Science of the Total Environment 470–471 (2014) 1040–1046

Contents lists available at ScienceDirect

Science of the Total Environment

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

Metabolic biotransformation half-lives in fish: QSAR modeling and consensus analysis

Ester Papa [a,*], Leon van der Wal [b], Jon A. Arnot [c,d], Paola Gramatica [a]

[a] QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Theoretical and Applied Sciences, University of Insubria, Varese, Italy
[b] REACH Mastery, Via Giovio 16, 22100 Como, Italy
[c] ARC Arnot Research & Consulting, 36 Sproat Avenue, Toronto, ON, Canada
[d] Department of Physical and Environmental Science, University of Toronto Scarborough, Toronto, ON, Canada

HIGHLIGHTS

- New QSARs provide insights in chemical structure and fish biotransformation rates
- Determination and analysis of the QSARs applicability domain
- Comparison with EPI Suite™ and development of consensus model
- The consensus approach reduced the number of possible false negative predictions
- New QSARs applicable using freely available software

ARTICLE INFO

Article history:
Received 20 August 2013
Received in revised form 22 October 2013
Accepted 22 October 2013
Available online 14 November 2013

Keywords:
Biotransformation rate
Metabolic half-life
QSAR
Consensus modeling
Risk assessment
Chemical prioritization

ABSTRACT

Bioaccumulation in fish is a function of competing rates of chemical uptake and elimination. For hydrophobic organic chemicals bioconcentration, bioaccumulation and biomagnification potential are high and the biotransformation rate constant is a key parameter. Few measured biotransformation rate constant data are available compared to the number of chemicals that are being evaluated for bioaccumulation hazard and for exposure and risk assessment. Three new Quantitative Structure–Activity Relationships (QSARs) for predicting whole body biotransformation half-lives (HL_N) in fish were developed and validated using theoretical molecular descriptors that seek to capture structural characteristics of the whole molecule and three data set splitting schemes. The new QSARs were developed using a minimal number of theoretical descriptors ($n = 9$) and compared to existing QSARs developed using fragment contribution methods that include up to 59 descriptors. The predictive statistics of the models are similar thus further corroborating the predictive performance of the different QSARs; Q^2_{ext} ranges from 0.75 to 0.77, CCC_{ext} ranges from 0.86 to 0.87, RMSE in prediction ranges from 0.56 to 0.58. The new QSARs provide additional mechanistic insights into the biotransformation capacity of organic chemicals in fish by including whole molecule descriptors and they also include information on the domain of applicability for the chemical of interest. Advantages of consensus modeling for improving overall prediction and minimizing false negative errors in chemical screening assessments, for identifying potential sources of residual error in the empirical HL_N database, and for identifying structural features that are not well represented in the HL_N dataset to prioritize future testing needs are illustrated.

1. Introduction

The bioaccumulation assessment of new and existing chemicals is a legal requirement for many regulatory agencies (Government of Canada, 1999; U.S. EPA, 1998; ECHA, 2006). Bioaccumulation information is also used in exposure and risk assessment. Bioaccumulation in aquatic organisms (*i.e.* fish) is the net result of competing rates of chemical uptake (*i.e.* at the surface of the gill and skin from water, in the gastrointestinal tract from food) and elimination processes (*i.e.* gill elimination, biotransformation, fecal egestion, growth dilution) (Gobas et al., 2009; Burkhard et al., 2012). At steady state, metrics for assessing bioaccumulation include the field-based bioaccumulation factor (BAF), biomagnification factor (BMF), and trophic magnification factor (TMF) and the laboratory derived bioconcentration factor (BCF) and BMF (Gobas et al., 2009; Burkhard et al., 2012). Field-based metrics include all routes of chemical exposure simultaneously whereas most controlled laboratory tests focus on one route of chemical exposure at a time

[*] Corresponding author at: QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Theoretical and Applied Sciences, University of Insubria, Via J.H. Dunant 3, 21100 Varese, Italy. Tel.: +39 0332421552; fax: +39 0332421554.
E-mail address: ester.papa@uninsubria.it (E. Papa).

0048-9697/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.scitotenv.2013.10.068

(*i.e.* water or diet for BCF and BMF, respectively). Standardized laboratory bioaccumulation tests (OECD, 2012) can also provide kinetic information such as uptake and elimination rate constants (*k*; /day) and half-lives (*HL*; days).

A few hundred chemicals have been subject to laboratory bioaccumulation tests; however, several thousand chemicals require assessment (Arnot and Gobas, 2006; Weisbrod et al., 2007). Bioconcentration and bioaccumulation models have been developed to address scientific data gaps in regulatory requirements and to provide mechanistic insights into the bioaccumulation phenomenon. Kinetic mass balance models that simulate bioconcentration and bioaccumulation use sub-models for the individual uptake and elimination rate constants. Various empirical, theoretical and Quantitative Structure–Activity Relationship (QSAR) models have been developed for gill uptake and elimination rate constants (*e.g.* Barber, 2003) and for dietary uptake and fecal egestion and growth dilution rate constants (*e.g.* Barber, 2008). The first-order, whole body, primary biotransformation rate constant (k_M) and corresponding *HL* are shown to be key parameters determining the overall bioaccumulation potential of a chemical in fish and in food webs, particularly for hydrophobic chemicals (Burkhard, 2003); however, few data and predictive tools for this parameter exist. A QSAR for predicting k_M from chemical structure has been developed (Arnot et al., 2009) and is included in the U.S. Environmental Protection Agency's Estimation Program Interface EPI Suite™ package Ver.4.1 (U.S. Environmental Protection Agency, 2011) and in the OECD Toolbox (http://www.oecd.org/document/54/0,3746,98en_2649_34379_42923638_1_1_1_1,00.html; (Organization for Economic Co-operation and Development, 2013)). The "k_M-QSAR" included in EPI Suite™ was developed using 57 molecular fragments, the octanol-water partition coefficient (K_{OW}) and molar mass (Arnot et al., 2009). Other QSAR approaches for predicting primary biotransformation half-lives have recently been developed (Brown et al., 2012) and on-going research seeks to estimate biotransformation rate constants *in vitro* and then scale these rates to liver clearance and then to whole body k_M values (Nichols et al., 2007, 2006). Given the relative importance of k_M on the bioaccumulation process, the limited availability of measured data, the extensive costs associated with bioaccumulation testing, and the need to reduce animal testing (Weisbrod et al., 2007), there is a need to develop and evaluate more *in silico* tools for predicting k_M.

The main objective of the present study was to develop and evaluate new, simple and statistically valid QSAR models to predict k_M in fish. The new QSARs were developed and evaluated with an *in vivo* k_M database using a limited number of theoretical molecular descriptors (excluding K_{OW} and molar mass), and in compliance with the OECD Principles for development and validation of QSARs (Organization for Economic Co-operation and Development, 2004). The new molecular descriptors provide additional insights into chemical structure and biotransformation rate capacity in fish. The new QSARs also include the ability to determine the applicability domain for predictions thus providing some guidance on the applicability of the model for chemicals that may not be well represented in the QSAR training and test sets. Finally, the new QSARs developed here are compared with the "k_M-QSAR" in EPI Suite™ and the concept of consensus modeling (Papa and Gramatica, 2010; Fernandez et al., 2012) is illustrated for fish k_M predictions.

2. Material and methods

2.1. Input data

A mass balance method was developed to estimate *in vivo* whole body metabolic biotransformation rate constants (*i.e.* k_M; /day) in fish from laboratory bioaccumulation data (Arnot et al., 2008a). The method includes a screening level uncertainty analysis for the k_M estimates and was applied to a database of evaluated laboratory bioaccumulation test data to derive an *in vivo* k_M database (Arnot et al., 2008b, and available on-line via www.arnotresearch.com). The biotransformation rate constants from a range of fish species, body sizes and temperatures were normalized to rate constants for fish with a body weight of 0.01 kg at 15 °C, *i.e.*, noted as $k_{M,N}$ (Arnot et al., 2008b) and subsequently converted to normalized biotransformation half-life values (HL_N; days) as $HL_N = \ln 2/k_{M,N}$. The half-life is selected as the parameter for modeling because half-lives (expressed in terms of time) are intuitively easier to comprehend and compare than rate constants (expressed in terms of inverse time). The HL_N are further expressed in base 10 logarithmic units ($\log HL_N$), with a range from −1.6 to 3.0, to normalize the response. The *in vivo* k_M database used to generate the QSAR models includes values for 632 chemicals. The chemicals in the data set have molecular weight values ranging from 68.8 to 959.2 g/mol, and $\log K_{OW}$ values ranging from 0.31 to 8.70. The data set is highly complex and heterogeneous, including structurally diverse chemicals such as halogenated organics (polychlorinated biphenyls; PCBs, dioxins; PCDDs, and furans; PCDFs), aliphatic and aromatic hydrocarbons (polycyclic aromatic hydrocarbons; PAHs), amines, imides, alcohols, phenols, ethers, ketones and esters. Therefore the dataset is considered generally representative of metabolic half-lives in fish for most principal functional groups present in pesticides and various organic pollutants of interest for their possible impacts on the environment and on human health.

Due to uncertainty in estimating HL_N for appreciably dissociated ionogenic organic chemicals (IOCs), limited HL_N data for such chemicals are included in the current database (Arnot et al., 2008b). Brown et al. (2012) examined their HL_N QSAR predictions for neutral and appreciably dissociated IOCs and found only minor differences in the predictive capacity of the QSARs for IOCs, based on available data. To further examine this issue with the new QSAR models developed here, IOCs were identified by calculating acidic and basic pKas using JChem (Instant JChem 5.5.0, Chem Axon (http://www.chemaxon.com)) (Brown et al., 2012), and ACD Labs software (Release 12.00 product version 12.5, Build 47877, April 2011).

2.2. Molecular structures and descriptors

Chemical structures of all the 632 compounds were manually designed, checked and energetically optimized by the software HyperChem ver. 7.3. The molecular structures were then used to calculate mono-, and bi-dimensional molecular descriptors, as well as PubChem molecular fragments, by the software PaDEL-Descriptors ver. 2.17 (Yap, 2011).

The use of three-dimensional descriptors was avoided since their calculation depends on 3D conformations of the studied chemicals and may give problems of reproducibility of the models. Constant descriptors, and descriptors found to be correlated pair-wise (correlation greater than 0.98) were excluded from the total amount of descriptors generated in PaDEL-Descriptors (*i.e.* 1567 descriptors), to minimize redundant information. The procedure of cleaning and reduction of the dataset was performed by the software QSARINS (Gramatica et al., 2013). A final set of 520 descriptors was used as input variables for the variable subset selection procedure, resulting in nine selected variables in each of the final QSAR models.

The list of the studied 632 chemicals and calculated values of the descriptors selected in the QSAR models developed in the present study are reported in Supplementary data (Tables S1–S2 Appendix B).

2.3. Data splitting

Data were split into three independent training sets (*i.e.* used to develop the models) and prediction sets (used to evaluate the predictive ability of the models), in a 2:1 proportion, using different approaches.

The first data splitting scheme (Split-1) was the same as the one used previously by Arnot (Arnot et al., 2009) (*i.e.*, 421 training, and 211 prediction chemicals). This data split was performed manually by analyzing the structural similarity of the compounds, so that training and prediction sets would cover the same structural domain (structurally biased procedure — Split-1). Chemicals with the highest and the lowest value of the response are included in the training set. This splitting guarantees that the prediction set spans the entire range of the experimental measurements and is numerically representative of the data set.

The second splitting scheme (Split-2) was obtained by random selection through ordered response of the samples (response biased procedure). In this approach chemicals are ordered according to descending values of the response, and then chemicals are randomly put in the training set (*i.e.*, 405 training, and 227 prediction chemicals). Chemicals with the highest and the lowest value of the response are also included in the training set. However such splitting does not guarantee that the training set represents the entire chemical space of the original dataset (structurally unbiased).

The third splitting scheme (Split-3) was the same as the one reported by Brown and colleagues (Brown et al., 2012) for the HL_N dataset (*i.e.*, 421 training, and 211 prediction chemicals) and took into account both the range of the response and the structural similarity (Brown et al., 2012).

The presence of molecules with different stereochemistry in the original data-split dataset was previously highlighted by Brown (Brown et al., 2012). However stereochemistry couldn't be captured by the 2D-descriptors calculated by Brown, nor by the molecular descriptors generated in this study. To solve this problem Brown et al. (2012) averaged the value of the endpoint and treated each unique 2D structure as a single data point (this reduced the total number of compounds from 632 to 619, of which 412 and 207 chemicals were used in the training and in the prediction set, respectively, to develop the IFS-HL_N model). In this paper different data reported for stereoisomers were not averaged in Split-1 and 3 in order to keep the original splitting schemes as reported in literature (Arnot et al., 2009; Brown et al., 2012). Differently, in splitting scheme 2 the worst case value (or values) of half-life (longest half-life) was kept in the training set for each group of stereoisomers. The three splitting schemes as well as molecules with identical 2D structures is reported in Supplementary data (Table S1-Appendix B).

2.4. QSAR models development and validation

Multiple linear regression analysis using the Ordinary Least Square (OLS) regression was performed to model the studied response ($\log HL_N$). Three populations of QSAR models were generated separately for each of the three training sets by selection, from the input pool of 520 molecular descriptors, of the combinations with the highest modeling performance, and according to their Q^2 leave-one-out (Q^2_{LOO}) values. The variable selection procedure was performed in two steps by the software QSARINS (Gramatica et al., 2013). First, the *All Subset*-method was applied to explore all possible combinations of the available descriptors to include up to two variables into the models. In a second step the *Genetic Algorithm*-method (GA) was applied to optimize the selection and to include up to nine variables into the models. The GA selection procedure was stopped when the inclusion of additional variables did not significantly increase the predictivity of the models.

To reduce the possibility of multi-collinearity, the regressions were only calculated for variable subsets with an acceptable multivariate correlation with response, established by applying the QUIK rule (Q Under Influence of K) (Todeschini et al., 1999). This procedure excludes models which have a K multivariate correlation index of the [X] variable block greater than the correlation within the [X + Y] block variables, where Y is the response variable. Y-scrambling was also applied (500 scrambles) to exclude the possibility of chance correlation (Q^2scr.).

The robustness of the models and their internal predictivity were evaluated by applying the "leave many out" method, leaving out 50% of the chemicals and calculating the Q^2 leave-many-out (Q^2_{LMO}) through 5000 iterations. The best internally validated models in the three populations developed by the variable selection were then verified for their external predictivity in order to be proposed as reliable and predictive QSAR models (Organization for Economic Co-operation and Development, 2004, 2007; Gramatica, 2007). Different parameters were compared to quantify the external predictivity of the models, such as Q^2_{ext} using three different formulas (Consonni et al., 2009), and the concordance correlation coefficient (CCC) (Chirico and Gramatica, 2011). Details and formulas used to calculate these parameters are reported in the Supplementary data (Table S3 — Appendix B). Additionally, to investigate the stability of the models in prediction, the residual mean squared errors (RMSE) were also calculated, and compared for the residual mean squared errors of the training (RMSET) and prediction sets (RMSEP).

2.5. Applicability domain

The applicability domain of the model was investigated by the identification of response outliers (*i.e.*, compounds with cross-validated standardized residuals greater than 2.5σ, where σ stands for standard deviation units), and structural outliers (*i.e.*, compounds with a leverage value (h^*) greater than 3 p'/n (h^*)). Following this method p' is the number of model variables plus one and n the number of the objects used to calculate the model that fall outside the chemical domain of the training set (Gramatica, 2007). High leverage chemicals in the training set can influence the selection of the parameters of the models, while data predicted for high leverage chemicals in the prediction set are extrapolated and could therefore be unreliable.

The applicability domain was also graphically investigated through the Williams plot which is the plot of hat values (h) *versus* standardized residuals. In addition, chemicals with residuals in the prediction set that were larger than 1 log unit were identified and compared individually across the analyzed models.

3. Results and discussion

3.1. Modeling results

The variable selection procedure was started separately for each of the training sets in order to obtain different models with different descriptors and domains. Every model was evolved until a maximum of nine variables were reached: the performance achieved at this level of complexity was considered as satisfactory and the number of descriptors was not further increased to avoid the possibility of over-fitting and to propose externally predictive models that are as simple as possible. This approach resulted in three models (one from each population of models evolved by Genetic Algorithm for each training set), which were chosen as the most predictive independent of the applied splitting scheme. Values of the descriptors selected in the three models and performances of individual models evaluated on different splitting schemes, are reported in Table S2-Appendix B and Table S4-Appendix B. The equations and comparison of internal and external predictivity across the selected models are reported in Table 1. Additional details regarding models statistics, graphs, predictions, residuals, diagonal $HAT_{i,i}$ values, and standardized residuals calculated for models M1–M3, are reported in Supplementary data (Appendix A, Tables S1–S3, Figs. S1–S3; Appendix B, Table S5).

Three descriptors that recur in all models listed in Table 1 are: a vertex adjacency index (VAdjMat), the number of halogens (nX), and the minimum E-State energy for hydrogen bond donors (minHBd). These descriptors are the most relevant for modeling of the selected response, which is confirmed by their standardized regression coefficients. The first two descriptors give information about molecular dimension, hydrophobicity and presence of halogen atoms (VAdjMat, nX). The correlation among VAdjMat and $\log K_{OW}$ (from EPI Suite) is 62%, and the correlation among nX and VAdjMat and molecular weight

Table 1
Best individual models selected by Genetic Algorithm on the basis of different splitting schemes.

Models	Splitting	Equation	$N_{vars.}$	R^2	Q^2_{LOO}	Range Q^2_{ext}	CCC_{ext}	RMSET	RMSEP
M1	Training Obj. 421 Test Obj. 211 (Split-1)	Log $HL_N = -4.081 + 1.082$ VAdjMat $- 0.122$ gmax $- 0.205$ nHBAcc $+ 0.119$ nX $- 0.116$ SaaaC $+ 0.387$ FP503 $+ 2.294$ FP29 $- 0.666$ minHBd $+ 0.241$ ndSCH	9	0.74	0.73	0.76–0.77	0.87	0.60	0.56
M2	Training Obj. 405 Test Obj. 227 (Split-2)	Log $HL_N = -3.883 + 0.400$ nX $+ 1.052$ VAdjMat $- 0.111$ gmax $- 0.147$ nHBa $- 0.007$ ATSm4 $+ 0.149$ MDEC-11 $- 0.853$ minHBd $- 0.095$ SaaaC $- 0.188$ nHCHnX	9	0.75	0.73	0.76–0.77	0.86	0.59	0.56
M3	Training Obj. 421 Test Obj. 211 (Split-3)	Log $HL_N = -4.19 + 1.058$ VAdjMat $- 0.254$ MAXDP $+ 0.112$ nX $- 0.154$ nHBAcc $+ 0.154$ MDEC-11 $+ 0.464$ FP362 $- 0.141$ naaaC $- 0.804$ minHBd $- 0.311$ FP376	9	0.76	0.74	0.75	0.86	0.57	0.58

is 61% and 75%, respectively. The last descriptor (minHBd) describes the ability of the chemical to participate in intramolecular interactions, such as hydrogen bonds.

Two other modeling descriptors of importance, the maximum electrotopological state (gmax Kier and Hall, 1999) and the maximum electrotopological positive variation (MAXDP), are alternatively selected in the three models because they are internally correlated (correlation = 0.98). These descriptors are calculated from the electronic environment of each atom which is related to its intrinsic electronic properties. Also the variables naaaC (count of atom-type E-State: ::C:) and SaaaC (sum of atom-type E-State: ::C:), which describe the number of ring juncture carbons in fused rings and their sum, respectively, are alternatively selected in the three models because they are internally correlated (correlation = 0.99). It should also be noted that the correlation among logK_{OW} and all above mentioned descriptors except VAdjMat is always lower than 45%.

The other variables selected in the proposed models encode for specific information, reported in Table 2, related to the structural domains for each of the training sets depending on its chemical composition.

These variables have the smallest standardized coefficients in their respective equations, and are counters of atom fingerprints (FP29, FP362, FP376 and FP503), functional groups and bonds (nHBAcc, nHCHnX, ndSCH, MDEC-11), the topological autocorrelation index ATSm4, and the E-State descriptor nHBa. These descriptors encode for ring systems (rings, fused rings, multiple ring systems…), presence of aromaticity, and specific atoms, such as chlorine and nitrogen, which are relevant to improve the modeling of each different training set.

Table 2
Descriptor details for less relevant descriptors used in QSAR models M1, M2 and M3.

Descriptor name	Descriptor definition
nHBa	Count of E-States for (strong) hydrogen bond acceptors
nHBAcc	Number of hydrogen bond acceptors (using CDK HBondAcceptorCountDescriptor algorithm)
ndSCH	Count of atom-type =CH–
nHCHnX	Count of atom-type H E-State: CHnX
MDEC-11	Molecular distance edge between all primary carbons
ATSm4	Autocorrelation descriptors weighted by atomic mass
FP29	Pubchem fingerprint — count of individual chemical atoms ≥2Si (ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt)
FP362	Pubchem fingerprint — simple atom nearest neighbors: C(~Cl)(:C) These bits test for the presence of atom nearest neighbor patterns, regardless of bond order (denoted by "~") or count, but where bond aromaticity (denoted by ":") is significant. (ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt)
FP376	Pubchem fingerprint — simple atom nearest neighbors: C(~N)(:C) These bits test for the presence of atom nearest neighbor patterns, regardless of bond order (denoted by "~") or count, but where bond aromaticity (denoted by ":") is significant. (ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt)
FP503	Pubchem fingerprint — simple smart pattern: Cl–C:C–[#1] These bits test for the presence of simple SMARTS patterns, regardless of count, but where bond orders are specific and bond aromaticity matches both single and double bonds. (ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt)

In general, moving from slowly biotransformed chemicals to chemicals that are relatively quickly biotransformed, an increase in MAXDP, gmax and minHBd values is observed while VAdjMat and nX values decrease (trends of the descriptors are reported in Supplementary data, Appendix A — Figs. S4a and S4b). In particular, Fig. 1 shows that most of the slowly metabolized compounds have VAdjMat values between 4.5 and 6, MAXDP values between 0 and 1 (gmax values between 1 and 3), and minHBd = 0.

This means that in the current dataset slower biotransformation (*i.e.*, longer HL_N) is associated to chemicals with no, or limited, ability to participate in non-covalent intramolecular interactions. These chemicals are characterized by large hydrophobic, halogenated, structures (mainly chlorinated), with few or no ramifications, and one or more aromatic rings. Large, non-aromatic, ring systems are also included in this group of chemicals such as hexabromocyclododecane (HBCD), chlordane and siloxanes, as well as long aliphatic chains (more than ten carbon atoms). It is interesting to note that the models appear to be sensitive also within structurally similar groups of chemicals, *i.e.*, they are able to correctly predict PCBs (5–10 Cl atoms) with longer HL_N than polybrominated diphenyl ethers (PBDEs), polychlorinated diphenyl ethers (PCDEs), PCDDs and PCDFs with the same range of halogen atoms. The prediction of HL_N of fluorinated compounds and siloxanes seems to be more difficult and may reflect the fact that these structures have limited representation in the current HL_N dataset. In particular M1 is the only model able to predict all the chemicals containing Si atoms with residuals <1 log unit.

The increasing presence of polar and ionizable groups as well as the number and variety of reactive functional groups simultaneously present in the molecules is, in the studied dataset, generally associated with faster biotransformation rates (*i.e.*, shorter HL_N).

Descriptors SaaaC and naaaC allowed for the distinction among fused aromatic ring systems and other aromatic compounds. This was particularly relevant for the correct prediction of HL_N for specific classes of chemicals such as PAHs, for which the empirical HL_N is lower compared to other classes of halogenated aromatic compounds (*i.e.* PCBs, PBDEs, PCDEs, PCDDs, PCDFs, *etc.*).

The statistical parameters calculated for the models reported in Table 1 reflect the satisfactory ability to fit the data points in the training sets (R^2 range: 0.74–0.76) as well as to predict the log HL_N values of prediction set chemicals not used in the development of the respective models (Q^2_{ext} range: 0.75–0.77; CCC range: 0.86–0.87).

All the models appear to be robust and externally predictive also when they are applied to generate predictions on different splitting schemes. For all the models presented here the graphic output for the analysis of the applicability domain, the calculated standardized residuals in the prediction, and the leverage values for all molecules are available in the Supplementary data (Appendix A Tables S1–3, Figs. S1–3). In particular, 13 chemicals were found to be structurally influential in all of the models, *i.e.*, Octaethylene glycol monotridecyl ether (no. 2), Benzo[a]pyrene (no. 15), Dibenzo[a,h]anthracene (no. 17), Tetradecamethylcycloheptasiloxane (D7; no.161), Perylene

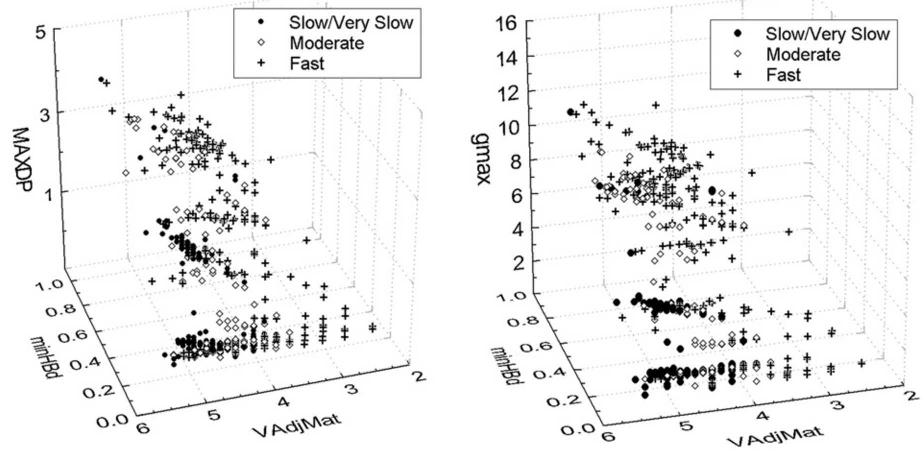

Fig. 1. 3D plots of the values of the main descriptors included in M1–M3 models for chemicals labeled according to their biotransformation half-life: Slow/Very Slow (half-life > 10 day), Moderate (1 < half-life < 10 day) and Fast (half-life < 1 day) (Arnot et al., 2009).

(no. 224), Benzo(k)fluoranthene (no. 226), Benzo[b]chrysene (no. 228), Dodecamethylcyclohexasiloxane (no. 258), Decamethylcyclopentasiloxane (D5; no. 259), 175 Factor L (no. 620), 175 Factor J (no. 621), Spinosad factor D (no. 622), Spinosad factor A (no. 623). Most of these chemicals have particular structures with complex aromatic (or heteroaromatic) and non-aromatic ring systems (*i.e.*, large macrocyclics), and/or large molecular weight (up to 760 g/mol) and represent different classes of compounds such as PAHs, siloxanes, pesticides, and (non-ionic) surfactants. It should also be noted that in general these chemicals are well fitted and predicted by the three models. Only D7, D5 and Spinosad factor A, have residuals that slightly exceeded 1 log unit in two out of three models. Therefore, the 13 aforementioned chemicals broaden the structural domain of the models thereby improving the capacity to predict HL_N for new chemicals with comparable structures. Furthermore, four compounds with standardized residuals that were always slightly larger than 2.5 σ (response outliers), *i.e.*, 1,1,1-(chloromethylidyne)tris-Benzene (no. 43), Dibenzofuran (no. 214), Oxirane [(dibromomethylphenoxy)methyl] (no. 468), and Cis 1,1,3,5 tetramethyl cyclohexane (no. 522). These common outliers, which are all training set chemicals, have a negative effect on the calculation of the fitting parameters of the respective models. However, the removal of these chemicals from the training set is not recommended since it only slightly increased the model fitting statistics but had no effect on the external predictivity of the models.

The new results are comparable to those reported for the IFS-HL_N model developed by Brown et al. (2012) that is based on molecular fragments, and has the following characteristics:

$$\text{Training Obj.: 412, Test Obj.: 207, no. of variables: 36, } R^2 \text{ training: 0.79,}$$

$$R^2\text{test: 0.75; CCC}_{ext}\text{: 0.86, RMSET: 0.53, RMSEP: 0.58.}$$

It is interesting to note that, with the exception of Trihexylsilanol and 5-methyl-2-(1-methylethyl)-cyclohexanol, 14 compounds, already flagged by Brown et al. (2012) as having poor predictions in common with the EPI Suite-HL_N model (Brown et al., 2012, Table SI3), were predicted with absolute residuals greater than one by at least one of the new QSARs developed in the present study. Five of these 14 chemicals show common prediction errors by all the models : Phosphoric acid, 2-chloro-1-(2,4,5-trichlorophenyl)ethenyl dimethyl ester, and [(dibromomethylphenoxy)methyl]-oxirane, are always overestimated, while Dibenzofuran, 1,3,5-trimethyl cyclohexane, and Cis 1,1,3,5 tetramethyl cyclohexane are always underestimated. These underestimations should be highlighted as errors of potential concern. The commonality of the prediction errors suggests that there may also be unresolved errors in the expected (empirical) data.

A final analysis was performed to test the potential effect of the degree of chemical ionization on the performance of the HL_N models based on the currently available dataset. 161 of the chemicals had predicted pKas (*i.e.*, considered to have ionogenic potential) and about 20% of all the ionogenic substances in the dataset (*i.e.*, 37 compounds) had basic or acidic pKa values greater than 5.5 or smaller than 9.5, respectively. These 37 chemicals were considered relevant to examine model performance for appreciably ionized organics compared to the neutral organics. The list of these chemicals is reported in Table S6 — Appendix B. These 37 ionogenic chemicals represent about 6% of the total dataset. By comparison highly hydrophobic chemicals (logK_{OW} values > 7) represent the 11% of the total dataset (*i.e.*, 73 compounds), therefore the ratio ionogenic/highly hydrophobic of about 1:2.

In the new QSARs only seven of these 37 compounds had residuals slightly larger than 1 log unit in at least one of the new QSARs. For these 37 chemicals, the RMSET values calculated by M1, M2 and M3 were 0.54, 0.60 and 0.56 respectively, and RMSEP values were 0.61, 0.65 and 0.65 calculated by M1, M2 and M3, respectively. These results are comparable with those calculated for M1, M2 and M3 training and prediction sets using all chemicals indicate that based on available data the models contain descriptors which are able to adequately predict HL_N for many chemicals with ionogenic features.

3.2. Comparison with EPI Suite-HL_N model

The new QSARs for HL_N were compared with the BCFBAF module from EPI Suite (Arnot et al., 2009):

$$\log HL_N = -1.537 + 0.307 \log K_{OW} - 0.003 \ \text{mol wt.} + \Sigma(\text{Fi} * \text{ni})$$

$$\text{Training Obj.: 421, Test Obj.: 211, no. of variables: 59, } R^2\text{: 0.82,}$$

$$Q^2_{LOO}\text{: 0.75; } Q^2_{ext}\text{: 0.72–0.74, CCC}_{ext}\text{: 0.86, RMSET: 0.49, RMSEP: 0.60} \quad (1)$$

where the element Σ(Fi∗ni) means the sum of coefficients calculated for the molecular fragments included in the model.

The comparison of values reported in Table 1 for models M1, M2 and M3 and the EPI Suite model (referred to below as M4) reported in Eq. (1), shows that M4 has better performance in fitting, due to its higher complexity (*i.e.*, it is based on 57 fragments plus two molecular properties, see Eq. (1)), in comparison to the new QSARs, which are based on only nine molecular descriptors. The external predictivity of

M4 was further evaluated using different measures, *i.e.*, three different external Q^2, the CCC, and a comparison of RMSE values. All the calculated parameters show good accuracy in prediction (Q^2_{ext} range 0.72–0.76, CCC range 0.85–0.86) and consistent RMSE values for the training and the prediction sets of all the models.

In order to analyze compounds which are poorly predicted by the models, an additional evaluation was performed to identify errors in prediction larger than a factor of 10. Ninety three of the 632 compounds (Table S7—Appendix B) were found to have prediction errors larger than a factor of 10 in one or more of the models. The number of overestimations and underestimations is about equally distributed among the models. In particular 42 chemicals, which were previously highlighted (Arnot et al., 2009), shared poor predictions among EPI Suite and the new QSARs. Some considerations are proposed here, and in Appendix B — Table S7, for the 93 chemicals in addition to those already suggested (Arnot et al., 2009). First, 10 chemicals are probably affected by high uncertainty in the empirical data, in fact they are poorly predicted by all the models although the models are based on different structural domains (poor predictions classified as "Data Quality—errors"). Additional testing is needed to confirm the reliability of these data. Second, 49 chemicals are poorly predicted only by some of the models and this might be due by both the presence of uncertainty in the experimental data and by a lack of experimental/structural information in the models (poor predictions classified as "Data Quality—noise"). Third, the remaining 34 chemicals are poorly predicted by only one of the models, therefore these prediction errors appear to depend on the model's domain (poor predictions classified as "Model Domain") and can be easily replaced by choosing a different model. For instance, D5 and D7 are both correctly predicted by at least one of the currently developed models M1–M3 and poorly predicted by EPI Suite. This can, at least partially, be explained by the absence of a specific fragment encoding for Si atoms in the EPI Suite model, and because siloxanes are poorly represented in the training set and in the studied dataset in general. This highlights the need for additional empirical data for *HL* of specific classes of compounds, which are currently missing or limited in the datasets, such as siloxanes. During QSAR model evaluation a thorough examination of the structural and response domain can lead to targeted identification of research needs.

3.3. Consensus model

Consensus modeling can provide more accurate predictions and take into account a wider applicability domain than individual models (Zhu et al., 2008). In this approach the predictions calculated by different models, which are based on a variety of descriptors encoding for different aspects of the molecular structure, are averaged. This is to avoid the fact that any individual QSAR model can overemphasize some of its aspects, underestimate others, or completely ignore important features, depending on the domain of its response and on the structural features included in the training set.

Predictions by all the models reported in Table 1 in addition to predictions by EPI Suite-HL_N model (M4) were combined into a prediction by consensus.

Averaged residuals were calculated across predictions in order to identify those chemicals which HL_N would still be poorly predicted in an averaged by consensus approach. Thirty eight chemicals reported in Supplementary data (Table S8—Appendix B) still have residuals larger than 1 log unit, but only 13 of them are larger than 1.2. As mentioned before, these predictions included about an equal amount of underestimated and overestimated chemicals; however, the 16 underestimated chemicals are of highest concern with respect to bioaccumulation hazard screening since the biotransformation rate predictions for these chemicals are faster than expected according to the corresponding empirical (expected) data. From a regulatory point of view such errors (*i.e.* false negatives) could be more problematic than an overestimation of the biotransformation half-life when screening chemicals for bioaccumulation potential. Among the substances that consistently have HL_N under predicted by the models, two chemicals have errors larger than 1.5, and up to 1.68 log unit (*i.e.* Dibenzofuran, and Cis 1,1,3,5 tetramethyl cyclohexane).

An additional evaluation of the performance of the consensus model *versus* the individual models was carried out considering three general qualitative classes of relative biotransformation rate (or half-life) proposed earlier (Arnot et al., 2009). The percentages of chemicals correctly classified in each class by the different models are reported in Table 3. Additional results are reported in the Supplementary data (Table S9—Appendix B, Fig. S5—Appendix A).

Compared to the single performance of individual models, the consensus approach show a clear improvement for correctly identifying slow/very slow biotransformed compounds (*i.e.* substances of potential regulatory concern). Thus applying the consensus modeling could reduce the number of possible false negatives that would be predicted when using the individual models separately. Furthermore it improves the single ability of the QSAR models M1–M3 to predict slow/very slow and moderate compounds up to levels comparable to the EPI Suite-HL_N model (M4). Many chemicals that fall into the "moderate" category likely have some structural features contributing to facilitated biotransformation and some contributing to reduced biotransformation capacity, thus making reliable model predictions challenging. While in this study the consensus results suggest reduced overall success in categorizing chemicals with short biotransformation half-lives, these types of chemicals will not be bioaccumulative hazards (Burkhard, 2003; Arnot et al., 2008b).

4. Conclusions

The aim of this research was to develop statistically valid multivariate models based on a small amount of theoretical molecular descriptors for the prediction of biotransformation half-lives in fish and to develop a consensus model in order to combine the strengths of the currently available prediction models. An advantage of the models is that they can be used for chemicals before their actual synthesis thus promoting green chemistry. To facilitate the application of the new models they were developed using freely available online software to calculate mono-dimensional and bi-dimensional descriptors.

The chemical applicability domain of the models and the reliability of the predictions were always verified by the leverage approach. A group of five descriptors selected in the three independent runs of the GA for the different splitting schemes identified the basic set of structural information necessary to model the logHL_N response. These variables were associated to: degree of halogenation, molecular dimension and complexity (branching and presence of multiple reactive groups), the tendency to be involved in intramolecular interactions (such as hydrogen bonds), and electronic distribution.

The comparison of the new QSAR models with the existing QSAR in EPI Suite illustrates a new external validation of this widely used software developed by US EPA. The validation was executed by calculating different statistical parameters which demonstrated the robustness and the validity of the EPI Suite model within its domain. However the new QSARs have much lower complexity than the EPI Suite HL_N model, which is based on 59 descriptors.

The analysis of chemicals with poor predictions made by all of the models allowed for the identification of chemicals that may have

Table 3
Percentage of chemicals correctly classified into qualitative biotransformation rate categories.

	M1	M2	M3	M4	Consensus
Very slow + slow ($HL_N > 10$ day)	80%	81%	79%	78%	83%
Moderate	71%	67%	69%	75%	74%
Fast + very fast ($HL_N < 1$ day)	73%	74%	73%	78%	72%

residual error in the dataset and for which additional testing is suggested.

The consensus modeling provided an increase in the prediction accuracy for slow and very slow biotransformed compounds, demonstrating the usefulness of combining different modeling approaches based on different domains to reduce potential errors in screening assessments.

The newly developed QSARs can be used independently or with other QSARs for consensus modeling in order to predict biotransformation rate constants in fish so that such data can be used for bioaccumulation, exposure and risk assessment screening and priority setting when no, or limited empirical data are available.

In order to facilitate these procedures, the QSAR models will be implemented in the software QSARINS (Gramatica et al., 2013) developed within the QSAR Research Unit in Environmental Chemistry and Ecotoxicology at University of Insubria (http://www.qsar.it).

Acknowledgments

The authors would like to kindly thank Dr. Trevor Brown for providing additional information useful for this study. JAA acknowledges funding support from Cefic-LRI ECO.15 project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.scitotenv.2013.10.068.

References

Arnot JA, Gobas FAPC. A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. Environ Rev 2006;14:257–97.

Arnot JA, Mackay D, Bonnell M. Estimating metabolic biotransformation rates in fish from laboratory data. Environ Toxicol Chem 2008a;27:341–51.

Arnot JA, Mackay D, Parkerton TF, Bonnell M. A database of fish biotransformation rates for organic chemicals. Environ Toxicol Chem 2008b;27:2263–70.

Arnot JA, Meylan W, Tunkel J, Howard P, Mackay D, Bonnell M, et al. Quantitative structure activity relationship for predicting metabolic biotransformation rates for organic chemicals in fish. Environ Toxicol Chem 2009;28:1168–77.

Barber MC. A review and comparison of models for predicting dynamic chemical bioconcentration in fish. Environ Toxicol Chem 2003;22:1963–92.

Barber MC. Dietary uptake models used for modeling the bioaccumulation of organic contaminants in fish. Environ Toxicol Chem 2008;27:755–77.

Brown T, Arnot JA, Wania F. Iterative fragment selection: a group contribution approach to predicting fish biotransformation half-lives. Environ Sci Technol 2012;46:8253–60.

Burkhard L. Factors influencing the design of bioaccumulation factor and biota-sediment accumulation factor field studies. Environ Toxicol Chem 2003;22:351–60.

Burkhard LP, Arnot JA, Embry MR, Farley KJ, Hoke RA, Kitano M, et al. Comparing laboratory and field measured bioaccumulation endpoints. IEAM 2012;8:17–31.

Chirico N, Gramatica P. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model 2011;51:2320–35.

Consonni V, Ballabio D, Todeschini R. Comments on the definition of the Q2 parameter for QSAR validation. J Chem Inf Model 2009;49:1669–78.

European Chemical Agency (ECHA). REACH Regulation (EC) No 1907/2006. http://eur-lex.europa.eu/LexUriServ/site/en/oj/2006/l_396/l_39620061230en00010849.pdf, 2006.

Fernandez A, Lombardo A, Rallo R, Roncaglioni A, Giralt F, Benfenati E. Quantitative consensus of bioaccumulation models for integrated testing strategies. Environ Int 2012;45:51–8.

Gobas FAPC, de Wolf W, Burkhard LP, Verbruggen E, Plotzke K. Revisiting bioaccumulation criteria for POPs and PBT assessments. IEAM 2009;5:624–37.

Government of Canada. Canadian Environmental Protection Act, 1999. Canada Gazette part III. Statutes of Canada chapter 33Ottawa, ON: Department of Justice; 1999.

Gramatica P. Principles of QSAR models validation: internal and external. QSAR Comb Sci 2007;26:694–701.

Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S. QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. J Comput Chem 2013;34:2121–32.

HyperChem, rel. 7.03 for Windows. Sausalito, CA (USA): Autodesk, Inc.; 2002.

Instant JChem 5.5.0, Chem Axon. http://www.chemaxon.com. [accessed 2011].

Kier LB, Hall LH. Molecule structure description: the electrotopological state. New York: Academic Press; 1999.

Nichols J, Schultz I, Fitzsimmons P. *In vitro–in vivo* extrapolation of quantitative hepatic biotransformation data for fish — I. A review of methods, and strategies for incorporating intrinsic clearance estimates into chemical kinetic models. Aquat Toxicol 2006;78:74–90.

Nichols J, Fitzsimmons P, Burkhard L. *In vitro–in vivo* extrapolation of quantitative hepatic biotransformation data for fish. II. Modeled effects on chemical bioaccumulation. Environ Toxicol Chem 2007;26:1304–19.

Organization for Economic Co-operation and Development (OECD). OECD Principles for the validation, for regulatory purposes, of (quantitative) structure–activity relationship models. http://www.oecd.org/dataoecd/33/37/37849783.pdf, 2004, [Paris].

Organization for Economic Co-operation and Development (OECD). Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono%282007%292&doclanguage=en, 2007, [Paris].

Organization for Economic Co-operation and Development (OECD). OECD guidelines for testing chemicals. Test no. 305: bioaccumulation in fish: aqueous and dietary exposure; 2012 [Paris].

Organization for Economic Co-operation and Development (OECD). The OECD QSAR toolbox for grouping chemicals into categories. http://www.qsartoolbox.org/index.html, 2013, [Paris].

Papa E, Gramatica P. QSPR as a support for the EU REACH regulation and rational design of environmentally safer chemicals: PBT identification from molecular structure. Green Chem 2010;12:836–43.

Todeschini R, Maiocchi A, Consonni V. The K correlation index: theory development and its application in chemometrics. Chemometr Intell Lab 1999;46:13–29.

U.S. Environmental Protection Agency. Proposed category for persistent, bioaccumulative, and toxic chemical substances. Fed Regist 1998;63:53417–23. [Washington, D.C.].

U.S. Environmental Protection Agency. Estimation Programs Interface (EPI) Suite for Microsoft® Windows, Ver. 4.1; 2011 [Washington, D.C.].

Weisbrod A, Burkhard L, Arnot J, Mekenyan O, Howard P, Russom C, et al. Workgroup report: review of fish bioaccumulation databases used to identify persistent, bioaccumulative, toxic substances. Environ Health Perspect 2007;115:255–61.

Yap C. PaDEL-Descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 2011;32:1466–74.

Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, et al. Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. J Chem Inf Model 2008;48:766–84.