect-related data. This should provide a solid basis for better chemical understanding and future model development purposes.

Received 8 November 2019, accepted 3 February 2020, published online 23 April 2020

Introduction

Per- and polyfluoroalkyl substances (PFASs) are man-made high production-volume chemicals that have been used in industry and consumer products worldwide since the 1950s (Buck et al. 2011). PFASs are a broad group of chemicals with different properties and applications. The chemical and biological knowledge is generally based on rather a limited number of specific substances given the large number and range of different PFASs. Numerous PFASs are being used in non-stick cookware, water-repellent clothing, stain-resistant fabrics and carpets, cosmetics, firefighting foams, and products that resist grease, water, and oil (Buck et al. 2011). These compounds have received increasing public attention owing to their persistence, bioaccumulation, and possible adverse effects in humans and wildlife (Ahrens and Bundschuh 2014; Giesy et al. 2010; Martin et al. 2003; Patlewicz et al. 2019). Certain PFASs are ubiquitous chemicals in the environment and have been detected in the air, surface water, groundwater, soil, sediment, biota, and food (Blaine et al. 2014; Gewurtz et al. 2013). Shorter chain PFASs ($C < 8$) dominate in the aqueous phase and in plants (Ahrens and Bundschuh 2014; Gobelius et al. 2017), whereas longer chain PFASs ($C \geq 8$) are mainly associated with soils and sediments (Higgins and Luthy 2006). However, PFASs, such as fluorotelomer alcohols (FTOHs), fluorotelomer acrylates (FTACs), perfluorooctane sulfonamides (FOSAs), and sulfonamidoethanols (FOSEs), are generally more volatile and can be transported in the atmosphere (Ahrens et al. 2011; Jahnke et al. 2007). Some of these PFASs can be degraded via gas-phase peroxy radical

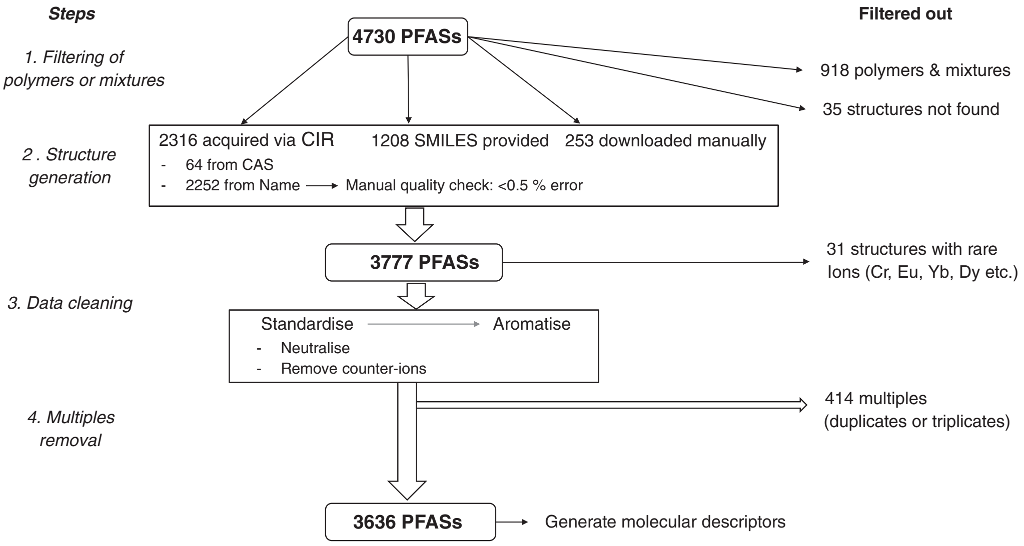


Fig. 1. PFASs database pre-processing methodology.

cross-reactions or bio-degraded under aerobic and anaerobic conditions to perfluoroalkyl carboxylates (PFCAs) or perfluoroalkane sulfonates (PFSAs), which result in degradation products with significantly higher persistence in the environment (Ellis et al. 2004; Lee et al. 2010; Liu and Mejia Avendaño 2013). Production of PFASs has changed significantly in recent years after the voluntary phasing-out and banning of the C8-based PFASs, such as perfluorooctane sulfonate (PFOS) and perfluorooctanoic acid (PFOA), and manufacturers have shifted to alternative fluorinated substances (Mejia-Avendaño et al. 2017; Xiao et al. 2017). Currently PFOS, PFOA and their related salts and precursors are regulated by the Stockholm Convention (Paul et al. 2009; Stockholm Convention 2019) as persistent organic pollutants, and recently, the European Food Safety Authority (EFSA) suggested lower tolerable weekly intake levels for PFOA and PFOS (EFSA 2018). In addition, PFOA, perfluorononanoic acid (PFNA), and, most recently, perfluorodecanoic acid (PFDA) and perfluorobutanesulfonic acid (PFBS) have been added to the European Chemicals Agency (ECHA) Candidate List of Substances of Very High Concern for authorisation (ECHA 2019).

The Danish EPA (Kjølholt et al. 2015) and the Nordic Council of Ministers (Posner et al. 2013), as well as the US Agency for Toxic Substances and Disease Registry (ATSDR 2018) have concluded that there are considerable knowledge gaps regarding PFASs other than PFOS and PFOA, and that there is an urgent need to acquire data on the physical and chemical properties as well as on the toxicity for a broader range of PFASs. The Organization for Economic Co-operation and Development (OECD) has recently published a compilation of 4730 PFASs for which the environmental and human health risks are mostly unknown (OECD 2018*a*, 2018*b*). To address the lack of data, the US EPA recently completed a study on a large range of PFASs with the aim to create a library of PFASs for high-throughput screening based on a chemical category approach (Patlewicz et al. 2019). Computational approaches are a means to rank and prioritise such large chemical inventories based on their mobility, persistence, and bioaccumulation potential (Brown and Wania 2008; Dürig et al. 2019; Pizzo et al. 2016), which has previously been done only for smaller datasets of PFASs (Arp et al. 2006; Ding and Peijnenburg 2013; Gomis et al. 2015; Wang et al. 2011).

The present study aims to increase our understanding of the chemical and structural variation of PFASs of the recently compiled OECD database. Curated structural data in combination with multivariate statistics and hierarchical clustering was used to guide the selection of structurally diverse training sets of PFASs. Chemicals were selected for possible future testing aiming for the largest structural variation possible while also taking into consideration commercial availability. Lastly, the performance of available models for estimating physicochemical properties and environmental fate characteristics was studied related to their applicability domain and predictive accuracy in regard to PFASs.

Experimental

OECD database and data curation

This study was based on the data inventory of PFASs published by the OECD/UNEP Global PFC Group (OECD 2018*a*). The comprehensive database includes chemicals that have perfluoroalkyl moieties with three or more carbons or a perfluoroalkylether moiety with two or more carbons. This data inventory contains a total of 4730 entries with CAS numbers, chemical names, and structural categorisation. The open-source software package *Konstanz Information Miner* (*KNIME* Version 3.6.0 and 4.0) (KNIME 2019*a*) was used for data curation and as a generic modelling framework.

The data curation was performed in four steps (Fig. 1). The first step was to omit entries labelled as mixtures or polymers because they were unsuitable for the computational approaches applied in this project. Among the 3809 chemicals remaining after the first step, 1208 had simplified molecular-input line-entry systems (SMILES) primarily provided by the Swedish Chemicals Agency (KEMI), while the rest were missing and therefore had to be acquired from other sources. The second step of the process was to generate and check the quality of the structural information. The chemical identifier resolver node in *KNIME* (National Cancer Institute 2019) was used to acquire SMILES based on compound name and CAS information, which resulted in 2252 and 64 additional structures, respectively. A randomly selected subset of 225 was inspected for quality assessment and an error rate of 0.5 % was determined.

Furthermore, 288 structures could not be acquired using the method above, and therefore had to be manually downloaded from SciFinder (CAS 2019) based on the CAS registry number. Among these, 35 (Table S1, Supplementary Material) did not have any structural information available on SciFinder, PubChem (National Center for Biotechnology Information 2019), or ChemSpider (ChemSpider 2018), and therefore were excluded from the dataset. Details on data correction and quality checking can be found in the Supplementary Material. In total, 3777 curated structures were converted to SDF MOL format using *OpenBabel* (KNIME 2019*b*).

The third step of the data processing involved structure standardisation using the *Indigo Toolkit* (EPAM Systems 2019) standardizer node to remove single atoms, charges, and smaller ions, to neutralise zwitterions, and to standardise cis/trans structural information. The *Indigo Aromatizer* was used to harmonise aromatic structures.

The cheminformatics software used was not able to process rare atoms (such as Eu, Dy, Yb, etc.); therefore, 31 structures from the PFASs database had to be omitted (Fourches et al. 2010) (Table S2, Supplementary Material). The remaining 3734 cleaned SDF structures contained many duplicates with a majority resulting from salts with the same PFAS ion but different counter-ions and some arising from stereoisomers. Because these PFASs will dissociate in the environment and form identical ions, the fourth and final step of the curation process was to identify and eliminate duplicate entities. To identify these, the SDFs were converted to InChIKeys, and then the GroupBy (KNIME 2019*c*) node in *KNIME* was used to merge all identical InChIKeys, which resulted in 3363 unique structures (Table S3, Supplementary Material).

Descriptor generation, principal component analysis, and clustering

A total of 64 chemical descriptors (Table S4, Supplementary Material) were generated in *MOE* (ver. 2015.1001) (Chemical Computing Group 2019) and were used as previously described (Rännar and Andersson 2010; Stenberg et al. 2009). These descriptors were selected owing to their interpretability, which made it relatively simple to discuss the chemistry in the multivariate statistical analysis. Descriptors for the log octanol-water partition coefficient (named logP(o/w) and SlogP in *MOE*), water solubility (logS) and molecular refractivity (name SMR and mr in *MOE*) (Table S4, Supplementary Material), were removed to avoid possible predictive errors arising from PFASs being out of the domain for these models. The 59 remaining descriptors were log-transformed ($\log(1+j)$) with the exception of *PEOE_PC*$^-$ (total negative partial charge), which was negative log-transformed ($\log(-j+1)$).

Principal component analysis (PCA) was used to analyse the chemical variation of PFASs and the hierarchical clustering in the selection of representative compounds (Rännar and Andersson 2010). PCA was performed using the PCA nodes in *KNIME*, and, before analysis, the log-transformed chemical descriptors were scaled by decimal scaling and then normalised using *z*-score normalisation.

The agglomerative hierarchical clustering was based on a distance matrix using the first five principal components (PCs) and Euclidean distances. Single-, complete-, and average linkage methods (Yim and Ramdeen 2015) of hierarchical clustering were used to test how evenly the data were split between the clusters set at 12 clusters. Complete linkage resulted in the most even spread and was therefore chosen for further analysis. We strived to identify an optimum number of clusters with a high variance ratio criterion (VRC) (Caliński and Harabasz 1974; Downs and Barnard 2002), a low number of clusters and a low root mean square error (RMSE) for even cluster distribution (Fig. S1, Supplementary Material). The final number of clusters and the distribution of PFASs within them are shown in Table 1.

Environmental fate properties

The *EPI Suite 4.1* (US EPA 2015) package was used for the prediction of log octanol-water partition coefficient (K_{OW}), water solubility (S_w), bioconcentration factor (BCF) and vapour pressure (V_p) (model name and version are shown in Table 2). Predictions could not be performed for two compounds (Fig. S2, Supplementary Material) owing to the very large, complex structures, and, therefore, the property calculations were only performed on 3361 structures. In addition, log *D* and acid dissociation constant (pK_a) values were predicted using the JChem extension in *ChemAxon* version 19.21.0 (ChemAxon 2019*a*; ChemAxon 2019*b*) for *Microsoft Office 2016* and pK_a was also estimated using *MOE* (Chemical Computing Group 2019). Furthermore, three BCF models (CAESAR, Meylan and KNN) available in *VEGA QSAR* ver. 1.1.5 (VEGA HUB 2019) were applied for BCF prediction to assess whether performance varied.

For the models where training set information was available, the range of molecular weights and the number of fluorine atoms or fluorine-containing fragments were considered as the basis for three different evaluations of the applicability domains. For the first evaluation, PFASs with molecular weights higher or lower than the training set were considered to be out of the domain. For fragment-based models, PFASs were defined as out of the domain if they had a greater number of any fluorine-containing fragments than the maximum in the training set, for which correction factors were applied in the model. The same approach was used for the number of fluorine atoms, i.e. PFASs with a greater number of fluorine atoms than members of the training set were considered to be out of the domain with an exception for the V_p model. For the V_p models, one extreme compound was found in the training set with 10 fluorine atoms more, as compared with the second most fluorinated compound. In this case, the second most fluorinated compound was used to set the upper limit. Information on the training set of the applied log *D* model was lacking, but personal communication with the developer confirmed that no fluorinated compounds were used in the model development (ChemAxon 2018).

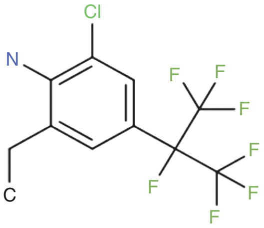
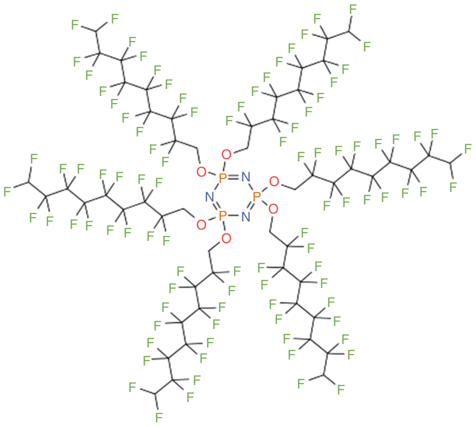
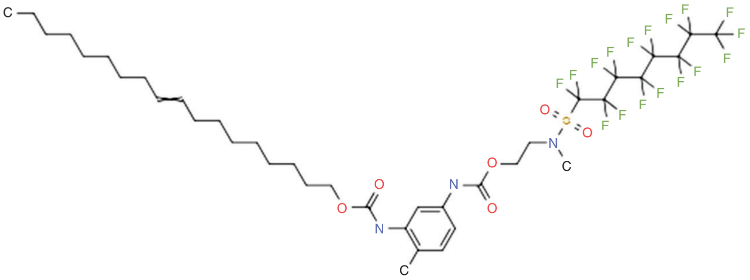
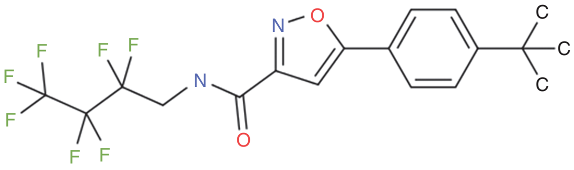
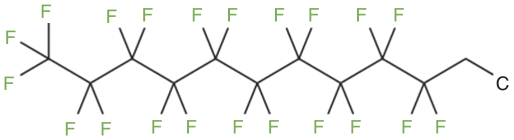
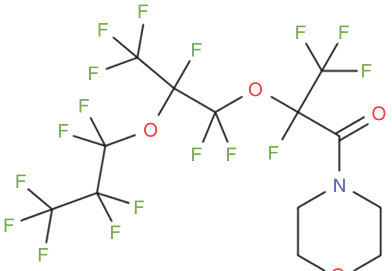
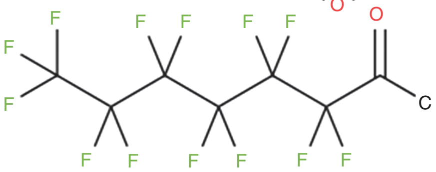
Results and discussion

Chemical variation of PFASs

The comprehensive OECD database of 4730 entries was curated and resulted in an inventory of 3363 unique PFAS structures. Around 19 % of the original OECD set ($n = 918$) was composed of mixtures (4 %, $n = 196$) or polymer mixtures (15 %, $n = 724$). The remaining PFASs (81 %) in the database had molecular weights ranging from 150 Da to 3217 Da (mean = 502 Da), and the number of fluorine atoms varied between 5 and 102. These PFASs contained a large number of molecular functionalities including acids, esters, ketones, aldehydes, linear and branched structures, aromatic ring structures, and in some cases, other halogens such as chlorine, bromine, and iodine.

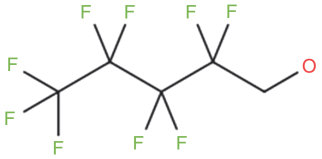
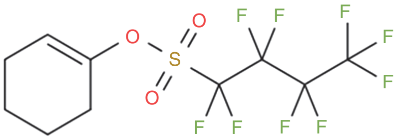
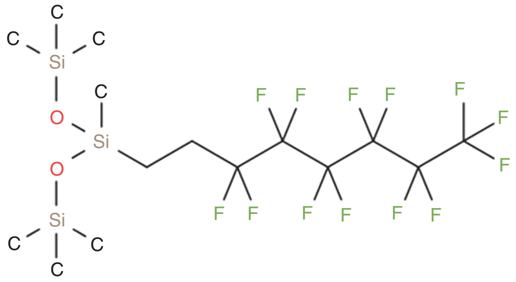
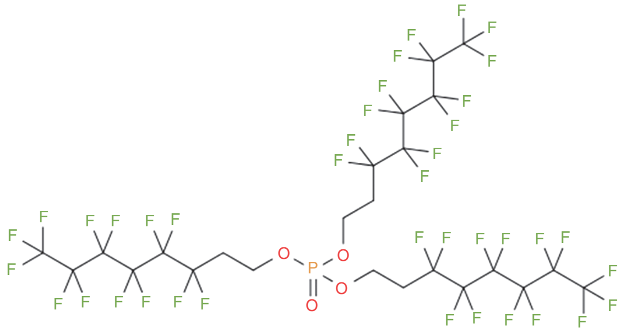
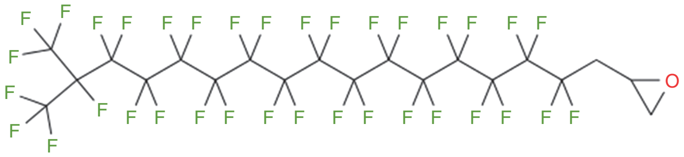
A detailed investigation of the chemical variation of the 3363 PFASs was undertaken using PCA and 59 chemical descriptors (Table S4, Supplementary Material) purely based

Table 1. Cluster summary including number of compounds (*N*), well-studied PFASs and one representative structure per cluster

Cluster	*N*	Well-studied PFAS[A]	Representative compound[B]
1	174		
2	7		
3	28		
4	180		
5	362	10 : 2 FTOH, 8 : 2 FTOH, PFTrDA, 12 : 2 FTOH, PFTeDA	
6	642	EtFOSA, PFOPA, PFDPA, PFDoA, EtFOSE, PFDS, C6/C6 PFPiA, PFDA, MeFOSE, PFUnA, MeFOSA, PFOS	
7	741	PFHxPA, HFPO-DA (GenX), 6 : 2 FTOH, PFHxA, PFHxS, PFOA, PFNA, PFHpA	

(*Continued*)

Table 1. (Continued)

Cluster	*N*	Well-studied PFAS[A]	Representative compound[B]
8	329	PFPeA, 4 : 2 FTOH, PFBS, PFBA	
9	429	MeFBSE, MeFBSA	
10	209		
11	131		
12	131	8 : 2 diPAP, C8/C8 PFPiA, 6 : 2 diPAP	

[A] PFASs that have been previously studied and have more than 10 citations according to Wang et al. (2017).

[B] Molecular structures of compounds located closest to the centre of each cluster based on Euclidian distance.

on non-empirical structural features (Fig. 2 for PC1 and 2; Fig. S3 (Supplementary Material) for PC3 and 4). Chemical structures of PFASs with extreme values for each PC are shown in Fig. S4 (Supplementary Material) and the descriptors with highest weights for each PC are shown in Table S5 (Supplementary Material). The first principal component (PC1) explained 45 % of the variance and had, in general, high weights for descriptors related to molecular size and surface area such as the Wiener path number (wienerPath) (Balaban 1979), the area of van der Waals surface (vdw_area) or the first kappa shape index (Kier 1) (Hall and Kier 1991). In this first dimension, small PFASs showed high PC1 values (e.g. CAS 697–11–0) and large PFASs showed low values (e.g. CAS 956790–67–3). The second PC (17 % of variance) was mostly related to relative density, number of fluorine atoms and aromaticity, which meant that poorly fluorinated, low density molecules with aromatic rings had high values (e.g. CAS 862133–14–0) while dense perfluorinated or other PFAS also containing other halogens (e.g. CAS 335–48–8) displayed low values. The third PC (13 % of variance) was related to polarity descriptors, where polar PFASs with acidic groups showed high values (e.g. CAS 109669–84–3) in contrast to more nonpolar and hydrophobic PFASs (e.g. CAS 190394–25–3). The fourth PC (8 % of variance) described a variance in the number of rotatable bonds and ring structures with high values for aromatic (e.g. CAS 956790–67–3) and low values for linear PFASs (e.g. CAS 400–57–7). The fifth PC (3 % of variance) was related to double bonds and hydrogen bond donors, with high values for PFAS with a large number of H-donors, e.g. sulfonic acid groups (e.g. CAS 375–73–5), and low values for PFASs with many C-C double bonds and a lack of ionisable groups (e.g. CAS 685–63–2).

Clustering and selecting training sets of PFASs

Clustering of the PFASs was performed based on the first five principal components aimed at finding groups of chemicals

Table 2. Applicability domain and accuracy for models predicting physicochemical and fate data

Parameter	Model	Domain (F)[A]		Domain (MW)[B]	R^2 [C]	Experimental data	
		F Fragments (%)	No. of F (%)	(%)		Total	Training[D]
log K_{OW}	*KOWWIN* ver. 1.68[E]	1.0	0.8	87	0.8	22	2
S_w (mg L^{-1})	*WSKOWWIN* ver. 1.42[E]	1.1 %	0.1	80	−0.4	14	3
pK_a	*JChem* for *Excel* ver. 18.27.0.403[F]	- [G]	-	-		6	0
log *D*	*JChem* for *Excel* ver. 18.27.0.403[F]	NA[H]	0.0	-	0.8	7	0
log K_{OC}	*KOCWIN* ver. 2.00[E]	0.0	0.0	84	0.8	17	0
BCF	*BCFBAF* ver. 3.01[E]	0.0	22	95	−2.1	11	0
V_p (mm Hg)	*MpbpWIN* ver. 1.43[E]	NA	78	94	0.9[I]	20	18

[A] Percentage of compounds (out of 3363) within the domain based on F count. Different domains were calculated depending on fragments or number of F atoms, as explained in the Methods section.

[B] Percentage of PFASs within the domain based on the molecular weight (MW) range of the training data for each model.

[C] Predicted versus experimental data (that was not used in model training).

[D] Number of experimental data, which was included in model development and therefore not used for comparison with predicted data.

[E] US EPA (2015).

[F] ChemAxon (2019*b*).

[G] Data not available.

[H] Not applicable.

[I] Calculated using all data including that used for model development.

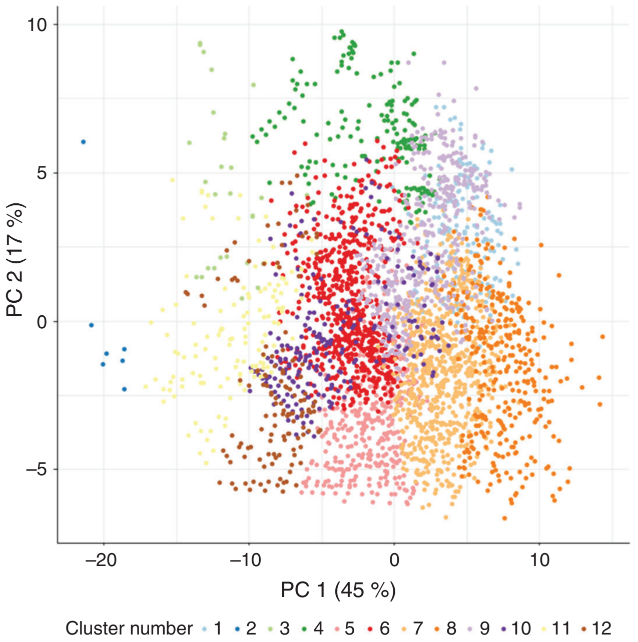


Fig. 2. Principal component analysis of the PFASs database including 3363 chemicals and 59 descriptors with the first two principal components (PC1 and PC2) indicating, in colour, the 12 clusters defined in the hierarchical cluster analysis.

sharing structural and chemical properties. The distribution was shown to be most even at 10 and above clusters, while the VRC, also known as the Calinski-Harabasz (CH) index (Caliński and Harabasz 1974; Downs and Barnard 2002), showed local maxima at 12 and 15 clusters (Fig. S1, Supplementary Material); therefore, 12 was selected as it was the lowest optimum number of clusters. Half of the clusters included well-studied PFASs (Table 1) (Wang et al. 2017), defined as those with more than 10 citations in the recent review by Wang et al. (2017). These have been detected in various environmental compartments such as air (Ahrens et al. 2011), surface water (Munoz et al. 2017; Pan et al. 2018), groundwater (Gobelius et al. 2018), and soil (Dalahmeh et al. 2018; Plassmann and Berger 2013). Notably the six clusters including the well-studied PFASs were large and covered 78 % ($n = 2634$) of the database, and these represented mainly small to medium-sized, linear, highly fluorinated, and non-polar or bipolar PFASs. This means that considerable knowledge gaps exist regarding environmental fate and effects and human health risks of, for example, aromatic, large, highly polar, and branched PFASs. The centroid chemical per cluster, calculated using Euclidean distances, is presented in Table 1 as a cluster representative.

To select a representative set of PFASs for future testing, all chemicals in each cluster were studied in cluster-specific PCA models (Rännar and Andersson 2010). The 5 % (alternatively, one compound if $n < 20$) of compounds found closest to the centre were selected as representatives of that particular chemical domain, which yielded, in total, 165 chemicals spread over 1–37 individual chemicals per cluster proportional to the cluster size (Table S6, Supplementary Material). This approach enabled us to represent the chemical space by cluster-typical chemicals and thus to avoid more unique chemicals that might be found at the cluster edge (Rännar and Andersson 2010). This large number of suggested chemicals offers several options for the design of structurally varied training sets considering possible constraints such as commercial availability, experimental design, ease of chemical analysis, etc. The number of compounds can be varied as long as each cluster is represented and most of the chemical space is covered, and here, we denote this as the theoretical training set.

A large share of the chemicals in the OECD data inventory is not likely to be procurable because these include, for example, patent records, and thus commercial availability should be addressed in a procurable training set as opposed to the theoretical set. Recently, an initiative to select PFASs for toxicity testing has been communicated by Patlewicz et al. (2019), which addresses issues such as availability and solubility for testing purposes. The methodology for selecting diversity was expert-based, in contrast to the cheminformatics-based approach discussed in this study, and the selection was based on a different PFAS inventory. The test set suggested by Patlewicz et al.

(2019) was inspected in terms of cluster distribution and showed representation in 5 of the 12 clusters, similar to the well-studied PFASs. Furthermore, we inspected the Norman suspect screening list for PFASs (Trier and Lunderberg 2015) and found matches in 10 of the clusters (3–12) with a similar distribution of compounds as for the whole OECD database. To select a procurable training set with a larger coverage of the chemical domain of PFASs, information provided by Patlewicz et al. (2019) and an inventory provided by the Swedish Chemicals Agency, KEMI (KEMI 2019) was used (details in the SI) and 1–3 PFASs were selected from each cluster. In the cases where more than three compounds were available for one cluster, those found closest to the centre of the cluster were selected. Clusters 2 and 3 were not represented in the above-mentioned inventories, but procurable compounds were found and included in the test set. These two clusters, however, only contained very large structures that were unlikely to be water soluble, thus would not be suitable for many laboratory tests. Nonetheless, the procurable set contained 23 PFASs spanning over most of the chemical space (Table S7, Supplementary Material) because all 12 clusters were represented. Including the chemicals of the procurable or the theoretical training set in future screening programs on critical environmental and human health endpoints would increase our understanding of an important group of chemicals in relation to their structural and chemical variation and would form a basis for the development of new predictive models including fate and effect models.

Physicochemical data and fate properties of PFASs

Several commonly used computational models for predicting physicochemical properties and fate characteristics were studied to determine their applicability and accuracy for PFAS. Notably, molecular size as a determinant of the applicability domain of the models showed that the majority of studied PFASs were within the domain (53–95 %) (Table 2). Using molecular fragments with fluorine atoms or the number of fluorine atoms yielded a much stricter assessment of domain inclusion. It should also be stressed that much more rigorous estimations of the applicability domain are typically used in modelling (Tropsha 2010) and most of these models were developed for neutral organic compounds, thus unlikely to be suitable for ionisable molecules.

Using the fluorine-based applicability domain criteria, none of the PFASs were considered within the domain of the models predicting log K_{oc} and log D, whereas only 0.1–1.1 % and only 0.8–1.0 % of the PFASs were in the domain of the S_w and log K_{OW} models, respectively (Table 2). However, the V_p model had a much larger applicability domain that included 78 % of the PFASs. Training data for the pK_a model was not publicly available and therefore domain estimation was not possible. The BCF model does not incorporate any adjustment factors for fluorine-containing fragments; therefore, no compounds were considered in the domain for that assessment. However, some fluorinated chemicals were used to train that model yielding a representation of 22 % of chemicals within the database considering the number of fluorine atoms (same or lower).

The studied software used for predicting K_{OW} have only a few PFASs in their training sets, which likely yields uncertain and inaccurate results (Arp et al. 2006). A literature survey on the available experimental data on K_{OW} resulted in data for, in total, 18 PFASs (Arp et al. 2006; Carmosini and Lee 2008; de Voogt et al. 2012; Xiang et al. 2018). Several PFASs can possess both hydrophobic and hydrophilic character (Rayne and Forest 2009), which makes the K_{OW} difficult to measure experimentally (Xiang et al. 2018). Another issue with K_{OW} of PFAS is that they are ionisable, an issue that is further discussed below. The literature investigation showed low variability of log K_{OW} measurements between different studies of the same PFAS (Fig. 3). Furthermore, data seem to be linearly correlated with alkyl chain-length (Fig. S5, Supplementary Material). The experimental data correlated well with the estimated data from SLogP ($R^2 = 0.81$) and KOWWIN ($R^2 = 0.77$), while logP$_{(o/w)}$ highly overestimated K_{OW} ($R^2 = -1.56$) (Fig. 3 and Fig. S6, Supplementary Material). However, the experimental data cover PFASs only from four clusters and, to approximate the chemical representativity of these chemicals in more detail, a K-nearest neighbours analysis was performed. Euclidean distances were calculated and the five nearest neighbours identified (square root of data points rounded to the nearest integer (Jonsson and Wohlin 2004)) for each experimental data set, which resulted in a list of 65 PFASs after exclusion of the 18 initial data points and overlapping neighbours. This meant that 2 % of the PFASs in the entire dataset could assumingly be predicted with a high accuracy considering that the log K_{OW} predictions of the experimental data fitted well with both the SLogP and KOWWIN estimates. Estimated log K_{OW} by KOWWIN of these 65 PFASs ranged from 1.2 to 11 with a mean of 5.1, while the SlogP prediction ranged from 1.3 to 9.3 with a mean of 4.6 (Table S8, Supplementary Material).

Another critical physicochemical property that can be used to assess the mobility of emerging contaminants is water solubility. However, our inventory only revealed experimental data on 14 chemicals (Inoue et al. 2012; QSAR Toolbox Coordination Group 2019; US EPA 2015), and the correlation with predictions was poor (Table 2) and thus not reliable. Mobility and solubility are heavily dependent on the ionizability of studied chemicals, and a large share of PFASs are acids; therefore, reliable pK_a values are critical. Our inventory revealed six experimental pK_a values (Burns et al. 2008; López-Fontán et al. 2005; Moroi et al. 2001; QSAR Toolbox Coordination Group 2019) and, in addition, limits of acidity have been reported for three PFASs (i.e. no exact values) (Vierke et al. 2013). The performance of the applied prediction tools (MOE and JChem) was analysed using the six reported values (for PFBA, PFHxA, PFOA, PFDA, PFUnDA and EtFOSA) (the limit values were excluded in the analysis). Four were predicted by JChem below the reliability limit set at –1 (defined by JChem) and could therefore not be used for the estimation of model performance. *MOE*, however, predicted the pK_a of EtFOSA to 9 (reported 9.5) but assigned a value of around 1 for the remaining five PFASs. The variation in experimental data between different studies was very high (Fig. S7, Supplementary Material) with, for example, pK_a values of PFOA being reported between 0.5 and 3.8 (Burns et al. 2008; Vierke et al. 2013). Overall, the reliability of the pK_a models could not be assessed mainly owing to lacking and unreliable experimental data. Nevertheless, the calculated pK_a values indicated that 37 % of the PFASs (among the 1125 which were identified as ionisable by the model) had a pK_a below 6, which suggested that they might be in ionic form in both natural water bodies and human blood. This adds to the uncertainty of several physicochemical property predictions because most models are only valid for neutral species. The pK_a values of the studied chemicals were also reflected in the estimated log D values that were, on average, lower than the predicted log K_{OW} (5.8 and 6.5 respectively). For log D, only seven values (Rayne and Forest

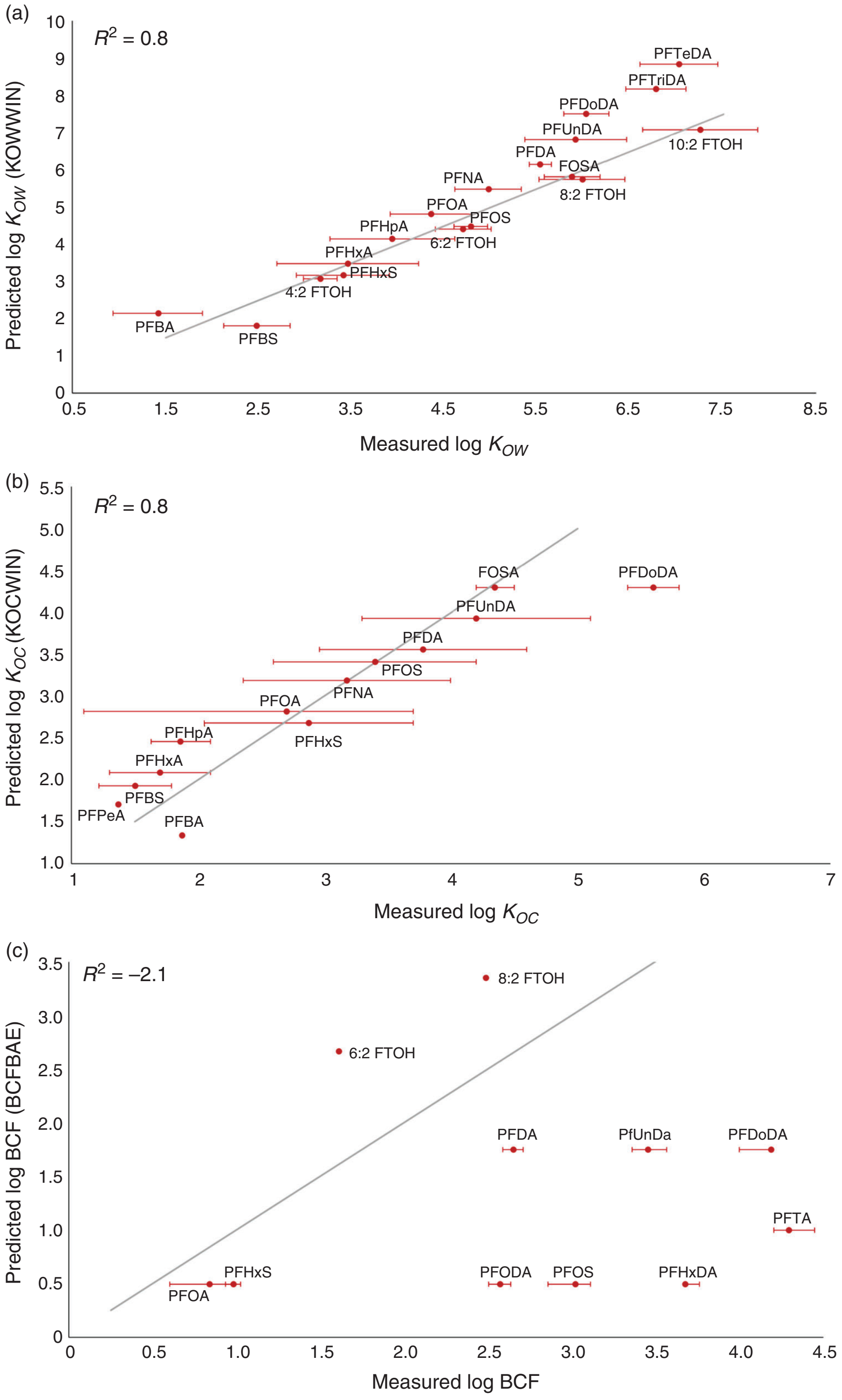


Fig. 3. (a) Average measured log K_{OW} for 18 PFASs (Arp et al. 2006; Carmosini and Lee 2008; de Voogt et al. 2012; Xiang et al. 2018) and their predicted log K_{OW} (*KOWWIN*). (b) Average measured log K_{OC} for 17 PFASs (Campos Pereira et al. 2018; Liu and Lee 2007). (c) Average measured BCF for 11 PFASs in fish (Inoue et al. 2012; Martin et al. 2003; QSAR Toolbox Coordination Group 2019) compared with predictions of BCFBAF. Error bars represent the minimum and maximum values from at least one study using one or several methodologies. Line represents a 1 : 1 correlation.

2009) were found and they correlated well with the experimental data (Fig. S8, Supplementary Material) but represent a very limited subset.

A data search for K_{OC} was performed, owing to it being a critical parameter to assess mobility and environmental fate of PFASs. A total of 17 experimental values (Campos Pereira et al. 2018; Liu and Lee 2007) were identified, which correlated well with the predicted data ($R^2 = 0.76$; Fig. 2), but only represented three of the twelve clusters. BCF predictions based on log K_{OW} could be overestimated considering the ionisation issue. Few

BCF values were found in the scientific literature (Inoue et al. 2012; Martin et al. 2003; QSAR Toolbox Coordination Group 2019) and the databases reviewed, and the four different BCF models applied generated data that was poorly correlated with experimental observations ($R^2 = -3.4$ to -0.5; Fig. 3 and Fig. S9, Supplementary Material). Some PFASs have been shown to bind strongly to albumin and other proteins (Jones et al. 2003) and this could cause this discrepancy between log K_{OW} and BCF despite a recent review indicating only a minor impact of protein binding for PFAS and other surfactants (Schlechtriem et al. 2014).

A rather large set of experimental data was compiled for V_p (Lei et al. 2004; QSAR Toolbox Coordination Group 2019; US EPA 2015), which, however, were almost exclusively used in the V_p model in *EPI Suite* and could therefore not be applied as a true external test set of the model. Thus, despite a high correlation with experimental data ($R^2 = 0.93$; Fig. S10, Supplementary Material), the use of the estimated V_p values cannot be recommended for PFAS. A search for data for air-water partitioning (K_{AW}) was also performed but only a small amount of data was identified (Lei et al. 2004; Rayne and Forest 2009).

The impressive inventory of PFASs by the OECD includes a huge variation in chemistry spanning from polymeric PFAS and mixtures to discrete small organofluorine chemicals. The curation of the data and the multivariate statistical analysis can hopefully serve as a starting point for further in-depth studies on these chemicals. We have clearly illustrated the gaps in the physicochemical properties and environmental fate data and also the imbalance in structure-related knowledge. The majority of studies are only performed on a handful of chemicals. The current study also highlights the need for improving available *in silico* fate and property models, which both warrant tailored models for these types of chemicals and new data for training said models. Sound models and accurate physicochemical properties are critical in understanding the environmental fate characteristics of PFASs and their potential hazards as a group, but primarily to enable high-throughput screening for the identification and prioritisation of the potentially most problematic PFASs.

Supplementary material

Extra information on descriptors, training sets, predicted data and various model performances are available on the Journal's website.

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgements

This research was financially supported by the Swedish Research Council Formas, grant no. 2017–00675, and the Swedish Research Council, grant no. 2017–01036.

References

Agency for Toxic Substances and Disease Registry (ATSDR) (2018). ATSDR – Toxicological Profile: Perfluoroalkyls. Available at https://www.atsdr.cdc.gov/toxprofiles/tp.asp?id=1117&tid=237 [verified 19 March 2019]

Ahrens L, Bundschuh M (2014). Fate and effects of poly- and perfluoroalkyl substances in the aquatic environment: A review. *Environmental Toxicology and Chemistry* **33**, 1921–1929. doi:10.1002/ETC.2663

Ahrens L, Shoeib M, Harner T, Lee SC, Guo R, Reiner EJ (2011). Wastewater Treatment Plant and Landfills as Sources of Polyfluoroalkyl Compounds to the Atmosphere. *Environmental Science & Technology* **45**, 8098–8105. doi:10.1021/ES1036173

Arp HPH, Niederer C, Goss K-U (2006). Predicting the partitioning behavior of various highly fluorinated compounds. *Environmental Science & Technology* **40**, 7298–7304. doi:10.1021/ES060744Y

Balaban AT (1979). Chemical graphs. *Theoretica Chimica Acta* **53**, 355–375. doi:10.1007/BF00555695

Blaine AC, Rich CD, Sedlacko EM, Hundal LS, Kumar K, Lau C, Mills MA, Harris KM, Higgins CP (2014). Perfluoroalkyl Acid Distribution in Various Plant Compartments of Edible Crops Grown in Biosolids-Amended soils. *Environmental Science & Technology* **48**, 7858–7865. doi:10.1021/ES500016S

Brown TN, Wania F (2008). Screening Chemicals for the Potential to be Persistent Organic Pollutants: A Case Study of Arctic Contaminants. *Environmental Science & Technology* **42**, 5202–5209. doi:10.1021/ES8004514

Buck RC, Franklin J, Berger U, Conder JM, Cousins IT, de Voogt P, Jensen AA, Kannan K, Mabury SA, van Leeuwen SP (2011). Perfluoroalkyl and polyfluoroalkyl substances in the environment: Terminology, classification, and origins. *Integrated Environmental Assessment and Management* **7**, 513–541. doi:10.1002/IEAM.258

Burns DC, Ellis DA, Li H, McMurdo CJ, Webster E (2008). Experimental pK_a Determination for Perfluorooctanoic Acid (PFOA) and the Potential Impact of pK_a Concentration Dependence on Laboratory-Measured Partitioning Phenomena and Environmental Modeling. *Environmental Science & Technology* **42**, 9283–9288. doi:10.1021/ES802047V

Caliński T, Harabasz J (1974). A dendrite method for cluster analysis. *Communications in Statistics* **3**, 1–27. doi:10.1080/03610927408827101

Campos Pereira H, Ullberg M, Kleja DB, Gustafsson JP, Ahrens L (2018). Sorption of perfluoroalkyl substances (PFASs) to an organic soil horizon – Effect of cation composition and pH. *Chemosphere* **207**, 183–191. doi:10.1016/J.CHEMOSPHERE.2018.05.012

Carmosini N, Lee LS (2008). Partitioning of fluorotelomer alcohols to octanol and different sources of dissolved organic carbon. *Environmental Science & Technology* **42**, 6559–6565. doi:10.1021/ES800263T

CAS (2019). SciFinder. Available at https://scifinder.cas.org/scifinder/login?TYPE=33554433&REALMOID=06-b7b15cf0-642b-1005-963a-830c809fff21&GUID=&SMAUTHREASON=0&METHOD=GET&SMAGENTNAME=-SM-8iKaCGmTCGHP7yOOI24GDUsJNzy%2bOcz79s1IdZR3o%2fpMdGZxHUbYH371HFTEMp2Z&TARGET=-SM-http%3a%2f%2fscifinder%2ecas%2eorg%3a443%2fscifinder%2f [verified 19 March 2019].

ChemAxon (2018). Applicability Domain assessment-pka, logD: Support Portal. Available at https://chemaxon.freshdesk.com/support/solutions/articles/43000569185-applicability-domain-assessment-pka-logd [verified 2 April 2020].

ChemAxon (2019*a*). Software solutions and services for chemistry & biology. Available at https://chemaxon.com/ [verified 14 January 2019].

ChemAxon (2019*b*). Jchem for Office. Available at https://chemaxon.com/products/jchem-for-office [verified 29 October 2019].

Chemical Computing Group (2019). Molecular Operating Environment (MOE). Available at https://www.chemcomp.com/Products.htm [verified 19 March 2019].

ChemSpider 2018. ChemSpider. Available at http://www.chemspider.com/ [verified 19 March 2019].

Dalahmeh S, Tirgani S, Komakech AJ, Niwagaba CB, Ahrens L (2018). Per- and polyfluoroalkyl substances (PFASs) in water, soil and plants in wetlands and agricultural areas in Kampala, Uganda. *The Science of the Total Environment* **631–632**, 660–667. doi:10.1016/J.SCITOTENV.2018.03.024

de Voogt P, Zurano L, Serne P, Haftka J (2012). Experimental hydrophobicity parameters of perfluorinated alkylated substances from reversed-phase high-performance liquid chromatography. *Environmental Chemistry* **9**, 564–570. doi:10.1071/EN12132

Ding G, Peijnenburg WJGM (2013). Physicochemical Properties and Aquatic Toxicity of Poly- and Perfluorinated Compounds. *Critical Reviews in Environmental Science and Technology* **43**, 598–678. doi:10.1080/10643389.2011.627016

Downs GM, Barnard JM (2002). Clustering methods and their uses in computational chemistry. *Reviews in Computational Chemistry* **18**, 1–40. doi:10.1002/0471433519.CH1

Dürig W, Tröger R, Andersson PL, Rybacka A, Fischer S, Wiberg K, Ahrens L (2019). Development of a suspect screening prioritization tool for organic compounds in water and biota. *Chemosphere* **222**, 904–912. doi:10.1016/J.CHEMOSPHERE.2019.02.021

Ellis DA, Martin JW, De Silva AO, Mabury SA, Hurley MD, Sulbaek Andersen MP, Wallington TJ (2004). Degradation of Fluorotelomer Alcohols: A Likely Atmospheric Source of Perfluorinated Carboxylic Acids. *Environmental Science & Technology* **38**, 3316–3321. doi:10.1021/ES049860W

EPAM Systems (2019). Indigo Toolkit. Available at http://lifescience.opensource.epam.com/indigo/ [verified 14 January 2019].

European Chemicals Agency (ECHA) (2019). Candidate list of substances of very high concern for authorisation. Available at https://echa.europa.eu/candidate-list-table [verified 28 October 2019].

European Food Safety Authority (EFSA) (2018). Contaminants update: first of two opinions on PFAS in food. Available at https://www.efsa.europa.eu/en/press/news/181213 [verified 25 April 2019].

Fourches D, Muratov E, Tropsha A (2010). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling* **50**, 1189–1204. doi:10.1021/CI100176X

Gewurtz SB, Backus SM, De Silva AO, Ahrens L, Armellin A, Evans M, Fraser S, Gledhill M, Guerra P, Harner T, Helm PA, Hung H, Khera N, Kim MG, King M, Lee SC, Letcher RJ, Martin P, Marvin C, McGoldrick DJ, Myers AL, Pelletier M, Pomeroy J, Reiner EJ, Rondeau M, Sauve M-C, Sekela M, Shoeib M, Smith DW, Smyth SA, Struger J, Spry D, Syrgiannis J, Waltho J (2013). Perfluoroalkyl acids in the Canadian environment: Multi-media assessment of current status and trends. *Environment International* **59**, 183–200. doi:10.1016/J.ENVINT.2013.05.008

Giesy JP, Naile JE, Khim JS, Jones PD, Newsted JL (2010). Aquatic toxicology of perfluorinated chemicals. *Reviews of Environmental Contamination and Toxicology* **202**, 1–52. doi:10.1007/978-1-4419-1157-5_1

Gobelius L, Lewis J, Ahrens L (2017). Plant Uptake of Per- and Polyfluoroalkyl Substances at a Contaminated Fire Training Facility to Evaluate the Phytoremediation Potential of Various Plant Species. *Environmental Science & Technology* **51**, 12602–12610. doi:10.1021/ACS.EST.7B02926

Gobelius L, Hedlund J, Dürig W, Tröger R, Lilja K, Wiberg K, Ahrens L (2018). Per- and Polyfluoroalkyl Substances in Swedish Groundwater and Surface Water: Implications for Environmental Quality Standards and Drinking Water Guidelines. *Environmental Science & Technology* **52**, 4340–4349. doi:10.1021/ACS.EST.7B05718

Gomis MI, Wang Z, Scheringer M, Cousins IT (2015). A modeling assessment of the physicochemical properties and environmental fate of emerging and novel per- and polyfluoroalkyl substances. *The Science of the Total Environment* **505**, 981–991. doi:10.1016/J.SCITOTENV.2014.10.062

Hall LH, Kier LB (1991). The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. In ‘Reviews in computational chemistry’. (Eds KB Lipkowitz, DB Boyd) pp. 367–422. (John Wiley & Sons, Ltd: Hoboken, NJ)

Higgins CP, Luthy RG (2006). Sorption of Perfluorinated Surfactants on Sediments. *Environmental Science & Technology* **40**, 7251–7256. doi:10.1021/ES061000N

Inoue Y, Hashizume N, Yakata N, Murakami H, Suzuki Y, Kikushima E, Otsuka M (2012). Unique Physicochemical Properties of Perfluorinated Compounds and Their Bioconcentration in Common Carp *Cyprinus carpio* L. *Archives of Environmental Contamination and Toxicology* **62**, 672–680. doi:10.1007/S00244-011-9730-7

Jahnke A, Ahrens L, Ebinghaus R, Temme C (2007). Urban versus Remote Air Concentrations of Fluorotelomer Alcohols and Other Polyfluorinated Alkyl Substances in Germany. *Environmental Science & Technology* **41**, 745–752. doi:10.1021/ES0619861

Jones PD, Hu W, Coen WD, Newsted JL, Giesy JP (2003). Binding of perfluorinated fatty acids to serum proteins. *Environmental Toxicology and Chemistry* **22**, 2639–2649. doi:10.1897/02-553

Jonsson P, Wohlin C (2004). An evaluation of k-nearest neighbour imputation using Likert data. In ‘10th International symposium on software metrics, 2004. Proceedings’. pp. 108–118. (IEEE: Chicago, IL)

KEMI (2019). Swedish Chemicals Agency. Available at https://www.kemi.se/ [verified 15 January 2019].

Kjølholt J, Jensen AA, Warming M (2015). Short-chain polyfluoroalkyl substances. A literature review of information on human health effects and environmental fate and effect aspects of short-chain PFAS (No. Environmental project No. 1707, 2015). The Danish Environmental Protection Agency.

KNIME (2019*a*). End to end data science. Available at https://www.knime.com/ [verified 14 January 2019].

KNIME (2019*b*). OpenBabel. https://nodepit.com/node/org.knime.ext.chem.openbabel.BabelFactory [verified 19 March 2019].

KNIME (2019*c*). GroupBy. Available at https://nodepit.com/node/org.knime.base.node.preproc.groupby.GroupByNodeFactory [verified 19 March 2019].

Lee H, D’eon J, Mabury SA (2010). Biodegradation of Polyfluoroalkyl Phosphates as a Source of Perfluorinated Acids to the Environment. *Environmental Science & Technology* **44**, 3305–3310. doi:10.1021/ES9028183

Lei YD, Wania F, Mathers D, Mabury SA (2004). Determination of Vapor Pressures, Octanol−Air, and Water−Air Partition Coefficients for Polyfluorinated Sulfonamide, Sulfonamidoethanols, and Telomer Alcohols. *Journal of Chemical & Engineering Data* **49**, 1013–1022. doi:10.1021/JE049949H

Liu J, Lee LS (2007). Effect of Fluorotelomer Alcohol Chain Length on Aqueous Solubility and Sorption by Soils. *Environmental Science & Technology* **41**, 5357–5362. doi:10.1021/ES070228N

Liu J, Mejia Avendaño S (2013). Microbial degradation of polyfluoroalkyl chemicals in the environment: A review. *Environment International* **61**, 98–114. doi:10.1016/J.ENVINT.2013.08.022

López-Fontán JL, Sarmiento F, Schulz PC (2005). The aggregation of sodium perfluorooctanoate in water. *Colloid & Polymer Science* **283**, 862–871. doi:10.1007/S00396-004-1228-7

Martin JW, Mabury SA, Solomon KR, Muir DCG (2003). Bioconcentration and tissue distribution of perfluorinated acids in rainbow trout (*Oncorhynchus mykiss*). *Environmental Toxicology and Chemistry* **22**, 196–204. doi:10.1002/ETC.5620220126

Mejia-Avendaño S, Munoz G, Vo Duy S, Desrosiers M, Benoît P, Sauvé S, Liu J (2017). Novel Fluoroalkylated Surfactants in Soils Following Firefighting Foam Deployment During the Lac-Mégantic Railway Accident. *Environmental Science & Technology* **51**, 8313–8323. doi:10.1021/ACS.EST.7B02028

Moroi Y, Yano H, Shibata O, Yonemitsu T (2001). Determination of Acidity Constants of Perfluoroalkanoic Acids. *Bulletin of the Chemical Society of Japan* **74**, 667–672. doi:10.1246/BCSJ.74.667

Munoz G, Labadie P, Botta F, Lestremau F, Lopez B, Geneste E, Pardon P, Dévier M-H, Budzinski H (2017). Occurrence survey and spatial distribution of perfluoroalkyl and polyfluoroalkyl surfactants in groundwater, surface water, and sediments from tropical environments. *The Science of the Total Environment* **607–608**, 243–252. doi:10.1016/J.SCITOTENV.2017.06.146

National Cancer Institute (2019). NCI/CADD Chemical Identifier Resolver. Available at https://cactus.nci.nih.gov/chemical/structure [verified 14 January 2019].

National Center for Biotechnology Information (2019). PubChem. Available at https://pubchem.ncbi.nlm.nih.gov/ [verified 26 April 2019].

OECD (2018*a*). Towards a new comprehensive global database of per- and polyfluoroalkyl substances (PFASs). Organisation for Economic Co-operation and Development (OECD).

OECD (2018*b*). Towards a new comprehensive global database of per- and polyfluoroalkyl substances (PFASs): summary report on updating the OECD 2007 list of per- and polyfluoroalkyl substances (PFASs) (No. JT03431231), series on risk management no. 39. Organisation for Economic Co-operation and Development (OECD).

Pan Y, Zhang H, Cui Q, Sheng N, Yeung LWY, Sun Y, Guo Y, Dai J (2018). Worldwide Distribution of Novel Perfluoroether Carboxylic and Sulfonic Acids in Surface Water. *Environmental Science & Technology* **52**, 7621–7629. doi:10.1021/ACS.EST.8B00829

Patlewicz G, Richard AM, Williams AJ, Grulke CM, Sams R, Lambert J, Noyes PD, DeVito MJ, Hines RN, Strynar M, Guiseppi-Elie A, Thomas RS (2019). A Chemical Category-Based Prioritization Approach for Selecting 75 Per- and Polyfluoroalkyl Substances (PFAS) for Tiered Toxicity and Toxicokinetic Testing. *Environmental Health Perspectives* **127**, 014501. doi:10.1289/EHP4555

Paul AG, Jones KC, Sweetman AJ (2009). A First Global Production, Emission, And Environmental Inventory For Perfluorooctane Sulfonate. *Environmental Science & Technology* **43**, 386–392. doi:10.1021/ES802216N

Pizzo F, Lombardo A, Brandt M, Manganaro A, Benfenati E (2016). A new integrated in silico strategy for the assessment and prioritization of persistence of chemicals under REACH. *Environment International* **88**, 250–260. doi:10.1016/J.ENVINT.2015.12.019

Plassmann MM, Berger U (2013). Perfluoroalkyl carboxylic acids with up to 22 carbon atoms in snow and soil samples from a ski area. *Chemosphere* **91**, 832–837. doi:10.1016/J.CHEMOSPHERE.2013.01.066

Posner S, Roos S, Brunn Poulsen P, Jörundsdottir HO, Gunnlaugsdottir H, Trier X, Astrup Jensen A, Katsogiannis AA, Herzke D, Bonefeld-Jörgensen EC, Jönsson C, Pedersen GA, Ghisari M, Jensen S (2013). Per- and polyfluorinated substances in the Nordic Countries: Use, occurence and toxicology (No. TemaNord 2013: 542). Nordic Council of Ministers.

QSAR Toolbox Coordination Group (2019). QSAR Toolbox. Available at https://qsartoolbox.org/ [verified 11 October 2019].

Rännar S, Andersson PL (2010). A Novel Approach Using Hierarchical Clustering To Select Industrial Chemicals for Environmental Impact Assessment. *Journal of Chemical Information and Modeling* **50**, 30–36. doi:10.1021/CI9003255

Rayne S, Forest K (2009). Perfluoroalkyl sulfonic and carboxylic acids: A critical review of physicochemical properties, levels and patterns in waters and wastewaters, and treatment methods. *Journal of Environmental Science and Health. Part A, Toxic/Hazardous Substances & Environmental Engineering* **44**, 1145–1199. doi:10.1080/10934520903139811

Schlechtriem DC, Nendza DM, Hahn DS, Zwintscher A, Schüürmann DG, Kühne DR (2014). Contribution of non-lipid based processes to the bioaccumulation of chemicals. Report No. UBA-FB 00. 92 pp.

Stenberg M, Linusson A, Tysklind M, Andersson PL (2009). A multivariate chemical map of industrial chemicals – Assessment of various protocols for identification of chemicals of potential concern. *Chemosphere* **76**, 878–884. doi:10.1016/J.CHEMOSPHERE.2009.05.011

Stockholm Convention (2019). Stockholm Convention. Available at http://www.pops.int/Home/tabid/2121/Default.aspx [verified 30 January 2020].

Trier X, Lunderberg D (2015). S9 | PFASTRIER | PFAS Suspect List: fluorinated substances. doi:10.5281/zenodo.3542121

Tropsha A (2010). Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **29**, 476–488. doi:10.1002/MINF.201000061

US EPA (2015). EPI Suite^{TM} - Estimation Program Interface v4.11. Available at https://www.epa.gov/tsca-screening-tools/download-epi-suitetm-estimation-program-interface-v411 [verified 14 January 2019].

VEGA HUB (2019). VEGA HUB. Available at https://www.vegahub.eu/ [verified 19 March 2019].

Vierke L, Berger U, Cousins IT (2013). Estimation of the Acid Dissociation Constant of Perfluoroalkyl Carboxylic Acids through an Experimental Investigation of their Water-to-Air Transport. *Environmental Science & Technology* **47**, 11032–11039. doi:10.1021/ES402691Z

Wang Z, MacLeod M, Cousins IT, Scheringer M, Hungerbühler K (2011). Using COSMOtherm to predict physicochemical properties of poly- and perfluorinated alkyl substances (PFASs). *Environmental Chemistry* **8**, 389–398. doi:10.1071/EN10143

Wang Z, DeWitt JC, Higgins CP, Cousins IT (2017). A Never-Ending Story of Per- and Polyfluoroalkyl Substances (PFASs)? *Environmental Science & Technology* **51**, 2508–2518. doi:10.1021/ACS.EST.6B04806

Xiang Q, Shan G, Wu W, Jin H, Zhu L (2018). Measuring log K_{ow} coefficients of neutral species of perfluoroalkyl carboxylic acids using reversed-phase high-performance liquid chromatography. *Environmental Pollution* **242**, 1283–1290. doi:10.1016/J.ENVPOL.2018.08.009

Xiao F, Golovko SA, Golovko MY (2017). Identification of novel non-ionic, cationic, zwitterionic, and anionic polyfluoroalkyl substances using UPLC–TOF–MSE high-resolution parent ion search. *Analytica Chimica Acta* **988**, 41–49. doi:10.1016/J.ACA.2017.08.016

Yim O, Ramdeen KT (2015). Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data. *The Quantitative Methods for Psychology* **11**, 8–21. doi:10.20982/TQMP.11.1.P008

Handling Editor: Kurunthachalam Kannan