Environment International 88 (2016) 198–205

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Investigating the mechanisms of bioconcentration through QSAR classification trees

Francesca Grisoni [a,b,*], Viviana Consonni [a,b], Marco Vighi [a], Sara Villa [a], Roberto Todeschini [a,b]

[a] University of Milano-Bicocca, Dept. of Earth and Environmental Sciences, Milano, Italy
[b] Milano Chemometrics and QSAR Research Group, Milano, Italy

ARTICLE INFO

Article history:
Received 31 October 2015
Received in revised form 18 December 2015
Accepted 19 December 2015
Available online 4 January 2016

Keywords:
Bioconcentration
CART
Mechanisms
QSAR
BCF
Molecular descriptors

ABSTRACT

This paper proposes a scheme to predict whether a compound (1) is mainly stored within lipid tissues, (2) has additional storage sites (e.g., proteins), or (3) is metabolized/eliminated with a reduced bioconcentration. The approach is based on two validated QSAR (Quantitative Structure–Activity Relationship) trees, whose salient features are: (a) descriptor interpretability and (b) simplicity. Trees were developed for 779 organic compounds, the TGD approach was used to quantify the lipid-driven bioconcentration, and a refined machine-learning optimization procedure was applied. We focused on molecular descriptor interpretation, which allowed us to gather new mechanistic insights into the bioconcentration mechanisms.

1. Introduction

One key aspect of environmental risk assessment is to quantify the extent to which contaminants, once emitted, will accumulate in biotic phases. Some xenobiotics can be stored within organisms at concentrations higher than measured in the environment and increase their concentration across the trophic chain, achieving harmful levels in fish, wildlife and humans (Chen et al., 1992; Cook et al., 2003; Gladen et al., 1988; Ratcliffe, 1967). At each trophic level, accumulation can occur through the skin or respiratory surfaces (e.g., lungs/gills), known as bioconcentration, or through the diet (dietary bioaccumulation) (Gobas and Morrison, 2000). This results in an increase in chemical concentration with increasing trophic level, exposing high-trophic level organisms, including humans, to adverse long-term effects difficult to predict, such as endocrine disruption (Geyer et al., 2000).

The growing concern over environmental and human-health risks connected with pollution has been reflected in massive regulatory efforts to identify and eliminate the most bioaccumulative chemicals. Regulatory authorities mainly rely on the Bioconcentration Factor (BCF) (Arnot and Gobas, 2006; Matthies et al., 2015), a proportionality constant comparing the concentration of a chemical within an organism (usually fish) to that in the water at steady-state, measured in a laboratory setting (Veith et al., 1979). BCF is the requested criterion for bioaccumulation assessment in many regulatory frameworks (Arnot and Gobas, 2006) but its determination is very expensive (more than 35,000 euros) and requires the use of more than 100 animals for each standard study (HESI, 2006), resulting in a general lack of data. This has necessitated the development of models to predict BCF: first using the *n*-octanol-water partition coefficient (K_{OW}) and, later, through more sophisticated methods, such as Quantitative Structure–Activity Relationships (QSARs). QSAR exploits statistical/mathematical techniques to quantitatively relate a biological property to molecular characteristics (e.g., structural features or physico-chemical properties), numerically encoded within the so-called molecular descriptors (Todeschini and Consonni, 2009). QSAR has been gaining increasing importance in international decision-making frameworks (Cronin et al., 2003) and is promoted by European REACH regulation (EC 1907/2006) for prioritization, data-gap filling and animal testing rationalization.

The majority of QSAR models for BCF are based on K_{OW} or related descriptors (Pavan et al., 2008). Bioconcentration in fact, mainly occurs as a thermodynamically driven partitioning between water and organism lipid phases, which is represented by K_{OW} (Mackay, 1982). However, other processes can significantly contribute to the observed

Abbreviations: QSAR, Quantitative Structure–Activity Relationship; BCF, bioconcentration factor; K_{OW}, octanol-water partition coefficient; CART, Classification and Regression Tree; TGD, Technical Guidance Document; GA, Genetic Algorithms.

[*] Corresponding author at: University of Milano-Bicocca, Dept. of Earth and Environmental Sciences, Milano, Italy.
E-mail address: f.grisoni@campus.unimib.it (F. Grisoni).

http://dx.doi.org/10.1016/j.envint.2015.12.024
0160-4120/© 2015 Elsevier Ltd. All rights reserved.

concentrations within organisms, such as metabolism, elimination, and specific interactions with tissues other than lipids (ECETOC, 1996). Chemicals that are metabolized into hydrophilic compounds can be eliminated faster and thus have lower BCF than predicted from K_{OW} (de Wolf et al., 1992; Muir et al., 1994). Nonetheless, chemicals that establish specific interactions with non-lipid tissues can have larger BCF than predicted, such as methylmercuric chloride, which has a low log K_{OW} but a very high BCF (up to 1,000,000 in fish) due to its association with protein sulfhydryl groups (Reinert et al., 1974). Correspondingly, K_{OW}-based QSAR models (of different degrees of complexity) can lead to underestimation or overestimation of the actual BCF (Grisoni et al., 2015).

To our knowledge, despite BCF having been widely modeled, the mechanisms that affect bioconcentration have never been investigated extensively in a QSAR setting and have never been integrated into a BCF assessment. This study stems from these considerations and aims to provide a scheme to: (1) assess the reliability of K_{OW}-based predictions of BCF, (2) understand which mechanisms can influence the fate of chemicals within organisms, and (3) gather new mechanistic insights into bioconcentration.

Wet weight BCF data and the TGD (Technical Guidance Document) model (European Commission, 2003) were used to assign 779 compounds to one of three mechanistic classes, viz.: (1) inert chemicals, which mainly accumulate within lipids, (2) specifically bioconcentrating chemicals, which have increased interactions with tissues, and (3) less bioconcentrating chemicals, which are metabolized or eliminated. These classes served as the model targets.

Our major goals were to: (1) use molecular description to investigate the structural features underlying the bioconcentration mechanisms and (2) obtain easily applicable and interpretable models to allow for widespread use and greater mechanistic understanding. To this end, we chose CART (Classification and Regression Trees) (Breiman et al., 1984) as a machine-learning algorithm. CART is based on a recursive partitioning of data using one variable at a time: at each univariate split (i.e. node), data are divided into two mutually exclusive groups (as homogeneous as possible) according to their variable values, and then the splitting procedure is further applied to each group separately. This procedure leads to a model that is graphically representable as a decision tree, where each node is a univariate split and leafs are the predicted classes for the objects that fall in that leaf. In addition to its simplicity and interpretability, the CART technique is able to deal with non-linear relationships between variables, thus being particularly well suited for complex biological problems.

Molecular description and CART classification were combined with growth-optimization, variable selection and model validation procedures, leading to the development of two QSAR classification trees, which offered new mechanistic insights and can be used for regulatory applications.

2. Materials and methods

2.1. Dataset

Our recently published dataset (wet weight BCF data on fish, Grisoni et al., 2015) was enriched with new experimental K_{OW} data, gathered from the work of Mansouri (2013) and manually curated (see Supplementary information). Only compounds with known experimental K_{OW} were retained, leading to a final dataset of 779 compounds.

2.2. Molecular descriptors

We calculated 0- to 2D molecular descriptors using Dragon 6 (Talete srl, 2012) and reduced the total pool (3763) by excluding those: (a) with at least one missing value; (b) constant or near-constant; (c) with a standard deviation less than 0.01; or (d) with a pairwise correlation larger or equal to 0.98 with other descriptors. In total, 1495 descriptors were retained as variables for model development.

2.3. Classification tree growth

CARTs were grown by:

1. Modeling a two-class problem (i.e., one class at a time against the others). This was done for classes 2 and 3 and gave considerably better results than the three-class approach.
2. Using an optimization approach, by varying (a) the misclassification cost (from 0 to 0.90 with a step of 0.10) and (b) the splitting criterion (Gini diversity index and cross-entropy) (Breiman et al., 1984).
3. Selecting the optimal tree complexity in cross-validation, by varying the minimum number of objects per leaf from 1 to 100 with a step of 10.
4. Implementing CART classification in a Genetic Algorithms setting (see 2.4).

2.4. Variable selection with Genetic Algorithms

CART automatically selects the optimal subset of variables in a stepwise manner. However, when a large number of variables are available, not all of their possible interactions are explored. For this reason, CART was integrated into Genetic Algorithms (GA) (Goldberg and Holland, 1988; Holland, 1992), a well-established technique for variable selection (e.g., Grisoni et al., 2014).

2.5. Software and code

CART calculations were performed with MATLAB (MATLAB, 2014), using the cartctree function as a core. Cross-validation, tree pruning, GA selection and model validation were performed using functions written by the Milano Chemometrics and QSAR Research Group.

3. Results and discussion

3.1. Class definition

The TGD model was used as a proxy for lipid-driven bioconcentration, as its results were the most accurate among the several equations tested (Grisoni et al., 2015). According to TGD, BCF is predicted as follows:

$$\begin{array}{ll} \log BCF_{TGD} = 0.15 & \text{if } \log K_{OW} < 1 \\ \log BCF_{TGD} = 0.85 \cdot \log K_{OW} - 0.70 & \text{if } 1 \le \log K_{OW} \le 6 \\ \log BCF_{TGD} = -0.20 \cdot (\log K_{OW})^2 + 2.74 \cdot \log K_{OW} - 4.72 & \text{if } 6 < \log K_{OW} < 10 \\ \log BCF_{TGD} = 2.68 & \text{if } \log K_{OW} \ge 10\,. \end{array} \tag{1}$$

Our hypothesis is that compounds reliably predicted from K_{OW} mainly accumulate within lipids, while others undergo additional processes. As a "region of reliability" for K_{OW}-based predictions, we used ±0.5 log units and hence defined the classes as:

1. Class 1 — Inert chemicals (460). Compounds whose $\log BCF_{exp}$ lies within a ± 0.5 log unit interval from $\log BCF_{TGD}$. These compounds mainly bioconcentrate within lipid tissues and thus partitioning-related models can be used.
2. Class 2 — Specifically bioconcentrating chemicals (64): if $\log BCF_{exp} \geq \log BCF_{TGD} + 0.5$. For these compounds, in addition to lipid storage, specific interactions with tissues can be hypothesized, which lead to an underestimation of BCF when K_{OW} or related parameters are used.
3. Class 3 — Less bioconcentrating chemicals (255 compounds): if $\log BCF_{exp} \leq \log BCF_{TGD} - 0.5$. The observed deficit of BCF could be

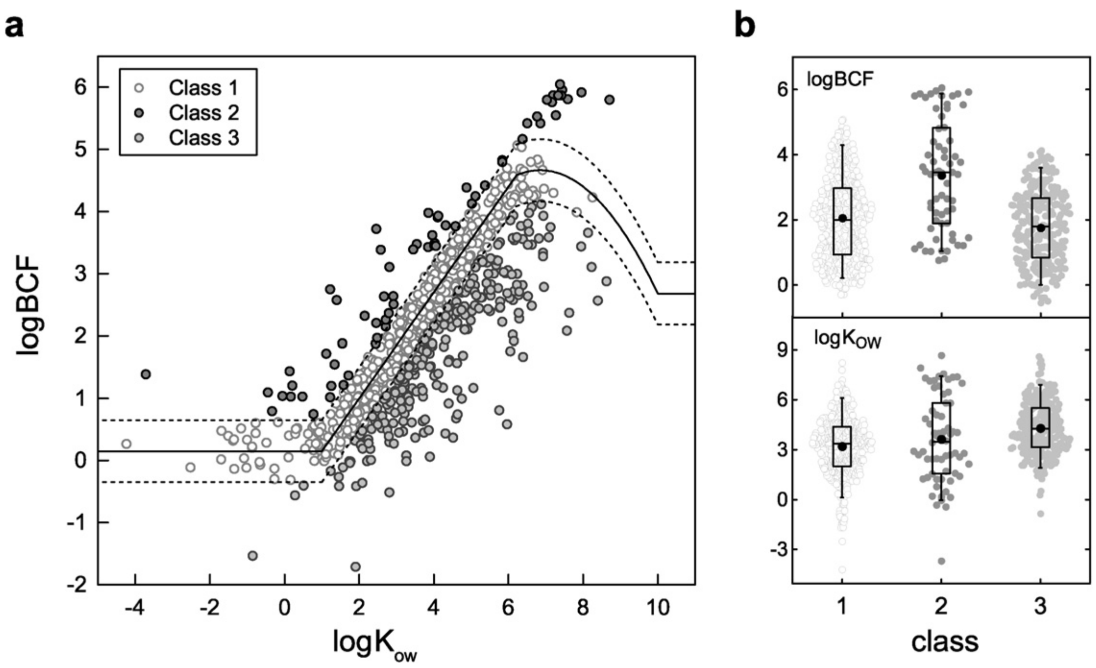


Fig. 1. Dataset compounds colored according to the assigned class: (a) K_{OW} vs BCF: the solid line represents the predicted BCF across the range of K_{OW} according to the TGD approach, while the dashed lines represent a ±0.5 log unit interval from BCF_{TGD}; and (b) distribution of K_{OW} and BCF data for the classes. Boxplots show median, 1st and 3rd quartiles (solid lines). The mean (black dots) and 5th–95th percentiles (whiskers) are also shown.

connected to biotransformation, which leads to faster elimination and/or to a bias in the measured BCF.

The threshold was calibrated on the basis of K_{OW} and BCF variability. We do not rule out the possibility that for some class 1 compounds, metabolism and/or interactions with tissues occur; however, deviations of less than 0.5 log units are not discernible from lipid-driven BCF. At the same time, some of the observed deviations could be due to model error or data uncertainty. Hence, where possible, we rationalized the obtained classifications through literature- and data-driven considerations.

The majority of compounds belong to class 1, with 33% and only 8% to classes 3 and 2, respectively (Fig. 1a). Class 2 chemicals lie mainly in the $\log K_{OW} > 0$ region and have a higher relative abundance of very bioaccumulative chemicals ($\log BCF > 3.5$, Fig. 1b), confirming that they are a class of concern.

The distribution within classes (Table 1) agrees with what is known about some environmental pollutants:

- PFAAs (perfluorinated alkyl acids) are hypothesized to have increased interactions with serum albumin, liver fatty acid proteins and phospholipids bilayers (Armitage et al., 2013; Jones et al., 2003; Lehmler and Bummer, 2004; Woodcroft et al., 2010; Xie et al., 2010). Accordingly, they were all assigned to class 2. Perfluorooctane Sulfonate showed the largest residual (2.35 log units) in the dataset.
- Some synthetic fragrances and organophosphorous compounds are known to be metabolized in fish (de Bruijn and Hermens, 1991; Rimkus, 1999) and they are correctly distributed in classes 1 and 3. Two organophosphates belong to class 2, as already highlighted in our previous study (Grisoni et al., 2015).
- Polycyclic Aromatic Hydrocarbons (PAHs) are effectively biotransformed into more hydrophilic compounds in fish liver and can be easily excreted (Lech and Vodicnik, 1985; Sijm and Opperhuizen, 1989). Furthermore, the influence of metabolism on their final BCF has already been reported (Jonsson et al., 2004); this agrees with their distribution in classes 1 (60%) and 3 (40%).
- Regarding Polychlorinated Biphenyls (PCBs), several studies reported a selective metabolic clearance in fish for a large number of congeners (Buckman et al., 2006; Melancon and Lech, 2013; White et al., 1997), with no observed elimination for hexa-, hepta- and octa-CBs (de Boer et al., 1993, 1994). These considerations agree to some extent with the data: 17 PCBs belong to class 2 and 14 of them are congeners with 6 to 8 Cl atoms. This could suggest a general excess of BCF for PCBs, particularly evident in those that are not metabolized/eliminated. Alternatively, biases could be ascribed to BCF-determination methodology, as already hypothesized by Wang et al. (2014)). Note that, for log $K_{OW} > 6$, only PCBs (15) belong to class 2 and, with the exception of two, they are all congeners with 6 to 8 Cl atoms.

Table 1
Class distribution of some environmentally relevant chemical classes of compounds. (PFAAs = perfluorinated alkyl acids, PCBs = polychlorinated biphenyls, PBBs = polybrominated biphenyls, PAHs = polycyclic aromatic hydrocarbons, IOCs = ionogenic organic compounds).

	Class			Total
	1	2	3	
Total number	460	64	255	779
PFAAs	0	7	0	7
PCBs	27	17	0	44
PBBs	2	0	0	2
Synthetic fragrances	1	0	2	3
Organophosphates	38	2	33	73
Synthetic pyrethroids	4	0	12	16
PAHs	9	0	6	15
IOCs*	182	26	110	318

* Selected according to the presence of ionizing functional groups (see Supplementary information).

Ionogenic Organic Compounds (IOCs) are a critical category of chemicals, whose uptake and elimination depend, in a complex way, on hydrophobicity, degree of ionization, electrostatic interactions and steric factors of both ionized and unionized forms. Moreover, some IOCs may interact with phospholipids and/or other macromolecules (Armitage et al., 2013). Despite a conceivable deviation from K_{OW}-based BCF, IOCs are about the 40% of the total, with almost equal relative distribution within the classes (Table 1). This means that no trend of overestimation/underestimation based on K_{OW} is directly associable with ionization. Therefore, we retained ionizing compounds for model development, to maximize the structural information available.

3.2. Modeling and results

Compounds were randomly split into a training set of 584 compounds (75%) and a test set of 195 compounds (25%), preserving the proportion between the classes. The training set was used for variable selection, tree pruning (both with 5-fold cross-validation), and model calibration. The test set only served to validate the final pool of models. The model's predictive ability was quantified using sensitivity (*Sn*), specificity (*Sp*) and non-error rate (*NER*), defined as follows:

$$\begin{aligned} Sn &= \frac{TP}{TP+FN} \\ Sp &= \frac{TN}{TN+FP} \\ NER &= \frac{\sum_{i=1}^{G} Sn_i}{G} \end{aligned} \tag{2}$$

where *TP*, *TN*, *FP* and *FN* are the number of true positives, true negatives, false positives and false negatives for each class, respectively; and *G* is the number of classes. Tree pruning and GA selection were performed in cross-validation.

Each critical class (i.e., 2 and 3) was modeled in turn against the remaining ones. In particular, for each optimization setting (Section 2.3), we carried out a three step procedure: (1) GA selection on the whole pool of descriptors; (2) GA selection on the most frequently retained descriptors from step 1 (maximum 150); and (3) generation of all of the possible combinations of the most relevant descriptors from step 2 (maximum 15). As a fitness-function, we chose the geometric mean between *Sn* and *Sp*, aiming to promote the most balanced models.

To predict test compounds, we characterized the model chemical space (Applicability Domain, AD) as a hyper-rectangle delimited by maximum and minimum values of each descriptor (reported in SI), using what is known as a “bounding-box” approach (Sahigara et al., 2012). Only compounds within the AD were predicted.

The best models were chosen according to: (1) performance in cross-validation and on the test set (in terms of *Sn*, *Sp* and *NER*), (2) complexity (the fewer the nodes, the better) and (3) descriptor interpretability. This was performed to produce predictive/robust models and to gather new mechanistic insights into the structural features that characterize the classes.

The selected models (Table 2) were very simple, comprising of a few univariate splits (from 4 to 9) and a few variables (up to 5). In both cases, only one test compound (mirex and carbon disulfide for class 2 and 3 trees, respectively) was outside the AD. The best performance was obtained for the most critical class. Class 2 compounds are, indeed, underestimated by K_{OW}. On the contrary, class 3 is less critical because it comprises compounds with reduced BCF. Therefore, the slightly lower performance on class 3 is acceptable.

Table 2
Statistics of selected classification trees: characteristics of the model (number of nodes, *k*, and descriptors, *p*) along with *Sn*, *Sp* and *NER* in fitting, cross-validation (5 deletion groups) and on the test set. For the test set, the number of compounds out of AD (out) was also reported. A detailed description of model settings can be found in SI.

ID	Target class	*k*	*p*	FIT			CV			TEST			
				NER	*Sn*	*Sp*	*NER*	*Sn*	*Sp*	out	*NER*	*Sn*	*Sp*
T2	2	9	5	0.85	0.94	0.76	0.75	0.80	0.71	1	0.75	0.75	0.75
T3	3	5	4	0.73	0.70	0.76	0.72	0.70	0.75	1	0.68	0.69	0.68

3.3. Class 2 tree (excess of BCF)

Fig. 2 depicts the tree targeting the classification of class 2 (excess BCF) compounds (labeled as T2). According to observations of PCBs, the first split distinguishes those having 6 to 8 Cl atoms and $K_{OW} > 6$, which are assigned to class 2. T2 is comprised of eight additional nodes based on five molecular descriptors resulting from GA selection, four of which are “manually calculable” from a known 2D molecular structure. Descriptors are briefly described below.

- PCD (Talete srl, 2012) is a path count descriptor, based on graph theory. The molecule is represented as a H-depleted molecular graph whose vertices are non-H atoms and edges are bonds, which can be characterized by paths (i.e., sequences of vertices without repetition). PCD is defined as:

$$\mathrm{PCD} = \ln\left(\frac{\mathrm{A} + \sum_{{}^{m}P_{ij}} w_{ij}}{\mathrm{TPC}}\right) \quad \text{where } w_{ij} = \prod_{b=1}^{m} \pi_{\mathrm{b}}^{*} \quad m \le 10 \tag{3}$$

where A is the number of vertices in the H-depleted molecular graph; ${}^{m}P_{ij}$ denotes a path of length *m* ($m \in [0, 10]$), from the *i*-th to the *j*-th vertex; w_{ij} is the path weight, calculated by multiplying the conventional bond order π_{b^*} (equal to 1, 2, 3 for single, double and triple bonds, respectively, and 1.5 for aromatic bonds) of all *m* edges of the path ${}^{m}P_{ij}$; and TPC is the total number of paths of any length (from 0 to 10). PCD relates to bond type/number and tends to increase with increasing number of multiple bonds (unsaturation) (Table 3).
- X2Av is the average valence connectivity index of order 2 (Kier and Hall, 1981), calculated as follows:

$$\mathrm{X2Av} = \frac{\sum_{j=1}^{K} \left(\prod_{i=1}^{3} \delta_i^V \right)_j^{-1/2}}{K} \tag{4}$$

where *j* runs over all of the *K* 2nd order paths of the molecular graph. Each path is weighted by the product of the valence vertex degree of the three vertices involved in the path (δ^V), which depends on the number of valence electrons (Z^v) and hydrogen atoms (*h*) bonded to an atom, as follows:

$$\delta^V = Z^v - h \tag{5}$$

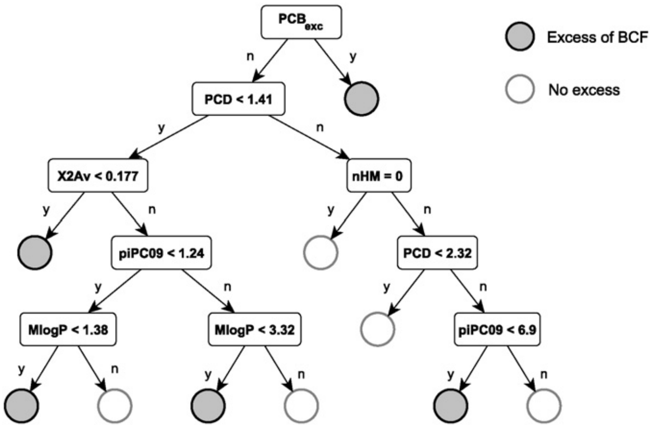
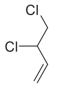
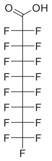
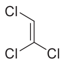
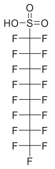
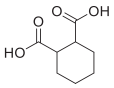
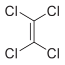
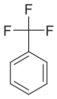
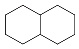
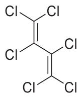
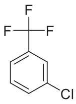
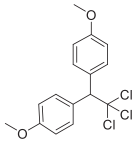
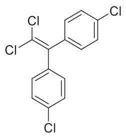
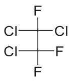
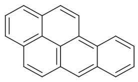


Fig. 2. Selected tree to discriminate class 2 compounds (gray) from other compounds (white). Square boxes denote univariate splits (i.e., nodes), while round boxes (i.e., leafs) denote the assigned class. PCB_{exc} refers to PCBs with 6 to 8 Cl atoms and log $K_{OW} > 6$. The other nodes are labeled according to the descriptor and the corresponding threshold value used for the split. Details about molecular descriptor calculations are given in the text.

Table 3
Examples of PCD, X2Av and piPC09 values. Molecules are sorted in ascending descriptor value order.

PCD		X2Av		piPC09	
Structure	Value	Structure	Value	Structure	Value
	0.205		0.094		0
	0.318		0.118		2.303
	0.343		0.133		2.565
	0.647		0.153		6.055
	2.145		0.225		9.316

- X2Av accounts for the presence of heteroatoms in the molecule as well as double and triple bonds (Table 3). This index decreases when increasing (a) the density of adjacent triplets of vertices with many valence electrons (e.g., F–C–F); (b) the number of cycles (high *K*); and (c) the density of unsaturated/aromatic bonds.
- piPC09 is the count of all of the paths of length 9 in the H-depleted molecular graph. In analogy with PCD (Eq. 3), each path is weighted by the product of the conventional bond order of the involved edges. piPC09 is influenced by molecular size and tends to be higher for polycyclic aromatic molecules, which are characterized by higher bond orders and more paths than aliphatic molecules (Table 3).
- MlogP and nHM are the two simplest descriptors. The former is K_{OW} predicted by the Moriguchi model (Moriguchi et al., 1994), which is based on a group contribution approach; the latter is the number of heavy atoms (i.e., halogens, P, S, Si and Sn in this dataset).

3.3.1. Rationalization

The left branch of T2 contains molecules with a small number of multiple bonds (PCD < 1.41) and among them, those with X2Av < 0.177 can show an excess of BCF. Part of the information encoded within PCD and X2Av is overlapping and opposite because low values of PCD correspond to few multiple bonds, while small values of X2Av correspond to a high density of multiple bonds. However, PCD is influenced more by the molecular dimension than X2Av is. In this leaf, therefore, lay small molecules (i.e., low PCD), with a high density of multiple bonds, in particular aromatic rings. Aromatic rings could be responsible for the excess in BCF, as they are involved in important biological intermolecular interactions, such as bonding between aromatic amino-acid side chains of proteins and (hetero)aromatic rings of small ligands (Meyer et al., 2003). In this leaf, we also find linear molecules, characterized by a high density of adjacent triplets/couples of heteroatoms and few multiple bonds. In this case, the excess could be ascribed to increased molecular flexibility (due to abundance of rotatable bonds), causing a structural rearrangement within organisms, to maximize the interactions between heteroatoms and tissues (Agatonovic-Kustrin et al., 2001). This leaf contains 25% of the class 2 compounds, e.g., all PFAAs, some carbamates and other N/O/Cl rich compounds, meaning that the structural features encoded by PCD and X2Av are relevant for BCF excess. Among the 20 small monocyclic aromatic compounds in this leaf, only the *meta*-substituted anilines and 2-aminopyridine show an excess of BCF.

Concerning the left branch (PCD < 1.41 and X2Av ≥ 0.177), the underestimation by TGD is more likely for low MlogP values. This could be related to errors in the TGD equation itself or to preferential storage within organisms' water phase (e.g., blood) (Dimitrov et al., 2002b; Wang et al., 2014).

The right part of the tree (PCD ≥ 1.41) contains large molecules with many multiple bonds and a low number of heteroatoms. This branch is characterized by a small number of BCF-excess compounds (19 out of 378) and this could be related to: (a) molecular rigidity due to double bond abundance; and (b) the stabilizing effect of neighboring carbon atoms, which could limit structural interaction with macromolecular targets and tissues. In particular, molecules with PCD ≥ 1.41 and no heavy atoms or 1.41 ≤ PCD ≤ 2.32 do not show an excess of BCF; these nodes have a purity of 100% and 98.2%, respectively. The remaining molecules (PCD ≥ 2.32 and nHM > 0) all have two or more aromatic rings and 77 out of 88 contain Cl atoms. They could show an excess of BCF if their piPC09 < 6.9. All except three of the molecules of this node that (a) have at least one O-Cl at topological distance of 4, or (b) have more than three circuits, have piPC09 ≥ 6.9 and do not show an excess of BCF.

3.4. Class 3 tree (deficit of BCF)

The tree for class 3 (T3) is comprised of five nodes and four molecular descriptors (Fig. 3), which are briefly described below.

- ON1V is the overall first-order modified Zagreb index (Bonchev and Trinajstič, 2001), it is a variant of the Kier-Hall valence connectivity index of the 1st order, defined as:

$$\mathrm{ON1V} = \sum_{b=1}^{B} \left(\delta^{V}_{b(1)} \cdot \delta^{V}_{b(2)}\right)^{-1} \qquad (6)$$

where B is the number of bonds, $\delta^V$is the valence vertex degree (Eq. 5) and $b(1)$ and $b(2)$ are the atoms connected by the b-th bond. ON1V increases when increasing the number of carbon atoms (Fig. 4),

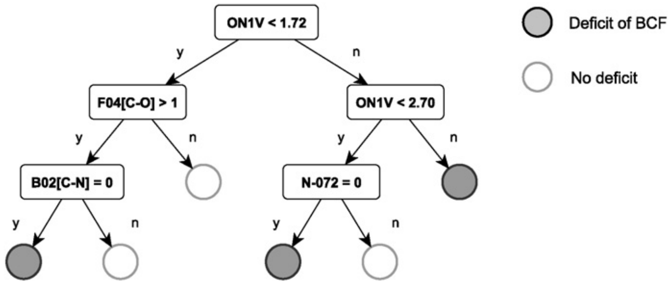


Fig. 3. Selected tree to discriminate class 3 compounds (gray) from other compounds (white). Square and round boxes denote univariate splits (i.e., nodes) and the assigned class (i.e., leaf), respectively. Nodes are labeled according to molecular descriptor acronyms, whose descriptions are given in the text.

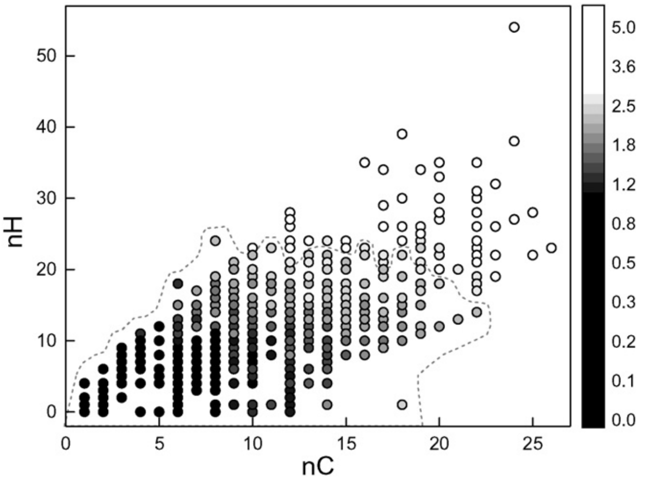
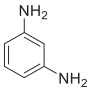
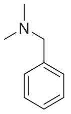
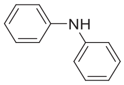
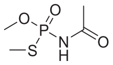
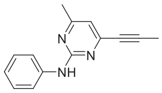
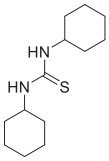
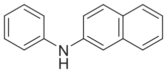
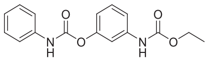
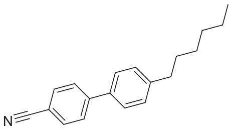
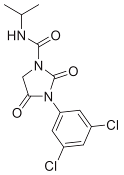


Fig. 4. Variation of ON1V (color map) according to the number of carbon (nC) and hydrogen (nH) atoms. The lighter the circle, the greater the ON1V values; the dashed line delimits the region of ON1V < 2.698. Note that nH is not influential in the ON1V calculation (as it derives from an H-depleted graph). However, nH represents the degree of branching, aromaticity and cyclicity, which tend to decrease the total number of hydrogen atoms bonded to carbon atoms.

Table 4
Examples of ON1V and N-072 descriptor values (in ascending order).

ON1V		N-072	
Structure	Value	Structure	Value
	0.776		0
	1.347		1
	1.728		2
	1.750		2
	2.911		3

reflecting molecular dimension, branching and the presence of heteroatoms (Table 4).

- F04[C–O] and B02[C–N] are 2D atom pairs descriptors (Carhart et al., 1985). F04[C–O] counts the occurrences of connected C and O atoms at a topological distance of 4; B02[C–N] is equal to 1 when there is at least one pair of C and N atoms separated by two bonds, and 0 otherwise.
- N-072 is an atom-centered fragment (Viswanadhan et al., 1989), which counts the occurrence of RCO-N<, RCS-N and >NCX = X (X being any electronegative atom) in the molecule (Table 4).

3.4.1. Rationalization

Large and branched compounds with few heteroatoms (ON1V ≥ 2.698) may show reduced BCF. This leaf, in fact, contains 75% of class 3 molecules (56), but only 25% of class 1 (14) and class 2 (5) compounds. The effect of molecular size on bioconcentration has been an object of debate in the last decades. While several works assert the influence of size on bioconcentration reduction (Dimitrov et al., 2005, 2002a; Opperhuizen et al., 1985), others ascribe the observed deviations to uncertain BCF data (Arnot et al., 2010) and/or to a decreased bioavailability due to sorption to particles of highly hydrophobic chemicals (Schrap and Opperhuizen, 1990). In our case, however, 22% of class 3 compounds have ON1V ≥ 2.698 across the whole range of K_{OW}, ostensibly supporting an effect of molecular size and branching on the reduced bioconcentration. As already hypothesized (Dimitrov et al., 2002a), effective diameter controls membrane permeability and, thus, large and branched compounds could have a limited diffusion through cell membranes, resulting in BCF values smaller than expected.

Among the molecules with ON1V < 2.698, the presence of C and O at lag 4 is often associated with a deficit in BCF. The presence of oxygen atoms has already been related to increased metabolic rates in fish (Arnot et al., 2009), in agreement with our model. It is important to note, however, that among the 310 molecules with F04[C–O] = 0, 46 belong to class 3. Among the molecules with F04[C–O] > 1, those with B02[C–N] = 0 may show a deficit in BCF. These molecules are characterized by a high abundance of aromatic oxygen atoms, phosphate esters, aliphatic ethers and aromatic/aliphatic ketones, fragments related to increased biotransformation rates (Arnot et al., 2009). Among the compounds with B02[C–N] = 1, half (56) contain at least one aromatic nitro-group and 46/56 belong to class 1, while the remaining to class 3 with small residues ($BCF_{TGD} - BCF_{exp} < 0.90$ log units). Hence, it can be stated that these structural features are related to small or no biotransformation. This is in contrast to the study of Arnot et al. (2009). It is important to note, however, that this structural feature concerns only compounds with ON1V < 1.724 and F04[C–O] > 1.

The terminal split of the right branch (N-072), in analogy with the left side of the tree, is based on an N-related descriptor. Even in this case, molecules without N-072 fragments may be metabolized; however, no shared structural characteristics are evident. All PAHs lying in this leaf (5 out of 6) are correctly classified. T3 misclassifies all PFAAs.

3.5. Consensus model

The T2 and T3 trees perform well at the individual class level; however, to allow for model application, one has to be able to classify a given compound into one of the three classes. To this end, we tested the combination of trees in a *consensus* manner, i.e., by assigning a compound to a class if and only if both models agreed. In cases of conflict (i.e., compound predicted as belonging to both classes 2 and 3) no class was assigned. This also allowed assignments of compounds to class 1, i.e., those predicted as both non-excess and non-deficit.

The resulting *consensus* model (Table 5) shows an increased *Sp* for class 3 compounds, meaning that it identifies compounds external to this class well. This is a prominent characteristic, as class 3 is characterized by a decreased bioaccumulation potential. For class 2 compounds, *Sp* increases with respect to T2, with a slight decrease in *Sn*. For the training set, this was caused mainly by the misclassification of PFAAs by T3 and their consequent non-prediction; while for the test set, this was caused by the rejection of two class 2 compounds (correctly

Table 5
Statistics of the *consensus* classification tree for each class on the training and test sets: *Sn*, *Sp* and the number of not predicted compounds (np) are reported.

Class	Training				Test set			
	NER	*Sn*	*Sp*	np	*NER*	*Sn*	*Sp*	np
1	0.69	0.62	0.76	20	0.62	0.57	0.66	9
2	0.87	0.90	0.84	17	0.76	0.71	0.81	2
3	0.74	0.65	0.83	28	0.66	0.58	0.74	16

predicted by T2). Despite class 1 not being modeled, the *consensus* model shows acceptable statistics, especially when considering the *Sp* values.

As an alternative, the models can be combined in a conservative manner: when a compound is predicted as belonging to both classes, it can be assigned to class 2, since this is the most critical class.

4. Conclusions

We presented a scheme to classify compounds as: (1) mainly stored within lipids, (2) affected by additional interactions with non-lipid tissues, or (3) metabolized/eliminated. The scheme is based on two QSAR classification trees, whose salient features are: (1) simplicity, (2) easy applicability, and (3) interpretability.

The best classification performance was shown for compounds with potential increased interactions with tissues (class 2). Structural features connected with their increased BCF are: (a) molecular flexibility and heteroatom density, or (b) the presence of aromatic rings in small molecules. The tree for metabolized/eliminated compounds (class 3) showed a high capability of rejecting false positives. This characteristic allows for a cautionary approach because these compounds show a reduced BCF due to metabolic clearance. Key structural features related to the deficit of BCF resulted are the molecular branching/dimension, and the presence of aromatic oxygen atoms, phosphate esters, aliphatic ethers and aromatic/aliphatic ketones.

Importantly, the trees can be combined into a *consensus* classification scheme, which can serve to assess the reliability of K_{OW}-based predictions of BCF.

CART was chosen for its simplicity at the expense of high model performance. Complex modeling strategies and/or complex molecular descriptors could lead to higher performances, but with a loss of interpretability and applicability. Because we aimed to investigate the mechanisms of bioconcentration, reaching high-performance classification by losing model/descriptor interpretability was beyond the scope of this work. Nonetheless, despite the simplicity of the selected trees and descriptors, the performances were adequate.

In our opinion, attention should be given to class 3 assignments: when a compound is assigned to class 3 and other weight-of-evidence is lacking, the K_{OW}-based BCF should be considered to pursue a cautionary approach. For perfluorinated compounds, the usage of T2 is recommended.

Finally, a major advantage lies in the possibility of calculating the selected descriptors with freely available software, such as E-Dragon (Tetko et al., 2005; VCCLAB, 2005), and, the majority of them, manually, if needed.

Data and software

The dataset, comprised of experimental wet weight fish BCF and K_{OW}, classes and selected descriptors values for 779 chemicals, is freely downloadable from Milano Chemometrics Website (http://michem.disat.unimib.it/chm/download/datasets.htm).

The classification tool will be soon provided as a freely available KNIME workflow, as well (http://michem.disat.unimib.it/chm/download/softwares.htm). Please contact FG for further information.

Acknowledgments

The authors would like to thank Dr. Andrea Mauri for his help with the mathematical definitions of molecular descriptors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.envint.2015.12.024.

References

Agatonovic-Kustrin, S., Beresford, R., Yusof, A.P.M., 2001. Theoretically-derived molecular descriptors important in human intestinal absorption. J. Pharm. Biomed. Anal. 25, 227–237. http://dx.doi.org/10.1016/S0731-7085(00)00492-1.

Armitage, J.M., Arnot, J.A., Wania, F., Mackay, D., 2013. Development and evaluation of a mechanistic bioconcentration model for ionogenic organic chemicals in fish. Environ. Toxicol. Chem. 32, 115–128. http://dx.doi.org/10.1002/etc.2020.

Arnot, J.A., Arnot, M.I., Mackay, D., Couillard, Y., MacDonald, D., Bonnell, M., Doyle, P., 2010. Molecular size cutoff criteria for screening bioaccumulation potential: Fact or fiction? Integr. Environ. Assess. Manag. 6, 210–224. http://dx.doi.org/10.1897/IEAM_2009-051.1.

Arnot, J.A., Gobas, F.A., 2006. A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. Environ. Rev. 14, 257–297.

Arnot, J.A., Meylan, W., Tunkel, J., Howard, P.H., Mackay, D., Bonnell, M., Boethling, R.S., 2009. A quantitative structure–activity relationship for predicting metabolic biotransformation rates for organic chemicals in fish. Environ. Toxicol. Chem. 28, 1168–1177. http://dx.doi.org/10.1897/08-289.1.

Bonchev, D., Trinajstič, N., 2001. Overall Molecular Descriptors. 3. Overall Zagreb Indices. SAR QSAR Environ. Res. 12, 213–236. http://dx.doi.org/10.1080/10629360108035379.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press.

Buckman, A.H., Wong, C.S., Chow, E.A., Brown, S.B., Solomon, K.R., Fisk, A.T., 2006. Biotransformation of polychlorinated biphenyls (PCBs) and bioformation of hydroxylated PCBs in fish. Aquat. Toxicol. 78, 176–185. http://dx.doi.org/10.1016/j.aquatox.2006.02.033.

Carhart, R.E., Smith, D.H., Venkataraghavan, R., 1985. Atom pairs as molecular features in structure-activity studies: definition and applications. J. Chem. Inf. Comput. Sci. 25, 64–73. http://dx.doi.org/10.1021/ci00046a002.

Chen, Y., Guo, Y., Hsu, C., Rogan, W.J., 1992. Cognitive development of yu-cheng ('oil disease') children prenatally exposed to heat-degraded PCBs. JAMA 268, 3213–3218. http://dx.doi.org/10.1001/jama.1992.03490220057028.

Cook, P.M., Robbins, J.A., Endicott, D.D., Lodge, K.B., Guiney, P.D., Walker, M.K., Zabel, E.W., Peterson, R.E., 2003. Effects of aryl hydrocarbon receptor-mediated early life stage toxicity on lake trout populations in Lake Ontario during the 20th century. Environ. Sci. Technol. 37, 3864–3877.

Cronin, M.T.D., Walker, J.D., Jaworska, J.S., Comber, M.H.I., Watts, C.D., Worth, A.P., 2003. Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. Environ. Health Perspect. 111, 1376–1390.

de Boer, J., Stronck, C.J.N., Traag, W.A., van der Meer, J., 1993. Non-ortho and mono-ortho substituted chlorobiphenyls and chlorinated dibenzo-p-dioxins and dibenzofurans in marine and freshwater fish and shellfish from The Netherlands. Chemosphere 26, 1823–1842. http://dx.doi.org/10.1016/0045-6535(93)90077-I.

de Boer, J., van der Valk, F., Kerkhoff, M.A.T., Hagel, P., Brinkman, U.A.T., 1994. An 8-year study on the elimination of PCBs and other organochlorine compounds from eel (*Anguilla anguilla*) under natural conditions. Environ. Sci. Technol. 28, 2242–2248. http://dx.doi.org/10.1021/es00062a007.

de Bruijn, J., Hermens, J., 1991. Uptake and elimination kinetics of organophosphorous pesticides in the guppy (*Poecilia reticulata*): correlations with the octanol/water partition coefficient. Environ. Toxicol. Chem. 10, 791–804. http://dx.doi.org/10.1002/etc.5620100610.

de Wolf, W., de Bruijn, J.H.M., Seinen, W., Hermens, J.L.M., 1992. Influence of biotransformation on the relationship between bioconcentration factors and octanol-water partition coefficients. Environ. Sci. Technol. 26, 1197–1201. http://dx.doi.org/10.1021/es50002a608.

Dimitrov, S.D., Dimitrova, N.C., Walker, J.D., Veith, G.D., Mekenyan, O.G., 2002a. Predicting bioconcentration factors of highly hydrophobic chemicals. Effects of molecular size. Pure Appl. Chem. 74, 1823–1830. http://dx.doi.org/10.1351/pac200274101823.

Dimitrov, S., Dimitrova, N., Parkerton, T., Comber, M., Bonnell, M., Mekenyan, O., 2005. Base-line model for identifying the bioaccumulation potential of chemicals. SAR QSAR Environ. Res. 16, 531–554. http://dx.doi.org/10.1080/10659360500474623.

Dimitrov, S.D., Mekenyan, O.G., Walker, J.D., 2002b. Non-linear modeling of bioconcentration using partition coefficients for narcotic chemicals. SAR QSAR Environ. Res. 13, 177–184. http://dx.doi.org/10.1080/10629360290002299.

ECETOC, 1996. The Role of Bioaccumulation in Environmental Risk Assessment: The Aquatic Environment and Related Food Webs. European Centre for Ecotoxicology and Toxicology of Chemicals. Technical Report No. 67, Brussels.

European Commission, 2003. Technical Guidance Document (TGD) on Risk Assessment in Support of Commission Directive 93/67/EEC on Risk Assessment for New Notified Substances and Commission Regulation (EC) No 1488/94 on Risk Assessment for Existing Substances and Directive 98/8/EC of the European Parliament and of the Council Concerning the Placing of Biocidal Products on the Market. Eur. Community Bruss, Belg.

Geyer, H.J., Rimkus, G.G., Scheunert, I., Kaune, A., Schramm, K.-W., Kettrup, A., Zeeman, M., Muir, D.C.G., Hansen, L.G., Mackay, D., 2000. Bioaccumulation and occurrence of endocrine-disrupting chemicals (EDCs), persistent organic pollutants (POPs), and other organic compounds in fish and other organisms including humans. In: Beek, B. (Ed.), Bioaccumulation — New Aspects and Developments, The Handbook of Environmental Chemistry. Springer, Berlin Heidelberg, pp. 1–166.

Gladen, B.C., Rogan, W.J., Hardy, P., Thullen, J., Tingelstad, J., Tully, M., 1988. Development after exposure to polychlorinated biphenyls and dichlorodiphenyl dichloroethene transplacentally and through human milk. J. Pediatr. 113, 991–995. http://dx.doi.org/10.1016/S0022-3476(88)80569-9.

Gobas, F., Morrison, H.A., 2000. Bioconcentration and biomagnification in the aquatic environment. In: Boethling, R., Mackay, S. (Eds.), Handb. Prop. Estim. Methods Chem.Environ. Health Sci.CRC Press, Boca Raton, USA

Goldberg, D.E., Holland, J.H., 1988. Genetic algorithms and machine learning. Mach. Learn. 3, 95–99. http://dx.doi.org/10.1023/A:1022602019183.

Grisoni, F., Cassotti, M., Todeschini, R., 2014. Reshaped sequential replacement for variable selection in qspr: comparison with other reference methods. J. Chemom. 28, 249–259. http://dx.doi.org/10.1002/cem.2603.

Grisoni, F., Consonni, V., Villa, S., Vighi, M., Todeschini, R., 2015. QSAR models for bioconcentration: is the increase in the complexity justified by more accurate predictions? Chemosphere 127, 171–179. http://dx.doi.org/10.1016/j.chemosphere.2015.01.047.

Hesi, I., 2006. JRC/SETAC-EU. pp. 5–6.

Holland, J.H., 1992. Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications to Biology, Control and Artificial Intelligence. MIT Press, Cambridge, MA.

Jones, P.D., Hu, W., De Coen, W., Newsted, J.L., Giesy, J.P., 2003. Binding of perfluorinated fatty acids to serum proteins. Environ. Toxicol. Chem. 22, 2639–2649. http://dx.doi.org/10.1897/02-553.

Jonsson, G., Bechmann, R.K., Bamber, S.D., Baussant, T., 2004. Bioconcentration, biotransformation, and elimination of polycyclic aromatic hydrocarbons in sheepshead minnows (*Cyprinodon variegatus*) Exposed to Contaminated Seawater. Environ. Toxicol. Chem. 23, 1538–1548. http://dx.doi.org/10.1897/03-173.

Kier, L.B., Hall, L.H., 1981. Derivation and significance of valence molecular connectivity. J. Pharm. Sci. 70, 583–589. http://dx.doi.org/10.1002/jps.2600700602.

Lech, J., Vodicnik, M., 1985. Biotransformation. Fundam. Aquat. Toxicol. Methods Appl.Hemisphere Publ. Corp., Wash. DC, pp. 526–557 (16 Fig 6 Tab 50 Ref)

Lehmler, H.-J., Bummer, P.M., 2004. Mixing of perfluorinated carboxylic acids with dipalmitoylphosphatidylcholine. Biochim. Biophys. Acta Biomembr. 1664, 141–149. http://dx.doi.org/10.1016/j.bbamem.2004.05.002.

Mackay, D., 1982. Correlation of bioconcentration factors. Environ. Sci. Technol. 16, 274–278. http://dx.doi.org/10.1021/es00099a008.

Mansouri, K., 2013. Estimating Degradation and Fate of Organic Pollutants by QSAR Modeling: Contributing to the Implementation of REACH, the European Community Regulation on Chemicals. LAP Lambert Academic Publishing.

MATLAB, 2014. R2014a. The MathWorks Inc., Natick, Massachusetts.

Matthies, M., Solomon, K.R., Vighi, M., Gilman, A., Tarazona, J.V., 2015. On the origin of the criteria for pbt and pop assessment and the evolution of cut-off values. Integr. Environ. Assess. Manag. (in press).

Melancon, M.J., Lech, J.J., 2013. Isolation and identification of a polar metabolite of tetrachlorobiphenyl from bile of rainbow trout exposed to14C-tetrachlorobiphenyl. Bull. Environ. Contam. Toxicol. 15, 181–188. http://dx.doi.org/10.1007/BF01685158.

Meyer, E.A., Castellano, R.K., Diederich, F., 2003. Interactions with Aromatic Rings in Chemical and Biological Recognition. Angew. Chem. Int. Ed. 42, 1210–1250. http://dx.doi.org/10.1002/anie.200390319.

Moriguchi, I., Hirono, S., Nakagome, I., Hirano, H., 1994. Comparison of reliability of log p values for drugs calculated by several methods. Chem. Pharm. Bull. (Tokyo) 42, 976–978.

Muir, D.C.G., Hobden, B.R., Servos, M.R., 1994. Bioconcentration of pyrethroid insecticides and DDT by rainbow trout: uptake, depuration, and effect of dissolved organic carbon. Aquat. Toxicol. 29, 223–240. http://dx.doi.org/10.1016/0166-445X(94)90070-1.

Opperhuizen, A., Velde, E.W.v.d., Gobas, F.A.P.C., Liem, D.A.K., Steen, J.M.D.v.d., Hutzinger, O., 1985. Relationship between bioconcentration in fish and steric factors of hydrophobic chemicals. Chemosphere 14, 1871–1896. http://dx.doi.org/10.1016/0045-6535(85)90129-8.

Pavan, M., Netzeva, T.I., Worth, A.P., 2008. Review of literature-based quantitative structure–activity relationship models for bioconcentration. QSAR Comb. Sci. 27, 21–31. http://dx.doi.org/10.1002/qsar.200710102.

Ratcliffe, D.A., 1967. Decrease in eggshell weight in certain birds of prey. Nature 215, 208–210. http://dx.doi.org/10.1038/215208a0.

Reinert, R.E., Stone, L.J., Willford, W.A., 1974. Effect of temperature on accumulation of methyl mercuric chloride and p,p′-DDT by rainbow trout (*Salmo gairdneri*). J. Fish. Res. Board Can. 31, 1649–1652. http://dx.doi.org/10.1139/f74-207.

Rimkus, G.G., 1999. Polycyclic musk fragrances in the aquatic environment. Toxicol. Lett. 111, 37–56. http://dx.doi.org/10.1016/S0378-4274(99)00191-5.

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., Todeschini, R., 2012. Comparison of different approaches to define the applicability domain of QSAR models. Molecules 17, 4791–4810. http://dx.doi.org/10.3390/molecules17054791.

Schrap, S.M., Opperhuizen, A., 1990. Relationship between bioavailability and hydrophobicity: reduction of the uptake of organic chemicals by fish due to the sorption on particles. Environ. Toxicol. Chem. 9, 715–724. http://dx.doi.org/10.1002/etc.5620090604.

Sijm, D.T.H.M., Opperhuizen, A., 1989. Biotransformation of organic chemicals by fish: enzyme activities and reactions. Reactions and Processes, The Handbook of Environmental Chemistry. Springer, Berlin Heidelberg, pp. 163–235.

Talete srl, 2012. Dragon (Software for Molecular Descriptor Calculation) Version 6.0 — 2012. http://www.talete.mi.it.

Tetko, I.V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., Palyulin, V.A., Radchenko, E.V., Zefirov, N.S., Makarenko, A.S., Tanchuk, V.Y., Prokopenko, V.V., 2005. Virtual computational chemistry laboratory—design and description. J. Comput. Aided Mol. Des. 19, 453–463. http://dx.doi.org/10.1007/s10822-005-8694-y.

Todeschini, R., Consonni, V., 2009. Molecular Descriptors for Chemoinformatics (2 Volumes). Wiley-VCH.

VCCLAB, 2005. Virtual Computational Chemistry Laboratory. http://www.vcclab.org.

Veith, G.D., DeFoe, D.L., Bergstedt, B.V., 1979. Measuring and estimating the bioconcentration factor of chemicals in fish. J. Fish. Res. Board Can. 36, 1040–1048. http://dx.doi.org/10.1139/f79-146.

Viswanadhan, V.N., Ghose, A.K., Revankar, G.R., Robins, R.K., 1989. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. J. Chem. Inf. Comput. Sci. 29, 163–172.

Wang, Y., Wen, Y., Li, J.J., He, J., Qin, W.C., Su, L.M., Zhao, Y.H., 2014. Investigation on the relationship between bioconcentration factor and distribution coefficient based on class-based compounds: the factors that affect bioconcentration. Environ. Toxicol. Pharmacol. 38, 388–396. http://dx.doi.org/10.1016/j.etap.2014.07.003.

White, R.D., Shea, D., Stegeman, J.J., 1997. Metabolism of the aryl hydrocarbon receptor agonist 3,3′,4,4′-tetrachlorobiphenyl by the marine fish scup (*Stenotomus chrysops*) in vivo and in vitro. Drug Metab. Dispos. 25, 564–572.

Woodcroft, M.W., Ellis, D.A., Rafferty, S.P., Burns, D.C., March, R.E., Stock, N.L., Trumpour, K.S., Yee, J., Munro, K., 2010. Experimental characterization of the mechanism of perfluorocarboxylic acids' liver protein bioaccumulation: the key role of the neutral species. Environ. Toxicol. Chem. 29, 1669–1677. http://dx.doi.org/10.1002/etc.199.

Xie, W., Bothun, G.D., Lehmler, H.-J., 2010. Partitioning of perfluorooctanoate into phosphatidylcholine bilayers is chain length-independent. Chem. Phys. Lipids 163, 300–308. http://dx.doi.org/10.1016/j.chemphyslip.2010.01.003.