Environ Sci Pollut Res (2014) 21:2955–2965
DOI 10.1007/s11356-013-2247-z

RESEARCH ARTICLE

Modeling bioconcentration factor (BCF) using mechanistically interpretable descriptors computed from open source tool "PaDEL-Descriptor"

Subrata Pramanik · Kunal Roy

Received: 29 August 2013 / Accepted: 14 October 2013 / Published online: 30 October 2013

Abstract Predictive regression-based models for bioconcentration factor (BCF) have been developed using mechanistically interpretable descriptors computed from open source tool PaDEL-Descriptor (http://padel.nus.edu.sg/software/padeldescriptor/). A data set of 522 diverse chemicals has been used for this modeling study, and extended topochemical atom (ETA) indices developed by the present authors' group were chosen as the descriptors. Due to the importance of lipohilicity in modeling BCF, XLogP (computed partition coefficient) was also tried as an additional descriptor. Genetic function approximation followed by multiple linear regression algorithm was applied to select descriptors, and subsequent partial least squares analyses were performed to establish mathematical equations for BCF prediction. The model generated from only ETA indices shows importance of seven descriptors in model development, while the model generated from ETA descriptors along with XlogP shows importance of four descriptors in model development. In general, BCF depends on lipophilicity, presence of heteroatoms, presence of halogens, fused ring system, hydrogen bonding groups, etc. The developed models show excellent statistical qualities and predictive ability. The developed models were used also for prediction of an external data set available from the literature, and good quality of predictions (R^2_{pred}=0.812 and 0.826) was demonstrated. Thus, BCF can be predicted using ETA and XlogP descriptors calculated from open source PaDEL-Descriptor software in the context of aquatic chemical toxicity management.

Keywords QSAR · BCF · PaDEL-Descriptor · Mathematical modeling · ETA · XlogP

Responsible editor: Michael Matthies

Electronic supplementary material The online version of this article (doi:10.1007/s11356-013-2247-z) contains supplementary material, which is available to authorized users.

S. Pramanik · K. Roy (✉)
Drug Theoretics and Cheminformatics Laboratory, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India
e-mail: kunalroy_in@yahoo.com

K. Roy
e-mail: kroy@pharma.jdvu.ac.in
URL: http://sites.google.com/site/kunalroyindia

Introduction

In the context of assessment of toxicity of chemicals on the environment, bioconcentration factor (BCF) plays a significant role in the exploration of chemical accumulation to toxicity levels in living organisms in the aquatic ecosystems (Devier et al. 2003; Hartung 2009; Ahrens and Traas 2007; Tasmin et al. 2013). The intimate exchange of chemicals between the living and non-living members in the ecosystem is an interesting field of interdisciplinary area which draws much attention to many problems in environmental pollution and human health hazards (Arnot and Gobas 2006; Dearden and Hewitt 2010a; Scherb and Voigt 2011). In view of this, BCF plays a critical role to define the ratio of concentration of a particular chemical in a biological tissue to per unit concentration of that chemical in water surrounding that tissue (Roy et al. 2006; Hewitt et al. 2009).

$$\mathrm{BCF} = \frac{C_{\mathrm{organism/tissue}}}{C_{\mathrm{aqueous\ environment}}}$$

BCF assessment is important in the evaluation of risk that chemicals may pose to human and environment, and this has been the recent focus of different environment regulatory authorities (OECD Document 2007; Williams et al. 2009).

However, practical measurement of BCF using experimental models faces bioethical complicacy, lengthy period of experimentation protocol, socioeconomic crisis, and wilderness and complication of in vivo–in vitro experiments (Garrigues 2005; Ahrens and Traas 2007; Hu et al. 2013).

The properties of chemicals are functions of their structural, electronic, molecular microenvironment and molecular architecture (Haranczyk et al. 2012). Similar type of molecules with just a minor structural variation can manifest either different magnitudes of a particular toxicity or quite different types of toxicity (Jorgensen 2010). Over the last few decades, a few studies have been conducted on BCF using in silico methods. A number of studies were focused on experimentally determined physicochemical descriptors including hydrophobicity, computed descriptors like structural and connectivity indices and other types of descriptors; some of the studies used classification-based modeling (Lu et al. 2000a; Liu et al. 2005; Roy et al. 2006; Zhao et al. 2008; Katritzky et al. 2010a). For these simulation studies, in most of the cases, experimentally determined physicochemical properties and/or commercially available in silico tools for descriptor computation have been used. In case of non-availability of values of experimentally derived physicochemical properties and resources of commercial software tools, prediction of BCF values for new query compounds will be problematic. In this perspective, we have used here for modeling BCF data mechanistically interpretable descriptors computed from open source software tool PaDEL-Descriptor (Yap 2011) to identify important molecular features of chemicals contributing to BCF.

A previous report on BCF modeling for a set of 522 diverse chemicals using the formulation of Bayes theorem generated three classification models of chemicals according to their bioaccumulation potency such as non-bioaccumulative (BCF≤2,000 l/kg), bioaccumulative (2,000 l/kg<BCF≤5,000 l/kg), and very bioaccumulative (BCF>5,000 l/kg) chemicals (Fernandez et al. 2012). Discrete Bayes, continuous Bayes, and protective continuous Bayes models using class-dependent thresholds were implemented to classify the chemicals. The classification models were also compared with five established expert systems, which are: (1) (CAESAR, http://www.caesar-project.eu; Lombardo et al. 2010), (2) (TEST, http://www.epa.gov/nrmrl/std/qsar/qsar.html; Dimitrov et al. 2002) (3) BCFBAF/M (Meylan et al. 1999), (4) BCFBAF/A (Arnot and Gobas 2003; Arnot et al. 2008), and (5) (CHEMPROP, http://www.ufz.de/index.php?en=6738; Dimitrov et al. 2005). The classification models based on Bayes theory are important tools to have an estimate about the bioaccumulation potency of chemicals, still molecular properties of chemicals responsible for BCF variations need to be explored (Toropova et al. 2013). It will also be interesting to get an exact predicted BCF data rather than only a classification prediction. In this back ground, the present study has been conducted to develop mathematical regression models using extended topochemical atom (ETA) descriptors originally developed by the present authors' group. These descriptors have been computed from the open source software PaDEL-Descriptor (http://padel.nus.edu.sg/software/padeldescriptor/) for modeling BCF of a large number (N_{total}=522) of environmental chemicals. Due to the importance of lipohilicity in modeling BCF, XLogP (computed lipophilicity) was also tried as an additional descriptor. The objectives of the present study include development of reproducible and interpretable regression-based mathematical models using molecular information of chemicals which can be calculated by open source software PaDEL-Descriptor for predicting BCF values. The developed models were validated by using a test set data and also an external data set available from the literature.

A number of 2D and 3D descriptors were initially generated using open source software PaDEL-descriptor v2.11. The data set was divided using the k-means clustering technique. Genetic function approximation followed by multiple linear regression (GFA-MLR) (Rogers and Hopfinger 1994) and subsequent partial least squares (PLS; Eriksson et al. 2001; Wold 1995) regression were applied to establish linear relationships between structural and molecular information of chemicals and BCF values. Different validation methodologies have been implemented to verify the robustness and external predictability of the developed models. Based on external predictive criteria, PLS-derived models obtained from extended topochemical atom indices (ETA; Roy and Ghosh 2003; Roy and Das 2011) and combination of ETA with partition coefficient (XlogP) (Wang et al. 1997) evolved as the best models.

Materials and methods

In this study, the data set used is comprised of 522 diverse chemicals (n_{total}=522) which were reported by Fernandez et al. (2012; Table S1 in Supplementary materials). These chemicals are present in different aquatic compartments such as water/sediment systems. The chemical domain in this data set includes aliphatic and aromatic hydrocarbons, alcohols, phenols, ethers, esters, anilines, amines, nitriles, nitroaromatics, amides, cyano compounds, organophosphates, heteroaromatics, thiols, phosphate esters, sulfonic acids, organochlorines, halogenated derivatives, and plant secondary metabolites. The application diversity of chemicals covers pharmaceuticals, agrochemicals, industrial chemicals, plant secondary metabolites, and pollutants. Another data set containing 58 polychlorinated biphenyls (Katritzky et al. 2010b) was used to check external predictive quality of the developed models. Structures of all compounds were drawn using MarvinSketch

5.10.0 software (ChemAxon Ltd. http://www.chemaxon.com). The BCF values in log scale (logBCF) were used as the response variable. A set of 2D and 3D descriptors comprising of ETA indices and various non-ETA descriptors (Table S2 in Supplementary materials) were calculated using open source software PaDEL-Descriptor (PaDEL-Descriptor, http://padel.nus.edu.sg/software/padeldescriptor). However, the best models involved only ETA descriptors along with XlogP and these only will be detailed here.

Dataset splitting and model development

As commonly practised in the quantitative structure activity relationship (QSAR) literature, the original dataset (n_{total}=522) was divided into two parts, training set for model development and test set for model validation. The division of the data set into training and test sets was done based on the k-means clustering technique. From each cluster, about 60 % of compounds were selected as the training set members ($n_{training}$=324) and the remaining (40 %) as the test set members (n_{test}=198) (Everitt et al. 2001; Dougherty et al. 2002; Johnson and Wichern 2005). The splitting has been done in such a way that both the sets cover the total chemical space of the whole data set. Statistical techniques like GFA-MLR (Rogers and Hopfinger 1994) and PLS (Eriksson et al. 2001; Wold 1995) were applied to identify the structural and physicochemical features of chemicals contributing to their bioconcentration potency. The PLS technique was used to obviate the problem of inter-correlation among descriptors.

Software

MarvinSketch 5.10.0 software (MarvinSketch 5.10.0, <http://www.chemaxon.com>) was used to draw structures of all compounds. The open source PaDEL-Descriptor software (http://padel.nus.edu.sg/software/padeldescriptor/) was used to calculate ETA indices and non-ETA descriptors. SPSS (SPSS, http://www.spss.com) was used for k-means clustering analysis to divide the data set into training and test sets. Cerius2 version 4.10 (Cerius 2 Version 4.10, Accelrys Inc.) was used for the GFA-MLR analyses. PLS equations were developed using the MINITAB software (MINITAB, http://www.minitab.com). Variable Importance Plot (VIP) and Y-randomization test were carried out using the SIMCA-P software (SIMCA-P 10.0, Umetrics).

Validation metrics for regression-based QSAR models

The robustness and predictive ability of the models was verified by different types of statistical validation metrics. The validation of the developed models was carried out in three ways: (1) leave-one-out internal validation or cross-validation, (2) external validation using the test set chemicals, and (3) evaluation of predictive quality of developed models from prediction of an external data set (Katritzky et al. 2010a). The main objective of this modeling is that the developed model should be robust and reproducible to be able to make accurate and reliable predictions of BCF values of new chemicals. Therefore, mathematical models were developed from the training set and subsequently validated using new chemical entities of the test set and also an external data set (a data set obtained from a completely different source) for checking the predictive capacity of the developed models. The external data set used in this study contains 58 polychlorinated biphenyls compounds. There are 13 compounds common in both data sets (the data sets of 522 and 58 compounds, respectively). Therefore, remaining 45 compounds were used to check predictive quality of developed models. The validation strategies check the reliability of the developed models for their possible application on a new set of data and assess confidence of such predictions. Different statistical parameters like model fitness parameters R^2 and R_a^2; internal validation metrics Q^2, external validation metrics R^2_{pred}, $\overline{r^2_{m(test)}}$ and $\Delta r^2_{m(test)}$ were estimated (Kubinyi et al. 1998a; Roy et al. 2012, 2013) (the relevant mathematical equations have been described in Supplementary material section). Further, predictive qualities of the models were judged based on Golbraikh and Tropsha's approaches (Golbraikh and Tropsha 2002). According to the acceptance criteria set forth by Golbraikh and Tropsha, a model must follow the following conditions:

1. $Q^2>0.5$
2. $r^2>0.6$
3. $\left(r^2-r_o^2\right)/r^2 < 0.1$ or $(r^2-r_o'^2)/r^2<0.1$
4. $0.85\le k\le 1.15$ or $0.85\le k'\le 1.15$

Here, the r^2 and r_o^2 are squared correlation coefficient values between the observed and predicted values (Y and X axes, respectively) of the compounds with and without intercept, respectively. An interchange of the axes gives the value $r_o'^2$ instead of r_o^2. The plot of observed values (Y-axis) against the predicted values (X-axis) of the test set compounds setting the intercept to zero gives the slope of the fitted line as the value of k. The interchange of axes gives the value of k'. Additionally, the developed models were also subjected to a randomization test to check the robustness of the models. The logBCF (Y) values were randomly permuted keeping the descriptor matrix intact followed by a PLS run. The randomization and subsequent PLS analysis generates a new set of R^2 and Q^2 values, which were plotted against the correlation coefficient between the original Y values and the permuted Y values. The corrected R^2_p ($^cR^2_p$) was also calculated based on the results of the randomized data (Schuurmann et al. 2008; Mitra et al. 2010). The models will

be considered valid if R^2_{int}<0.4, Q^2_{int}<0.05 and ${}^cR^2_p$>0.5. The equations for the calculation of statistical parameters have been explained in the Supplementary materials section.

Results

A large number of descriptors were generated using freely available software PaDEL-descriptor (Table S2 in Supplementary materials). The pool of descriptors was subjected to thinning based on preliminary study using genetic algorithm. The molecular properties connected with extended topochemical atom indices (ETA) and partition coefficient (XlogP) were applied for the final mathematical model development as these models were superior in prediction quality than those involving other categories of descriptors. The final robust predictive models were developed in two ways, (1) considering only ETA indices and (2) using ETA along with XlogP as the independent variables. The XlogP is the computed partition coefficient of chemicals based on the summation of atomic contributions including correction factors for some intramolecular interactions. The GFA algorithm was applied to select relevant descriptors, and subsequent PLS analyses were performed to establish mathematical equations for BCF (Table 1). The PLS analysis was done to obviate the possibility of inter-correlation among the descriptors. According to the calculated metrics, the developed PLS equations have good statistical quality and acceptable predictive ability for BCF. Therefore, the best equations using only ETA (Eq. 1) and ETA along with XlogP (Eq. 2) have been further explained.

Model with ETA descriptors

$$\begin{aligned}\log\ \text{BCF} &= 6.340 + 35.356 \times \langle 0.040 - \Delta\varepsilon_D\rangle + 67.375 \times \langle \varepsilon_3 - 0.420\rangle + 11.470 \times \sum \beta'_{ns(\delta)} \\ &\quad -7.731 \times \langle 2.851 - \eta^{\text{local}}\rangle - 11.930 \times \sum \beta'_s - 3.767 \times \Delta\Psi_B + 1.567 \times \langle \Delta\varepsilon_A - 0.108\rangle \\ R^2 &= 0.641,\quad Q^2 = 0.620,\quad R^2_a = 0.638,\quad \text{RMSEP} = 0.780,\quad R^2_{\text{pred}} = 0.659,\quad \overline{r^2_{\text{m(test)}}} = 0.544, \\ \Delta r^2_{\text{m(test)}} &= 0.182,\quad \left(r^2 - r^2_o\right)\Big/ r^2 = 0.001 \text{ and } k = 0.996;\quad \left(r^2 - r'^2_o\right)\Big/ r^2 = 0.121 \text{ and } k' = 0.868,\ {}^cR^2_P = 0.629\end{aligned} \tag{1}$$

The developed model (Eq. 1) shows importance of seven descriptors in BCF prediction. The descriptors in (Eq. 1) have been arranged based on their order of contribution according to the VIP plot (Fig. 1a). These descriptors belong to the class of the extended topochemical atom (ETA) indices which have been developed in our laboratory (Roy and Ghosh 2004; Roy and Das 2011). The detailed explanation of the descriptors may be found in the original papers (Roy and Ghosh 2004; Roy and Das 2011). The descriptor $\Delta\varepsilon_D$ gives the molecular information related to contribution of hydrogen bond donor atoms in the chemicals. Hydrogen bonds are formed when a hydrogen bond donor group (like -OH, -NH_2, -NH-, etc.) donates its covalently bonded hydrogen atom to an electronegative "acceptor" atom (e.g., nitrogen, halogens, and oxygen). The hydrogen bond-contributing groups are important to determine bioconcentration factor of chemicals. In this particular equation, the spline function of $\Delta\varepsilon_D$ ($\langle 0.040 - \Delta\varepsilon_D\rangle$) has a positive coefficient for BCF prediction. A spline function will have a zero contribution when its value is negative. To exert the positive effect, the value of $\Delta\varepsilon_D$ should be less than 0.0394; a value greater than or equal to 0.0394 will show no effect in BCF prediction (a negative value of the spline terms is treated as zero). In summary, a compound with higher number of hydrogen bond donor groups will have lower logBCF value. Compound nos. 13 (3-amino-1,2,4-triazole) and 16 (thiourea) having high values the $\Delta\varepsilon_D$ descriptor (and hence containing hydrogen bond donor groups) have low BCF values. Again, it is observed that chemicals triflumizole (508) and acridine (522) having zero value for the $\Delta\varepsilon_D$ descriptor (and hence not containing hydrogen bond donor groups) have BCF values in higher side. The ε_3 descriptor is calculated from the reference alkane which is obtained by replacing any heteroatoms with carbon and removing multiple bonds. The value of the ε_3 descriptor will be higher for a fused ring systems. The spline term of this descriptor ($\langle \varepsilon_3 - 0.420\rangle$) has a positive contribution in model development. The value of ε_3 should be greater than 0.42 to exert a positive contribution. The heteroaromatics (e.g., dibenz(a,h)acridine and dibenzothiophene) having higher values of this spline term have high bioconcentration values. The increased concentration of these chemicals in organism may lead to their toxicity. The $\sum\beta'_{ns(\delta)}$ descriptor is a measure of lone electrons (not bonded or shared) entering into resonance relative to

Table 1 Statistical quality parameters for PLS equations (Eq. 1 developed from ETA descriptors and Eq. 2 developed from ETA descriptors and partition coefficient (XlogP))

Chemometric tools	Eq. no	No. of descriptors	LVs	R^2	Q^2	$R^2{}_a$	RMSEP	$R^2{}_{\text{pred}}$	$\overline{r^2_{\text{m(test)}}}$	$\Delta r^2_{\text{m(test)}}$
(1) GFA-MLR (spline) followed by PLS	1	7	5	0.641	0.620	0.638	0.780	0.659	0.544	0.182
(2) GFA-MLR (spline) followed by PLS	2	4	2	0.614	0.597	0.611	0.738	0.696	0.580	0.222

RMSEP root mean square error of prediction

molecular size of chemicals. The presence of functional groups like $–NH_2$, -OH, -Cl attached to aromatic nucleus allows the lone pair of electron contributing to their resonance property, and this has a significant role in BCF prediction (positive contribution). However, in the present data set, this descriptor is mostly represented by halogen atoms attached to an aromatic nucleus. For example, chemicals like 2,3′,5,5′-tretrachlorobiphenyl and 2,3′,4,4′,6-pentachlorobiphenyl with high number of chlorine atoms have higher values of $\sum\beta'_{ns(\delta)}$ and consequently have higher BCF values. Thus, the descriptor $\sum\beta'_{ns(\delta)}$ basically indicates the contribution of lipophilic halogen atoms to BCF. The η^{local} descriptor is a local ETA index considering only bonded interactions. The spline term of this descriptor ($\langle 2.851-\eta^{\text{local}}\rangle$) has a negative contribution. A value of η^{local} less than 2.851 will have a negative contribution to BCF, while a value greater than or equal to zero will have no contribution in BCF prediction. The measure of electronegative atom count of the molecules relative to molecular size expressed as $\sum\beta'_s$ has a negative contribution towards BCF. This descriptor has a low contribution in model development. Chemicals [e.g., isocyanuric acid (158) and 1,3,5-tris(2′-hydroxyethyl)cyanuric acid) (302)] with higher electronegative atom count relative to the molecular size have less bioconcentration factor values. Thus, an increase in polarity (due to presence of heteroatoms) tends to decrease bioconcentration factor. The measure of hydrogen-bonding propensity (considering both hydrogen bond donor and acceptor properties) of the molecules ($\Delta\psi_B$) (a high value means low hydrogen bonding propensity) has a negative role in BCF prediction, and this has, however, a small contribution in model development. It appears that this term ($\Delta\psi_B$) penalizes the other hydrogen bonding term of ($\langle 0.040-\Delta\varepsilon_D\rangle$) present in Eq. (1). Chemicals such as 1,1,2,2-tetrabromoethane (45) and 1,2-dibromoethane (147) have higher values of this descriptor and hence lower BCF values. The $\Delta\varepsilon_A$ descriptor is a measure of contribution of unsaturation and electronegative atom count of the chemicals. The spline term ($\langle\Delta\varepsilon_A-0.108\rangle$) of this descriptor has a positive contribution in model development. The positive effect will be accounted if the value of $\Delta\varepsilon_A$ is greater than 0.108, and there will be no contribution if $\Delta\varepsilon_A$ is less than or equal to 0.108. This descriptor has the least contribution in model development. Overall, it is observed that presence of hydrogen bonding groups, halogen atoms, electronegative atoms, and fused ring stem in chemicals are important factors for BCF prediction.

Model with ETA descriptors along with XlogP

$$\begin{aligned}\log\text{BCF} &= 1.600-0.364\times\langle 5.274-X\log P\rangle + 8.038\times\langle 0.129-\Delta\psi_A\rangle \\ &\quad + 12.910\times\langle\Delta\varepsilon_B-0.082\rangle + 0.336\times\sum\beta_{ns(\delta)} \\ R^2 &= 0.614,\ Q^2 = 0.597,\ R^2_a = 0.611,\ \text{RMSEP} = 0.738,\ R^2_{\text{pred}} = 0.696,\ \overline{r^2_{\text{m(test)}}} = 0.580, \\ \Delta r^2_{\text{m(test)}} &= 0.222,\ \left(r^2-r_o^2\right)/r^2 = 0.080 \text{ and } k = 1.029 \text{ or } \left(r^2-r_o'^2\right)/r^2 = 0.115 \text{ and } k' = 0.855,\ {}^cR^2_P = 0.648\end{aligned} \tag{2}$$

The developed model (Eq. 2) using ETA descriptors and partition coefficient (XlogP) shows the importance of only four descriptors in BCF prediction. The descriptors in Eq. (2) have been arranged based on their order of contribution according to the VIP plot (Fig. 1b).

The partition coefficient (logP) of solutes in octanol/water is assumed to be the summation of the contributions of each atom of the molecules.

$$\log P = \sum_i a_i A_i$$

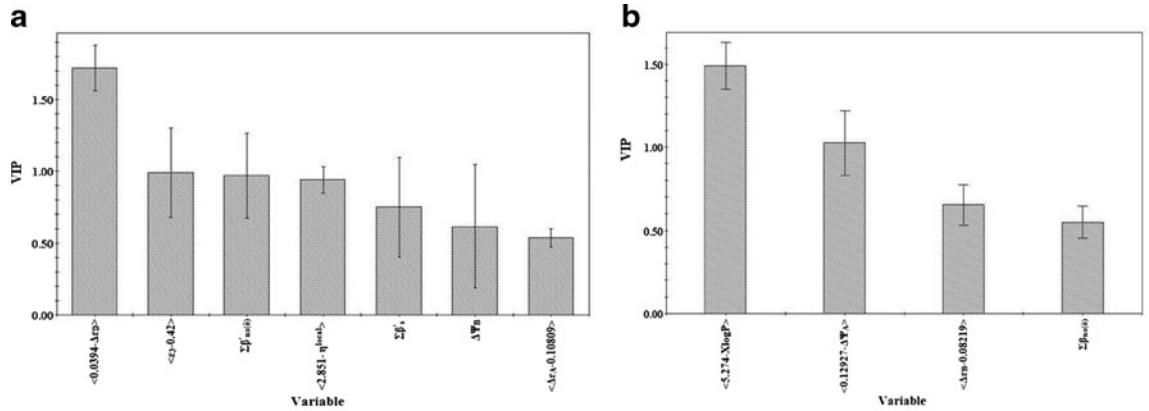

Fig. 1 Variable importance plots (VIP) to show the contributing features: **a** Contribution of descriptors for the model developed from ETA descriptors. **b** Contribution of descriptors for the model developed from ETA descriptors and XlogP

In the above equation, a_i is the contribution of the ith atom, and A_i is the number of occurrences of the ith atom type. This equation ignores the possible interactions among features within the molecule. Considering molecular interaction, the corrected partition coefficient (XlogP) can be described as

$$X\log P = \sum_i a_i A_i + \sum_j b_j B_j$$

Here, a_i and b_j are regression coefficients, A_i is the number of occurrences of the ith atom type, B_j is the number of occurrence of the jth correction factor identified by Wang et al. (1997, 2000). However, this fragmentation scheme is not the part of the current paper, and we have used the XlogP values as calculated from the PaDEL-Descriptor software. The spline term of this descriptor ($\langle 5.274 - X\log P\rangle$) has a negative contribution in the BCF model development. The molecules having values of XlogP less than 5.274 are less concentrated in the aquatic organism (negative values of the spline term are treated as zero). Therefore, it can be assumed that molecules with higher lipophilic property (XlogP>5.274) are highly concentrated in organisms and may produce toxic effects. It is observed from our result that chemicals (e.g., 1,3,5-triazine-2,4,6-triamine, picloram, and pentaerythritol) having higher values of the $\langle 5.274 - X\log P\rangle$ descriptor are less concentrated in aquatic organisms (BCF values are less). The spline term of XlogP suggests a nonlinear dependence of logBCF on the partition coeffiecient values. The $\Delta\psi_A$ descriptor is a measure of hydrogen-bonding propensity of the molecules. The spline term of this descriptor ($\langle 0.129 - \Delta\psi_A\rangle$) has a positive effect in model development. The molecules having hydrogen bonding propensity ($\Delta\psi_A$) less than 0.129 have high bioconcentration factor values. Our results also suggest that chemicals (e.g., 2,3,4,5,6-pentachlorobiphenyl, 2-isopropylnaphthalene, and 4,4′-dibromobiphenyl) with low hydrogen bonding propensity have low BCF potential. The $\Delta\varepsilon_B$ descriptor is a measure of

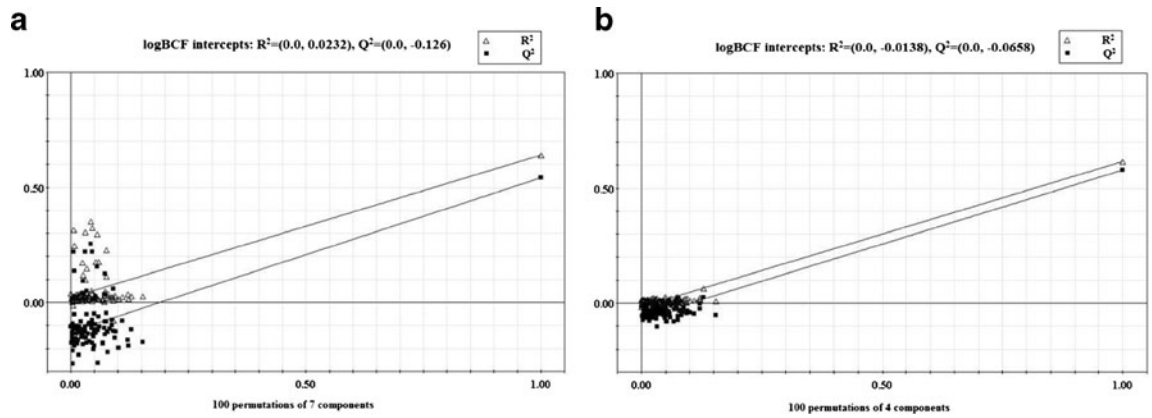

Fig. 2 Randomization test of the developed models (Y randomization at 100 randomization cycles). **a** Validation plot of the developed model from ETA descriptors. **b** Validation plot of the developed model from ETA descriptors and XlogP

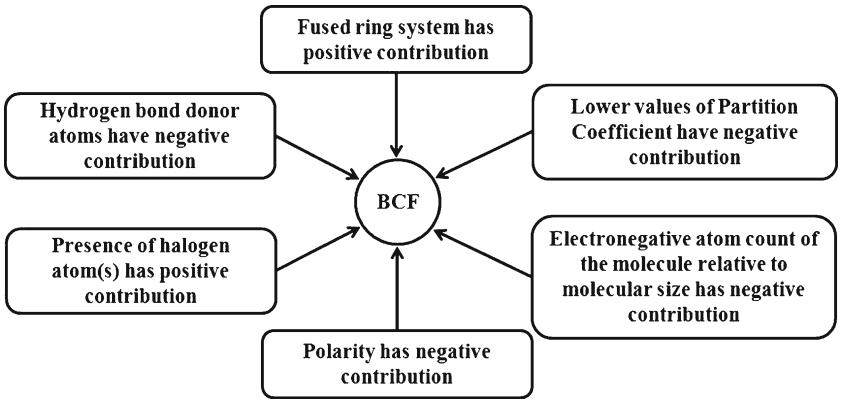

Fig. 3 Molecular features responsible for BCF of the molecules

contribution of unsaturation of molecules. The spline term of this descriptor ($\langle \Delta\varepsilon_B - 0.082 \rangle$) has a positive contribution in model development. The positive effect will be accounted if the value of $\Delta\varepsilon_B$ is greater than 0.082. Our results show that a number of chemicals (e.g., pentachlorobenzene, 1,2,4,5-tetrachlorobenzene, and 1,2,3,5-tetrachlorobenzene) having higher values of this spline term have higher BCF values. It appears that aromatic systems will have higher BCF values. This is on consonance with our earlier finding of positive contribution of fused (aromatic) ring system to BCF in case of Eq. 1. The measure of lone electrons entering into resonance ($\sum \beta_{ns(\delta)}$) has a positive contribution in BCF prediction. This has been discussed in case of Eq. 1, and using the similar logic, this can be attributed to presence of halogen atoms.

The contributions of the descriptors for both models based on variable importance plots (VIP) have been shown in Fig. 1. The Y-randomization test to check robustness of the developed models has been performed, and the results are shown in Fig. 2. The contribution molecular features for BCF prediction are summarized in Fig. 3. The predictive quality and overall performances of the developed models were verified using the scatter plots (Fig. 4).

Discussion

Model validation

The model fitness parameters R^2 and R_a^2 for both models are significant. Acceptable values for the internal (Q^2>0.6) and external (R^2_{pred}, $\overline{r^2_{\text{m(test)}}}$, and $\Delta r^2_{\text{m(test)}}$) validation metrics reflect the predictive potential of the developed models which can be efficiently applied for prediction of new set of chemicals (Table 1). The r_m^2 metrics for judging predictive potential of the models have been computed based on scaling of the response data (Roy et al. 2013). Moreover, external

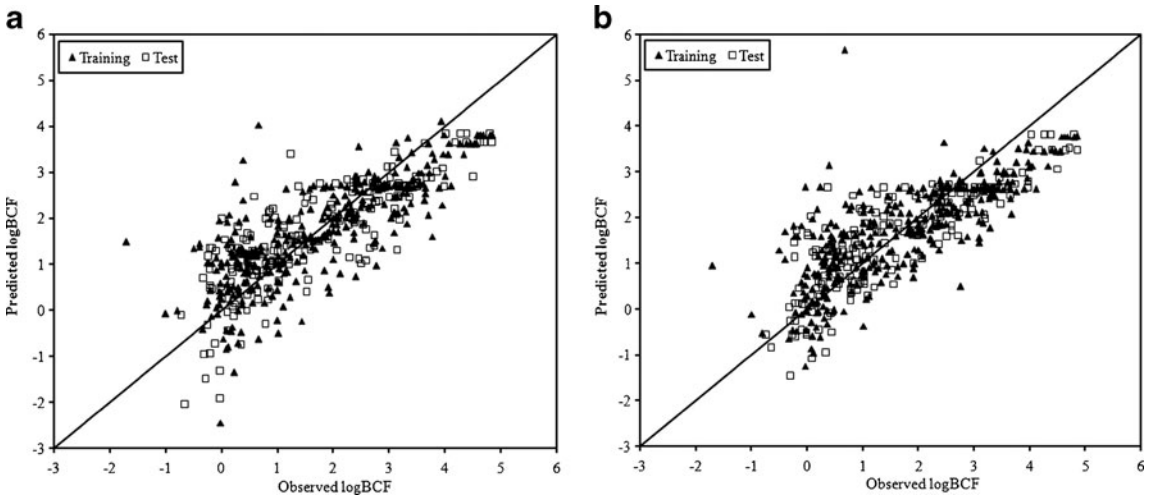

Fig. 4 Scatter plots for **a** the model developed from ETA descriptors and **b** the model developed from ETA descriptors and XlogP

Table 2 Comparison of quality of predictions of the presently developed models for logBCF values of the present test set with those of five established expert models (CAESAR, TEST, BCFBAF/M, BCFBAF/A, and CHEMPROP)

Response variable	Different models	R^2_{pred}	$\overline{r^2_{m(test)}}$	$\Delta r_m^2{}_{(test)}$
log BCF	1. CAESAR	0.828	0.754	0.137
	2. TEST	0.830	0.703	0.143
	3. BCFBAF/M	0.761	0.643	0.186
	4. BCFBAF/A	0.788	0.720	0.154
	5. CHEMPROP	0.625	0.487	0.264
	6. Model developed from ETA descriptors	0.659	0.544	0.182
	7. Model developed from ETA and XlogP descriptors	0.696	0.580	0.222

validation criteria according to Golbraikh and Tropsha approaches have also been checked for both the models.

The robustness of the models was again verified using Y-randomization test. Each randomization and subsequent PLS analysis generates a new set of R^2 and Q^2 values, which were plotted against the correlation coefficient between the original Y values and the permuted Y values (logBCF; Fig. 2). The developed models have accepted values of the intercepts: R^2=(0.0, 0.0232), Q^2=(0.0, −0.126) for the model developed with ETA descriptors, R^2=(0.0, −0.0138), Q^2=(0.0, −0.0658) for the model developed with ETA descriptors and partition coefficient (XlogP). From the Y randomization study, ${}^cR^2{}_p$ values were calculated for both models, which show that models are not obtained by chance.

Comparison of the developed models with other established models

There are a few previous reports on quantitative structure activity relationship (QSAR) studies of BCF modeling. The studies were focused on physicochemical descriptors including hydrophobicity along with structural, connectivity, and other types of descriptors and sometimes using classification-based modeling (Lu et al. 2000b; Liu et al. 2006; Dearden and Hewitt 2010b). External predictive qualities have been to some extent neglected in some of the studies. In the present study, open source software PaDEL-Descriptor has been used for the descriptor generation. In case of non-availability of experimental physicochemical information of chemicals and lack of resources of commercial software tools for computation of descriptors, our developed models can be applied for BCF prediction of new query chemical in the context of aquatic toxicity assessment. The developed models obtained using only ETA descriptors and ETA descriptors combined with partition coefficient (XlogP) have good quality for the test set predictions. The BCF data of the test set used in our study was also predicted using the following expert models: (1) CAESAR, (2) TEST, (3) BCFBAF/M, (4) BCFBAF/A, and (5) CHEMPROP (Table 2). From the comparison of the external prediction capability of our models with those of different expert systems, it appears that our models perform well, though they cannot supersede the qualities of all of the available expert systems. When we have applied our developed models on a true external data set (Katritzky et al. 2010a) for BCF prediction, a very good correlation between observed and predicted BCF values is obtained (Table 3). The model developed with ETA descriptors shows predicted variance, $R^2_{(pred)}$=0.812, and the model developed with ETA descriptors and partition coefficient (XlogP) show predicted variance $R^2_{(pred)}$=0.826. Therefore, the developed models can be applied on diverse environmental pollutant chemicals to predict their BCF in connection with aquatic toxicity control. However, it may be mentioned here that the external set used here belongs to the class of biphenyl derivatives and does not cover the entire chemical domain.

Conclusions

This study suggests that the extended topochemical atom indices (ETA) along with partition coefficient (XlogP) can be used for efficient modeling of bioconcentration factor (BCF). The exchange of chemicals between living and non-living entities in the ecosystem and their concentration into organic tissue can be predicted using our mathematical models. The bioconcentration factor of molecules depends on lipophilicity, presence of heteroatoms, presence of halogens, fused ring system, hydrogen bonding groups, etc. The descriptors used in the present study can be easily computed from the open source software tool PaDEL-Descriptor. In case of the non-availability of experimental physicochemical properties and lack of suitable commercial software for computation of descriptors, our developed models can be applied for BCF calculation of query chemical in the context of aquatic toxicity management.

Table 3 Predictions for an external data set (Katritzky et al. 2010a)

Sl. no.	Compounds	CAS number	Experimental logBCF	Prediction from model 1	Prediction from model 2
1	Biphenyl	92-52-4	2.64	2.166	2.640
2	Biphenyl, 4-chloro-	2051-62-9	2.77	2.558	2.740
3	Biphenyl, 2,3′ -dichloro-	25569-80-6	3.8	2.979	2.829
4	Biphenyl, 2,4-dichloro-	33284-50-3	3.55	4.651	5.501
5	Biphenyl, 2,40-dichloro-	34883-43-7	3.57	4.339	4.731
6	Biphenyl, 4,4′-dichloro-	2050-68-2	3.28	3.616	3.293
7	Biphenyl, 2,2′,5-trichloro-	37680-65-2	4.11	3.381	3.223
8	Biphenyl, 2,4,5-trichloro-	15862-07-4	4.26	3.616	3.293
9	Biphenyl, 2,2′,3,5′ -tetrachloro-	41464-39-5	4.84	2.980	2.829
10	Biphenyl, 2,2′,4,4′-tetrachloro-	2437-79-8	4.85	3.616	3.293
11	Biphenyl, 2,2′,4,5-tetrachloro-	70362-47-9	5	3.997	3.989
12	Biphenyl, 2,2′,6,6′-tetrachloro-	15968-05-5	3.85	2.980	2.963
13	Biphenyl, 2,3,4′,6-tetrachloro-	52663-58-8	4.6	3.997	3.989
14	Biphenyl, 2,3′,4′,5-tetrachloro-	32598-11-1	4.77	2.980	2.753
15	Biphenyl, 2,2′,3,4,5′-pentachloro-	38380-02-8	5.38	3.997	3.989
16	Biphenyl, 2,2′,3,4′,5-pentachloro-	68194-07-0	5	3.997	3.989
17	Biphenyl, 2,2′,3′,4,5-pentachloro-	41464-51-1	5.43	4.172	4.356
18	Biphenyl, 2,2′,4,4′,5-pentachloro-	38380-01-7	5	3.997	3.989
19	Biphenyl, 2,2′,4,5,5′-pentachloro-	37680-73-2	5.4	4.339	4.731
20	Biphenyl, 2,3,3′,4,6-pentachloro-	74472-35-8	5	3.381	3.048
21	Biphenyl, 2,2′,3,3′,4,4′-hexachloro-	38380-07-3	5.77	3.813	3.634
22	Biphenyl, 2,2,3,3′,6,6′-hexachloro-	38411-22-2	5.43	3.813	3.634
23	Biphenyl, 2,2′,3,4,4′,5-hexachloro-	35694-06-5	5.88	3.813	3.634
24	Biphenyl, 2,2′,3,4,4′,5′-hexachloro-	35065-28-2	5.39	3.997	3.989
25	Biphenyl, 2,2′,3,4,5,5′-hexachloro-	52712-04-6	5.81	3.997	3.989
26	Biphenyl, 2,2′,3,4′,5,6′-hexachloro-	74472-41-6	5.39	3.997	3.989
27	Biphenyl, 2,2′,3,5,5′,6-hexachloro-	52663-63-5	5.54	4.172	4.356
28	Biphenyl, 2,2′,4,4′,5,5′-hexachloro-	35065-27-1	5.65	3.616	3.293
29	Biphenyl, 2,2′,4,4′,6,6′-hexachloro-	33979-03-2	4.93	3.813	3.634
30	Biphenyl, 2,3,3′,4,4′,5-hexachloro-	38380-08-4	5.39	4.339	4.731
31	Biphenyl, 2,3,3′,4,4′,5′-hexachloro-	69782-90-7	5.39	3.616	3.293
32	Biphenyl, 3,3′,4,4′,5,5′-hexachloro-	32774-16-6	5.97	3.997	3.989
33	Biphenyl, 2,2′,3,3′,4,5,6′-heptachloro-	38411-25-5	5.8	4.172	4.356
34	Biphenyl, 2,2′,3,4,4′,5,5′-heptachloro-	35065-29-3	5.8	4.172	4.356
35	Biphenyl, 2,2′,3,4,4′,5,6′-heptachloro-	60145-23-5	5.8	4.498	5.114
36	Biphenyl, 2,2′,3,4,4′,5′,6-heptachloro-	52663-69-1	5.84	4.339	4.731
37	Biphenyl, 2,2′,3,4′,5,5′,6-heptachloro-	52663-68-0	5.8	3.997	3.989
38	Biphenyl, 2,3,3′,4,4′,5′,6-heptachloro-	74472-50-7	5.84	4.172	4.356
39	Biphenyl, 2,2′,3,3′,4,4′,5,5′-octachloro-	35694-08-7	5.81	3.813	3.634
40	Biphenyl, 2,2′,3,3′,4,4′,5,6-octachloro-	52663-78-2	5.92	4.339	4.731
41	Biphenyl, 2,2′,3,3′,4,4′,5,6′-octachloro-	42740-50-1	5.92	3.997	3.989
42	Biphenyl, 2,2′,3,3′,4,5,5′,6-octachloro-	68194-17-2	5.88	3.616	3.370
43	Biphenyl, 2,2′,3,3′,5,5′,6,6′-octachloro-	2136-99-4	5.82	3.813	3.711
44	Biphenyl, 2,2′,3,3′,4,5,5′,6,6′-nonachloro-	52663-77-1	5.71	3.997	3.989
45	Decachlorobiphenyl	2051-24-3	5.44	4.172	4.356

The predictions were done using the model developed with ETA descriptors (model 1) and the model developed with ETA descriptors and partition coefficient (XlogP) (model 2). R^2_{pred} for model 1=0.812; R^2_{pred} for model 2=0.826

Acknowledgments The authors thank Council of Scientific and Industrial Research (CSIR), New Delhi for awarding a major research project (no. 01 (2546)/11/EMR-II) to KR and a senior research fellowship to SP.

Declaration of interest None declared

References

Ahrens A, Traas TP (2007) Environmental exposure scenarios: development, challenges and possible solutions. J Expo Sci Environ Epidemiol 17:S7–S15. doi:10.1038/sj.jes.7500602

Arnot JA, Gobas FAPC (2003) A generic QSAR for assessing the bioaccumulation potential of organic chemicals in aquatic food webs. QSAR Comb Sci 22(3):337–345. doi:10.1002/qsar.200390023

Arnot JA, Gobas FAPC (2006) A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. Environ Rev 14(4):257–297. doi:10.1139/a06-005

Arnot JA, Mackay D, Parkerton TF, Bonnell M (2008) A database of fish biotransformation rates for organic chemicals. Environ Toxicol Chem 27(11):2263–2270. doi:10.1897/08-058.1

CAESAR. http://www.caesar-project.eu. Accessed 23 Oct 2013

Cerius2 version 4.10. Cerius 2 Version 4.10 is a product of Accelrys Inc., San Diego, CA

CHEMPROP. http://www.ufz.de/index.php?en=6738. Accessed 23 Oct 2013

Dearden JC, Hewitt M (2010a) QSAR modelling of bioconcentration factor using hydrophobicity, hydrogen bonding and topological descriptors. SAR QSAR Environ Res 21(7–8):671–680. doi:10.1080/1062936X.2010.528235

Dearden JC, Hewitt M (2010b) QSAR modelling of bioconcentration factor using hydrophobicity, hydrogen bonding and topological descriptors. SAR QSAR Environ Res 21(7–8):671–680. doi:10.1080/1062936X.2010.528235

Devier MH, Augagneur S, Budzinski H, Mora P, Narbonne JF, Garrigues P (2003) Microcosm tributyltin bioaccumulation and multibiomarker assessment in the blue mussel *Mytilus edulis*. Environ Toxicol Chem 22(11):2679–2687. doi:10.1897/02-413

Dimitrov S, Breton R, MacDonald D, Walker JD, Mekenyan O (2002) Quantitative prediction of biodegradability, metabolite distribution and toxicity of stable metabolites. SAR QSAR Environ Res 13(3–4):445–455. doi:10.1080/10629360290014313

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(6):531–554. doi:10.1080/10659360500474623

Dougherty ER, Barrera J, Brun M, Kim S, Cesar RM, Chen Y, Bittner M, Trent JM (2002) Inference from clustering with application to gene-expression microarrays. J Comput Biol 9(1):105–126. doi:10.1089/10665270252833217

Eriksson L, Johansson E, Kettaneh-Wold N, Wold S (2001) Multi-and megavariate data analysis: principles and applications, 2nd edn. Umetrics Academy, Umetrics, Umea, Sweden

Everitt B, Landau S, Leese M (2001) Cluster Analysis. Arnold, London

Fernandez A, Lombardo A, Rallo R, Roncaglioni A, Giralt F, Benfenati E (2012) Quantitative consensus of bioaccumulation models for integrated testing strategies. Environ Int 45:51–58. doi:10.1016/j.envint.2012.03.004

Garrigues P (2005) Environmental chemistry: the ultimate challenge in analytical chemistry. Anal Bioanal Chem 381(1):3–4. doi:10.1007/s00216-004-2931-4

Golbraikh A, Tropsha A (2002) Beware of q^2! J Mol Graphics Modell 20(4):269–276. doi:10.1016/S1093-3263(01)00123-1

Haranczyk M, Urbaszek P, Ng GE, Puzyn P (2012) Combinatorial×computational×cheminformatics (C^3) approach to characterization of congeneric libraries of organic pollutants. J Chem Inf Model 52(11):2902–2909. doi:10.1021/ci300289b

Hartung T (2009) Toxicology for the twenty-first century. Nature 460:208–212. doi:10.1038/460208a

Hewitt M, Cronin MT, Enoch SJ, Madden JC, Roberts DW, Dearden JC (2009) In silico prediction of aqueous solubility: the solubility challenge. J Chem Inf Model 49(11):2572–2587. doi:10.1021/ci900286s

Hu C, Liu X, Li X, Zhao Y (2013) Evaluation of growth and biochemical indicators of *Salvinia natans* exposed to zinc oxide nanoparticles and zinc accumulation in plants. Environ Sci Pollut Res. doi:10.1007/s11356-013-1970-9

Johnson AR, Wichern WD (2005) Applied multivariate statistical analysis. Pearson, Delhi

Jorgensen WL (2010) Drug discovery: pulled from a protein's embrace. Nature 466:42–43. doi:10.1038/466042a

Katritzky AR, Radzvilovits M, Slavov S, Kasemets K, Tamm K, Karelson M (2010a) Quantitative structure-activity relationship modeling of bioconcentration factors of polychlorinated biphenyls. Toxicol Environ Chem 92(7):1233–1247. doi:10.1080/02772240903306417

Katritzky AR, Radzvilovits M, Slavov S, Kasemets K, Tamm K, Karelson M (2010b) Quantitative structure–activity relationship modeling of bioconcentration factors of polychlorinated biphenyls. Toxicol Environ Chem 92(7):1233–1247. doi:10.1080/02772240903306417

Kubinyi H, Hamprecht FA, Mietzner T (1998a) Three-dimensional quantitative similarity–activity relationships (3D QSiAR) from SEAL similarity matrices. J Med Chem 41(14):2553–2564. doi:10.1021/jm970732a

Liu HX, Yao XJ, Zhang RS, Liu MC, Hu ZD, Fan BT (2005) Prediction of the tissue/blood partition coefficients of organic compounds based on the molecular structure using least-squares support vector machines. J Comput Aided Mol Des 19(7):499–508. doi:10.1007/s10822-005-9003-5

Liu H, Yao X, Zhang R, Liu M, Hu Z, Fan B (2006) The accurate QSPR models to predict the bioconcentration factors of nonionic organic compounds based on the heuristic method and support vector machine. Chemosphere 63(5):722–733. doi:10.1016/j.chemosphere.2005.08.031

Lombardo A, Roncaglioni A, Boriani E, Milan C, Benfenati E (2010) Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish. Cent Eur J Chem 4(Suppl 1):S1. doi:10.1186/1752-153X-4-S1-S1

Lu X, Tao S, Hu H, Dawson RW (2000a) Estimation of bioconcentration factors of nonionic organic compounds in fish by molecular connectivity indices and polarity correction factors. Chemosphere 41(10):1675–88. doi:10.1016/S0045-6535(00)00050-3

Lu X, Tao S, Hu H, Dawson RW (2000b) Estimation of bioconcentration factors of nonionic organic compounds in fish by molecular connectivity indices and polarity correction factors. Chemosphere 41(10):1675–1688. doi:10.1016/j.bbr.2011.03.031

Meylan WM, Howard PH, Boethling RS, Aronson D, Printup H, Gouchie S (1999) Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient. Environ Toxicol Chem 18(4):664–672. doi:10.1002/etc.5620180412

Mitra I, Saha A, Roy K (2010) Chemometric modeling of free radical scavenging activity of flavone derivatives. Eur J Med Chem 45(11):5071–5079. doi:10.1016/j.ejmech.2010.08.016

OECD Document (2007) Guidance Document on the Validation of (Quantitative) 1226. http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en. Accessed 23 Oct 2013

Rogers D, Hopfinger AJ (1994) Application of genetic function approximation to quantitative structure activity relationships and

quantitative structure property relationships. J Chem Inf Comput Sci 34(4):854–866. doi:10.1021/ci00020a020

Roy K, Das RN (2011) On some novel extended topochemical atom (ETA) parameters for effective encoding of chemical information and modeling of fundamental physicochemical properties. SAR QSAR Environ Res 22(5–6):451–472. doi:10.1080/1062936X.2011.569900

Roy K, Ghosh G (2003) Introduction of Extended topochemical atom (ETA) indices in the valence electron mobile (VEM) environment as tools for QSAR/QSPR studies. Internet Electron J Mol Des 2(9): 599–620. http://biochempress.com/Files/iejmd_2003_2_0599.pdf

Roy K, Ghosh G (2004) QSTR with extended topochemical atom indices. 2. Fish toxicity of substituted benzenes. J Chem Inf Comput Sci 44(2):559–567. doi:10.1021/ci0342066

Roy K, Sanyal I, Roy PP (2006) QSPR of the bioconcentration factors of non-ionic organic compounds in fish using extended topochemical atom (ETA) indices. SAR QSAR Environ Res 17(6):563–582. doi: 10.1080/10629360601033499

Roy K, Mitra I, Kar S, Ojha PK, Das RN, Kabir H (2012) Comparative studies on some metrics for external validation of QSPR models. J Chem Inf Model 52(2):396–408. doi:10.1021/ci200520g

Roy K, Chakraborty P, Mitra I, Ojha PK, Kar S, Das RN (2013) Some case studies on application of "rm2" metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data. J Comput Chem 34(12):1071–1082. doi: 10.1002/jcc.23231

Scherb H, Voigt K (2011) Adverse genetic effects induced by chemical or physical environmental pollution. Environ Sci Pollut Res 18(5): 695–696. doi:10.1007/s11356-010-0332-0

Schuurmann G, Ebert RU, Chen J, Wang B, Kuhne R (2008) External validation and prediction employing the predictive squared correlation coefficients test set activity mean vs training set activity mean. J Chem Inf Model 48(11):2140–2145. doi:10.1021/ci800253u

Tasmin R, Shimasaki Y, Tsuyama M, Qiu X, Khalil F, Okino N, Yamada N, Fukuda S, Kang IJ, Oshima Y (2013) Elevated water temperature reduces the acute toxicity of the widely used herbicide diuron to a green alga, *Pseudokirchneriella subcapitata*. Environ Sci Pollut Res Int. doi:10.1007/s11356-013-1989-y

Toropova AP, Toropov AA, Martyanov SE, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2013) CORAL: Monte Carlo method as a tool for the prediction of the bioconcentration factor of industrial pollutants. Mol Inf 32(2):145–154. doi:10.1002/minf.201200069

Wang R, Fu Y, Lai L (1997) A new atom-additive method for calculating partition coefficients. J Chem Inf Comput Sci 37(3):615–621. doi: 10.1021/ci960169p

Wang R, Gao Y, Lai L (2000) Calculating partition coefficient by atom-additive method. Perspect Drug Discovery Des 19(1):47–66. doi:10.1023/A:1008763405023

Williams ES, Panko J, Paustenbach DJ (2009) The European Union's REACH regulation: a review of its history and requirements. Crit Rev Toxicol 39(7):553–675. doi:10.1080/10408440903036056

Wold S (1995) PLS for multivariate linear modelling. In: van de Waterbeemd H (ed) Chemometric methods in molecular design. VCH, Weinheim, pp 195–218

Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32(7): 1466–1474. doi:10.1002/jcc.21707

Zhao C, Boriani E, Chana A, Roncaglioni A, Benfenati E (2008) A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). Chemosphere 73(11):1701–1707. doi:10.1016/j.chemosphere.2008.09.033