Food and Chemical Toxicology 69 (2014) 252–259

Contents lists available at ScienceDirect

Food and Chemical Toxicology

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

Predicting the bioconcentration factor of highly hydrophobic organic chemicals

Rajni Garg [a,*], Carr J. Smith [b]

[a] Computational Science Research Center, San Diego State University, San Diego, CA 92182-7720, USA
[b] Health, Safety & Environment, Albemarle Corporation, Baton Rouge, LA 70801-1765, USA

ARTICLE INFO

Article history:
Received 30 October 2013
Accepted 26 March 2014
Available online 20 April 2014

Keywords:
QSAR
Bioconcentration factor
Risk assessment
Organic chemicals
Hydrophobicity
Octanol–water partition coefficient

ABSTRACT

Bioconcentration refers to the process of uptake and buildup of chemicals in living organisms. Experimental measurement of bioconcentration factor (BCF) is time-consuming and expensive, and is not feasible for a large number of chemicals of regulatory concern. Quantitative structure–activity relationship (QSAR) models are used for estimating BCF values to help in risk assessment of a chemical. This paper presents the results of a QSAR study conducted to address an important problem encountered in the prediction of the BCF of highly hydrophobic chemicals. A new QSAR model is derived using a dataset of diverse organic chemicals previously tested in a United States Environmental Protection Agency laboratory. It is noted that the linear relationship between the BCF and hydrophobic parameter, i.e., calculated octanol–water partition coefficient (ClogP), breaks down for highly hydrophobic chemicals. The parabolic QSAR equation, $\log BCF = 3.036\ ClogP - 0.197\ ClogP^2 - 0.808\ MgVol$ ($n = 28$, $r^2 = 0.817$, $q^2 = 0.761$, $s = 0.558$) (experimental log BCF range = 0.44–5.29, ClogP range = 3.16–11.27), suggests that a non-linear relationship between BCF and the hydrophobic parameter, along with inclusion of additional molecular size, weight and/or volume parameters, should be considered while developing a QSAR model for more reliable prediction of the BCF of highly hydrophobic chemicals.

1. Introduction

Globally, regulatory agencies are developing methods and criteria for hazard and risk assessment of chemicals (ASTM, 1993; ECETOC, 1995; ECHA, 2012; OECD, 2007). Bioaccumulation and bioconcentration refer to the process of uptake and buildup of chemicals in living organisms. The bioaccumulation factor (BAF) parameter is used as a measure of a chemical's bioaccumulation potential. If the BAF value of a chemical is not available, its BCF is used to assess the bioaccumulation potential. Experimental measurement of BAF and BCF values is time-consuming and expensive, and is not feasible for a large number of chemicals of regulatory concern. Therefore, attention is focused on estimation of these values by using quantitative structure–activity relationship (QSAR) models. QSAR models are used as screening tools to assess the effect of a large number of chemicals on the environment and human health. These models establish empirical relationships between the molecular parameters (physico-chemical properties) of the organic chemicals and physiological responses in the organism. Based on a large number of QSAR studies, it has been noted that the dataset of chemicals should exhibit a wide range in their biological activities and parameter values for developing a robust QSAR model (Hansch and Leo, 1995; Hansch et al., 1995).

The most common method for estimating BCF value consists of developing QSAR models establishing correlations between BCF and hydrophobicity of a chemical as measured by the logarithm of the octanol–water partition coefficient (denoted by $\log P$ or $\log K_{ow}$). In regulatory context, the objective is to use parameters which are easy to calculate and compare (such as $\log P$) and develop simple models which could be used to predict the most accurate BCF value (ASTM, 1993; ECETOC, 1995; ECHA, 2012; Mackay and Fraser, 2000; OECD, 2007).

Several QSAR models have been proposed for predicting the BCF of organic chemicals, which use a linear, parabolic, bilinear or polynomial relationship, extensively reviewed in Arnot and Gobas (2006), Devillers et al. (1998), Müller and Nendza (2009), and Pavan et al. (2006). Most of the QSAR models reported for the prediction of BCF within a regulatory context are based on the correlation of log BCF with $\log K_{ow}$. For a chemical, the mechanistic basis underlying the relationship of BCF with $\log K_{ow}$ is the analogy

* Corresponding author. Address: 5500 Campanile Drive, San Diego, CA 92182-7720, USA. Tel.: +1 (858) 472 9132.
E-mail address: rg004747@yahoo.com (R. Garg).

http://dx.doi.org/10.1016/j.fct.2014.03.035
0278-6915/

between the partitioning process between a biological lipid membrane and water, and the partitioning process between n-octanol and water (Arnot and Gobas, 2003, 2006; Bintein et al., 1993; Dearden, 2004; Dearden and Hewitt, 2010; Devillers et al., 1998; Dimitrov et al., 2002; Jonker and van der Haijden, 2007; Kubinyi, 1976; Müller and Nendza, 2009; Pavan et al., 2006; USEPA, 2012; Veith et al., 1979). Few QSAR models have been reported using other experimentally derived parameters such as water solubility (S), and soil adsorption coefficients (Kenaga and Goring, 1980). However, their applicability is limited due to the problem of data availability. To avoid new tests, theoretical molecular descriptors (such as topological, connectivity indices, quantum, and other descriptors) have been used for developing BCF prediction models. Assessment of their ability to correctly predict the BCF value of a chemical resulted in a large number of incorrect classifications (Pavan et al., 2006).

Arnot and Gobas (2003) proposed a mechanistic QSAR model for predicting the BCF and BAF of organic chemicals in aquatic food webs. This model uses the $\log K_{ow}$ and a number of correction factors, but does not consider a chemical's molecular weight and size-related parameters. EPI Suite software from United States Environmental Protection Agency (USEPA) (2012) uses the BCFBAF program, based on the Arnot–Gobas model, to predict the BCF and BAF values of a chemical. The BCFBAF program uses two linear QSAR equations for predicting the BCF of a chemical. The first equation with a positive linear hydrophobic term indicates that the log BCF increases linearly with $\log K_{ow}$ values for $\log K_{ow} \leqslant 7.0$, while the second equation with a negative linear hydrophobic term shows a decreasing linear relationship for values of $\log K_{ow} > 7.0$. According to this model the decrease in BCF with increasing $\log K_{ow}$ (>7.0) for highly hydrophobic chemicals is mainly due to adsorption of chemical in the water phase and not due to biomagnification or steric factors affecting membrane permeability (Arnot and Gobas, 2003). In another study, Bintein et al. (1993) reported a comparative analysis of linear, parabolic and bilinear QSAR models to explain the nonlinear dependence of fish bioconcentration on $\log P$. These models indicate that the linear relationship between log BCF and hydrophobicity is unable to explain the low BCF of highly hydrophobic chemicals. The authors (Bintein et al., 1993) concluded that the parabolic model, and preferably the bilinear model (Kubinyi, 1976), is more useful.

The European Chemical Agencies (ECHA) guidance document indicates that the log BCF increases linearly with $\log K_{ow}$ values <5 and a decreasing linear relationship is observed for higher values of $\log K_{ow}$. It is noted that apart from experimental errors in the determination of BCF values for these very hydrophobic chemicals, reduced uptake due to the increasing molecular size may also be responsible for this relationship (ECHA, 2012). Dimitrov et al. (2002) established that the relationship between log BCF and $\log K_{ow}$ for highly hydrophobic chemicals can be explained by including the molecular size parameter in the QSAR model. The ECHA guidance document also suggests that the molecular weight parameter, even though not directly related to the molecular size of a compound, together with other information can be used to assess a chemical's bioaccumulation potential (ECHA, 2012). However, no experimental data have been reported to support a specific threshold for the molecular weight parameter.

To predict the BCF values of highly hydrophobic chemicals, we have derived a new QSAR model using a dataset of diverse organic chemicals whose experimental BCF values were measured in a USEPA laboratory (Veith et al., 1979). The developed model is validated using cross validation, Tropsha's metrics, r_m^2 metrics, y-randomization test, and applicability domain analysis. This new model is discussed below and also compared with other QSAR models reported in the literature.

2. Materials and methods

2.1. Selection of dataset

Experimental log BCF values of 29 chemicals used in this study are taken from Veith et al. (1979) (see Table 1). This study on a diverse group of organic chemicals tested for bioconcentration in fathead minnow (*Pimephales promelas*) was conducted at a USEPA Environmental Research Laboratory. This is a good dataset for QSAR study as it includes a diverse group of organic chemicals including halogenated, nonhalogenated, and phosphate containing chemicals displaying a wide range in the parameter values (experimental log BCF range = 0.44–5.29, ClogP range = 3.16–11.27). Earlier models based on this dataset are used as an example for BCF prediction in the European Union Technical Guidance Document on risk assessment (Pavan et al., 2006). Out of 55 chemicals for which the BCF data were reported (Veith et al., 1979), only 30 chemicals were tested at the USEPA laboratory and the others were taken from different sources. We have used the BCF data of chemicals tested in the USEPA laboratory. One chemical 'toluene diamine', out of these 30 chemicals, is not included in our study due to uncertainty as to its structure.

2.2. Calculation of molecular parameters

The $\log P$ values listed in Table 1 are taken from Veith et al. (1979) and are provided here for comparison. They were estimated by the reverse phase HPLC method (Veith and Morris, 1978). The ClogP and MgVol parameter values are calculated and auto loaded from the C-QSAR Program (2006). The utility of the C-QSAR program in comparative correlation analysis has been discussed in Hansch and Leo (1995). Within chemical families of structural congeners, biological activity is well predicted from a chemical structure by the C-QSAR program. The parameters used in this report have been discussed in detail along with their applications in Hansch and Leo (1995). Briefly, ClogP is the calculated $\log P$ and is a measure of hydrophobicity of a chemical (Leo et al., 1971; Leo, 1993), and MgVol is the molar volume calculated by the method of (Abraham, 1993; Zhao et al., 2003). Note that the ClogP values are for the neutral form of acids and bases that may be partially ionized. If the degree of ionization is about the same for a set of congeners, the ionization factor can be neglected; otherwise, good correlation can be obtained using electronic terms (Leo et al., 1971; Leo, 1993).

The correlation matrix for the parameters used in this study is given in Table 2. The correlation between experimental $\log P$ and ClogP values for 13,815 compounds in the CLOG program, which is a part of the C-QSAR Program (2006), is 0.98 (experimental $\log P = 1.00\ \text{ClogP} - 0.03$ ($n = 13{,}815$, $r = 0.98$, $s = 0.35$). Many programs are used for calculating octanol–water partition coefficients and are reviewed in Mannhold et al. (2009). However, we have used the ClogP parameter in this study as it has been widely used and cited by the QSAR community, both for environmental studies and drug design (Arnot and Gobas, 2006; Devillers et al., 1998; Garg et al., 1999; Hansch et al., 1989; Leo and Hansch, 1999; Müller and Nendza, 2009; Selassie et al., 2003; Smith et al., 2002, 2003, 2004, 2006), and a very high correlation ($r = 0.98$) between experimental $\log P$ and ClogP gives confidence in using ClogP values whenever experimental $\log P$ values are not available.

The QSAR stepwise multiple linear regression (MLR) analyses are executed with the C-QSAR program and all the parameters are auto loaded (C-QSAR, 2006). In all the QSAR equations reported in this report, n is the number of data points, r is the correlation coefficient, s is the standard deviation, and q^2 is the quality of fit of the data, calculated using Cramer et al.'s (1988) approach, which approaches the value of r^2 as the quality of fit improves.

3. Results and discussion

First, we developed a QSAR model for the whole dataset using stepwise MLR analysis. Next, we divided the whole dataset into a training set and a test set and performed internal and external validation studies. Cross validation techniques were utilized for internal validation, and the model developed using training set was used to predict the activity of test set chemicals. Tropsha's and r_m^2 metrics were also calculated to evaluate the internal and external predictive abilities of the QSAR model. To ensure the developed QSAR model is robust and not derived due to chance, the y-randomization test was performed. Lastly, the applicability domain of the developed QSAR model was evaluated to ascertain the reliability of the model.

3.1. Model development

Stepwise MLR analysis on whole dataset reported by Veith et al. (1979) (Table 1) resulted in Eqs. (1)–(3).

Table 1
Experimental and predicted BCF and molecular parameter values used in this study.

No.	Chemical	CAS no.	Expt. log BCF[a]	Pred. log BCF[b] (Eq. (1))	Pred. log BCF[b] (Eq. (3))	Pred. log BCF[b] (Eq. (5))	Exp. log*P*	ClogP	log*P*	MgVol
1	Heptachlor	76-44-8	3.98	3.40	3.90	3.87	5.58	5.45	5.44	1.96
2	Heptachlor epoxide	1024-57-3	4.16	3.05	2.74	2.73	4.98	4.39	5.40	1.96
3	p,p′ – DDE[d]	72-55-9	4.71	3.82	4.65	**4.61**	6.96	6.74	5.69	2.05
4	Pentachlorophenol	87-86-5	2.89	3.16	3.61	3.65	5.12	4.71	5.01	1.39
5	Hexabromobiphenyl	59261-08-4	4.26	4.24	4.56	4.51	NA	8.01	6.39	2.37
6	Methoxychlor[d]	72-43-5	3.92	3.31	3.31	**3.25**	5.08	5.18	4.30	2.37
7	Mirex	2385-85-5	4.26	3.48	3.72	3.64	7.13	5.69	6.89	2.44
8	Hexabromocyclododecane	3194-55-6	4.26	4.22	4.27	4.18	NA	7.95	5.81	2.74
9	Hexachlorocyclopentadiene[c]	77-47-4	1.47	3.27	4.00	4.04	5.04	5.04	5.51	1.35
10	Heptachloronorbornene	5202-36-8	4.05	3.19	3.48	3.49	NA	4.82	5.28	1.69
11	Hexachloronorbornadiene	3389-71-7	3.81	3.06	3.13	3.17	NA	4.42	5.28	1.53
12	Aroclor-1016	12674-11-2	4.63	3.55	4.50	4.49	5.62	5.92	5.88	1.69
13	Aroclor-1248	12672-29-6	4.85	3.75	4.75	4.73	NA	6.51	6.11	1.81
14	Aroclor-1254[d]	11097-69-1	5.00	3.94	4.86	**4.83**	NA	7.11	6.47	1.94
15	Aroclor-1260	11096-82-5	5.29	4.20	4.73	4.69	NA	7.90	6.91	2.18
16	Chlordane	12789-03-6	4.58	3.69	4.40	4.35	6.22	6.32	6.00	2.13
17	Octachlorostyrene	29082-74-4	4.52	4.00	4.90	4.87	NA	7.28	6.29	1.93
18	p,p′ – DDT	50-29-3	4.47	3.83	4.53	4.47	6.91	6.76	5.75	2.22
19	o,p′ – DDT	789-02-6	4.57	3.83	4.53	4.47	NA	6.76	5.75	2.22
20	Hexachlorobenzene	118-74-1	4.27	3.60	4.79	4.80	5.73	6.06	5.23	1.45
21	1,2,4-Trichlorobenzene	120-82-1	3.32	2.98	3.14	3.23	4.05	4.16	4.23	1.08
22	Lindane	58-89-9	2.26	2.84	2.13	2.19	3.72	3.75	3.89	1.58
23	5-Bromoindole	10075-50-0	1.15	2.65	1.51	1.65	3.00	3.16	2.97	1.12
24	2,4,6-Tribromoanisol[d]	607-99-8	2.94	3.05	3.17	3.21	4.48	4.40	4.48	1.44
25	N-phenyl-2-napthylamine	135-88-6	2.17	3.14	3.19	3.20	4.38	4.64	4.38	1.79
26	Tris(2,3-dibromopropyl) phosphate	126-72-7	0.44	2.74	0.59	0.53	3.71	3.44	4.98	2.87
27	Tricresyl-phosphate	1330-78-5	2.22	3.56	3.63	3.51	NA	5.95	3.42	2.79
28	Chlorinated-ecosane	112-95-8	1.69	5.31	1.64	1.70	NA	11.27	7.05	2.93
29	Diphenylamine[d]	122-39-4	1.48	2.75	1.80	**1.89**	3.50	3.47	3.42	1.42

NA = data not available.
[a] Experimental log BCF values are taken from Veith et al. (1979).
[b] Predicted log BCF values are calculated by QSAR Eqs. (1), (3) and (5).
[c] Not included in deriving QSAR Eqs. (3) and (5).
[d] Test compounds (#3, 6,14, 24, and 29) are not included in deriving Eqs. (4) and (5). Values in bold are predicted for these compounds using Eq. (5).

Table 2
Correlation matrix (r^2 and n) for the molecular parameters used in this study.

	Exp. log*P*	log*P*	ClogP	MgVol	MW
Exp. log*P*		0.789	0.891	0.185	0.073
log*P*	19		0.57	0.18	0.17
ClogP	19	29		0.348	0.065
MgVol	19	29	29		0.507
MW	19	29	29	29	

$$\text{Log BCF} = 0.328\ \text{ClogP} + 1.610$$
$$n = 29,\ r^2 = 0.187,\ q^2 = -0.324,\ s = 1.226 \quad (1)$$

$$\text{Log BCF} = 2.939\ \text{ClogP} - 0.200\ \text{ClogP}^2 - 6.134$$
$$n = 28,\ r^2 = 0.753,\ q^2 = 0.701,\ s = 0.672 \quad (2)$$
$$\text{Optimum ClogP } (\log P_0) = 7.370$$

$$\text{Log BCF} = 3.036\ \text{ClogP} - 0.197\ \text{ClogP}^2 - 0.808\ \text{MgVol} - 5.213$$
$$n = 28,\ r^2 = 0.817,\ q^2 = 0.761,\ s = 0.558 \quad (3)$$
$$\text{Optimum ClogP } (\log P_0) = 7.716$$

From these equations, it is obvious that the parabolic QSAR model given by quadratic Eq. (3) is the better regression model for this dataset (experimental log BCF range = 0.44–5.29, ClogP range = 3.16–11.27). The difference between r^2 and q^2 in Eq. (3) is 0.056, indicating the model has low overfitting effects. The linear model given by Eq. (1) is not able to explain the behavior of highly hydrophobic chemicals (Fig. 1). Note that although the ClogP based parabolic model given by Eq. (2) is significantly better than the linear model given by Eq. (1), addition of a volume parameter in Eq. (2) improves the quality of fit and standard deviation of the model as shown by Eq. (3). One chemical, hexachlorocyclopentadiene, was omitted in deriving Eq. (3) as these data were anomalous to the relationship between the log*P* and the log BCF shown in Fig. 1 (log BCF = 1.47, ClogP = 5.04). Veith et al. (1979) also noted this strange behavior of hexachlorocyclopentadiene and omitted it from their correlation analysis. Interestingly, two other chemicals, tris (2,3-dibromopropyl) phosphate and chlorinated ecosane, omitted from the Veith et al. (1979) analysis were predicted well by Eq. (3) (Table 1). We also assessed the bilinear relationship of hydrophobicity combined with MgVol and found that the predictability of bilinear model is lower ($n = 28$, $r^2 = 0.783$, $q^2 = 0.694$, $s = 0.655$).

Eq. (3) indicates that 7% of the variance in the data can be explained by including a volume related parameter. Correlation matrix in Table 2 shows that there is no mutual correlation between ClogP and MgVol parameter ($r^2 = 0.348$) eliminating the possibility of chance correlation. We performed an *F*-test to ensure the statistical significance of McGowan parameter in Eq. (3), and found $F_{3,\ 34} = 9.36$ (4.72). Here, F is the Fischer ratio between the variances of calculated and observed activities, and the value within the parenthesis is the standard *F*-value at 99% level (Lomax and Hahs-Vaughn, 2013). As the *F* ratio is greater than the standard *F*-value (9.36 > 4.72), it can be concluded that the McGowan volume parameter is significant and should be included in the equation used for predicting the BCF of the chemicals in this dataset.

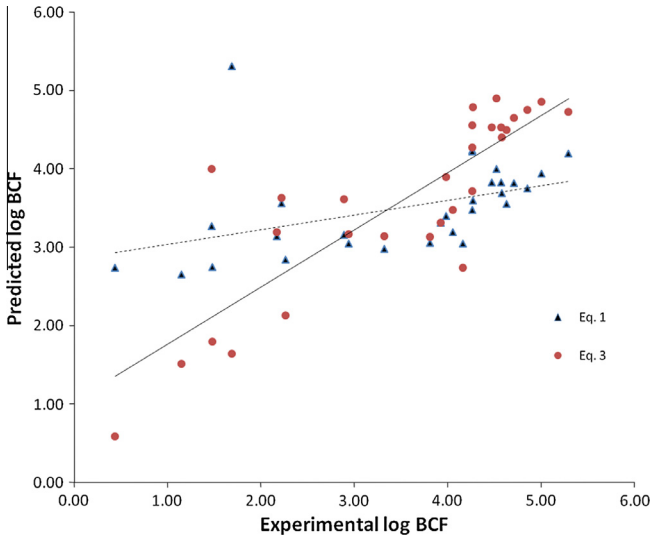

Fig. 1. Correlation of the experimental log BCF with predicted log BCF.

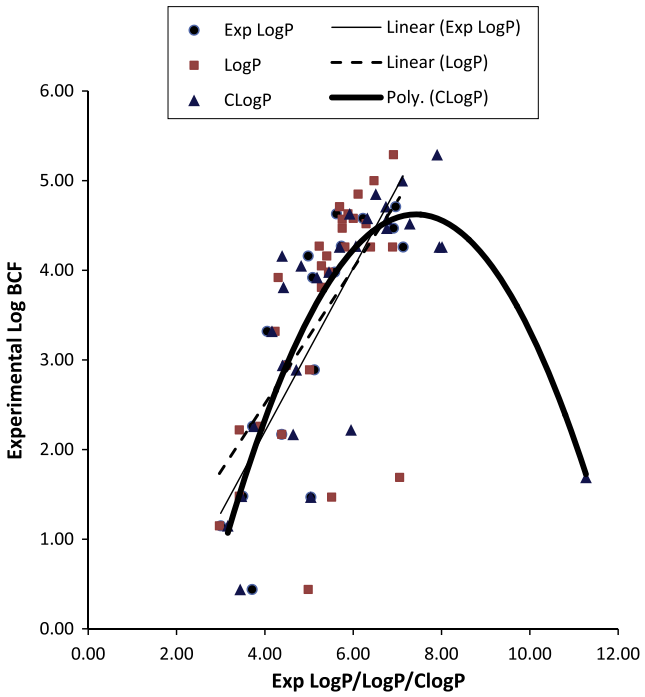

Fig. 2. Correlation of the experimental log BCF values with experimental log P, log P and ClogP.

Eq. (3) shows a parabolic dependence of bioconcentration on hydrophobicity indicating that the BCF value of a chemical first increases with increasing hydrophobicity up to an optimum value ($\log P_0 = 7.716$) and then decreases with increasing hydrophobicity. The presence of a negative MgVol term in Eq. (3) demonstrates that BCF decreases with increase in the size of a chemical. The MgVol parameter has been used in other QSAR studies to explain the correlation between biological activity and the size of a chemical (Hansch and Leo, 1995).

Our results show that the parabolic model (Eq. (3)) based on hydrophobic and molecular volume parameters is better able to explain the low log BCF values of highly hydrophobic chemicals in this dataset than the linear model (Eq. (1)) (Fig. 2). Eq. (3) underlines the presence of an optimum log P value in the range of 7–8. This is in agreement with optimum values reported in other QSAR studies related to BCF prediction (Arnot and Gobas, 2006; Dearden, 2004; Devillers et al., 1998; Müller and Nendza, 2009; Pavan et al., 2006). Most hydrophobic chemicals in this dataset including chlorinated ecosane (ClogP = 11.27, experimental log BCF = 1.69, pred. log BCF (Eq. (1)) = 5.31, pred. log BCF (Eq. (3)) = 1.64), are well predicted by the parabolic model in Eq. (3) (Table 1). Differences between experimental and predicted log BCF values (Table 1) for some of the chemicals (e.g., heptachlor epoxide, pentachlorophenol, N-phenyl-2-naphthylamine and Tricresyl-phosphate) could be due to their reactivity, acid/base properties or other factors. However, these factors are well considered in the calculation of ClogP parameter value (Leo, 1993). Note that ClogP values are calculated using the fragment constant method which employs a number of fragment constants based on a unique and simple set of rules and many correction factors to account for proximity effects due to multiple halogenation, hydrogen bonds, intra-molecular hydrogen-bonds involving oxygen and nitrogen atoms, electronic effects in aromatic systems, unsaturation, branching, chains, and rings (Hansch et al., 1995; Leo, 1993).

It is noteworthy that the parabolic models (Eqs. (2) and (3)) derived by us are similar to the parabolic model derived for 154 organic chemicals by Bintein et al. (1993). We did not observe a bilinear relationship for the dataset in Table 1. It could be due to a low number of chemicals in our dataset (29 vs. 154 in (Bintein et al., 1993)). Presence of a molecular volume related parameter in Eq. (3) is also supported by an earlier QSAR study by Dimitrov et al. (2002).

3.2. Model validation

To measure the predictive ability of the QSAR model, we performed internal and external validation studies as recommended by several researchers (Gramatica et al., 2013; Golbraikh and Tropsha, 2002; Golbraikh et al., 2003; Kiralj and Ferreira, 2009; Mitra et al., 2011; Roy and Mitra, 2011; Tropsha et al., 2003, Tropsha, 2010). Internal validation uses a training set to evaluate the predictivity of a model, whereas external validation uses a test set which is not included in building the training set model.

We divided the whole dataset into two sets, a training set comprised of 24 chemicals and a test set comprised of 5 chemicals (listed in Table 1), ensuring that the total range of experimental log BCF values is adequately tested (Golbraikh et al., 2003; Kennard and Stone, 1969). QSAR analysis on training set data gave following Eqs. (4) and (5).

$$\begin{aligned}&\text{Log BCF}=2.874\ \text{ClogP}-0.186\ \text{ClogP}^2-0.630\ \text{MgVol}-5.138\\&n=24,\ r^2=0.673,\ q^2=0.538,\ s=0.825\\&\text{Optimum ClogP}\ (\log P_0)=7.731\end{aligned}\tag{4}$$

$$\begin{aligned}&\text{Log BCF}=2.936\ \text{ClogP}-0.189\ \text{ClogP}^2-0.907\ \text{MgVol}-4.733\\&n=23,\ r^2=0.793,\ q^2=0.724,\ s=0.636\\&\text{Optimum ClogP}\ (\log P_0)=7.769\end{aligned}\tag{5}$$

Similar to Eq. (3) derived for the whole dataset, hexachlorocylcopentadiene was found as a misfit in deriving Eq. (4) and its removal from the training set improved the quality of fit and standard deviation of the model as shown by Eq. (5). In Eq. (5), the difference between r^2 and q^2 is 0.069, indicating the model has low overfitting effects. Note that Eq. (5) developed using the training set data is similar to Eq. (3) developed using the whole dataset.

Internal validation was done using cross validation techniques. First, we performed leave-one-out cross validation (LOO-CV), where one compound of the training set is removed and the model

is retrained using the remaining compounds. This process is repeated over all the samples in the training set. The model predictivity is measured by calculating the cross validated coefficient (q^2_{LOO}) as described in Roy and Mitra (2011), and Tropsha et al. (2003), and a model is considered satisfactory if $q^2_{LOO} > 0.5$. We developed a QSAR model using LOO-CV method and obtained a satisfactory predictive model, similar to Eq. (5), with $q^2_{LOO} = 0.53$.

Next, we performed leave-many-out cross validation (LMO-CV) studies as it has been suggested that LOO-CV alone may not be enough to assess the predictive power of a model (Tropsha et al., 2003; Roy and Mitra, 2011). LMO-CV is similar to LOO-CV except that in LMO-CV a definite number of training set data points are removed from model building in each cycle. After all the cycles are completed, the predicted activity values are used to calculate q^2_{LMO}. We obtained highly significant models with values of q^2_{LMO} varying from 0.52 to 0.57 after leaving out 2, 3 and 4 data points during cross validation.

For comparison, we also performed LOO-CV and LMO-CV analysis on the whole dataset and found that q^2_{LOO} and q^2_{LMO} of the final model varies from 0.757–0.761 after leaving out 1, 2, 3 and 4 data points during cross validation. These results indicate that for the dataset in Table 1, model developed using whole dataset (Eq. (3)) could be used for predicting the BCF of new compounds not included in developing the model.

External validation was done using the test set data to determine the predictive ability of the model for the compounds not included in the training set (Table 1). Predicted r^2_{pred} (also known as q^2_{ext}) calculated using the equation described in Tropsha et al. (2003) was found to be 0.925, and all the test set chemicals were very well predicted. Note that model with $r^2_{pred} > 0.5$ is considered to be valid for prediction.

Tropsha's metrics are used for analyzing the external predictability of the model and widely used to validate QSAR models (Roy and Mitra, 2011). We calculated k and k', slopes of the regression line of the predicted activity vs. experimental activity and *vice versa*, and r^2_0 and r'^2_0 correlation coefficient of regression between the predicted and experimental activity of the compounds in the test set and *vice versa* without using y-intercept, as described in Golbraikh and Tropsha (2002). A model is considered acceptable if the following conditions are satisfied: $r^2 > 0.6$, $0.85 \leqslant k \leqslant 1.15$ or $0.85 \leqslant k' \leqslant 1.15$, $(r^2 - r^2_0/r^2)$ or $(r'^2 - r^2_0/r^2) < 0.1$ (Golbraikh and Tropsha, 2002). Analysis of test set data show that all the values are within the specified range: $r^2 = 0.962$, $k = 1.06$, $k' = 0.94$, $(r^2 - r^2_0/r^2) = 0.007$, and $(r'^2 - r^2_0/r^2) = 0.01$. These values were calculated using DTC cheminformatics tools (http://dtclab.webs.com/software-tools). Although it is sometimes possible to obtain a high cross validated q^2 value due to many reasons, but only few of them are really found highly predictive when judged by these validation metrics. Thus, these results further validate the model developed in this study.

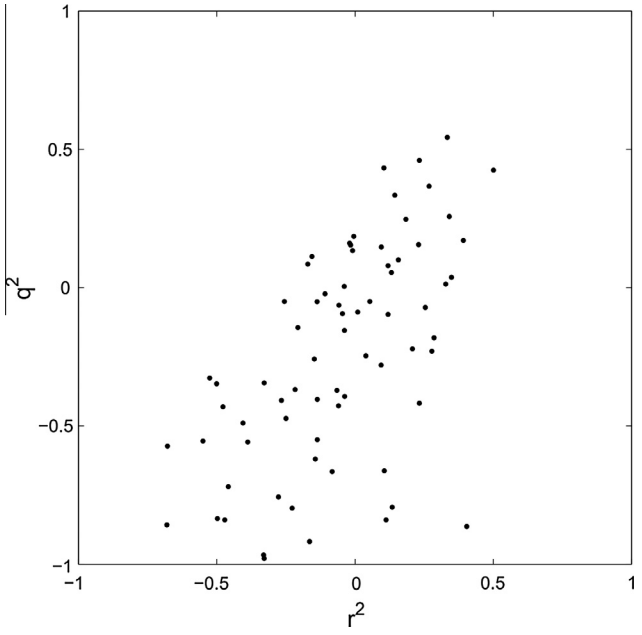

Fig. 3. y-Randomization plot of QSAR model.

r^2_m metrics (average $r_m{}^2$ and delta $r_m{}^2$) were developed to evaluate the internal and external predictive abilities of the QSAR model (Roy et al., 2012). It is suggested that for a QSAR model to be acceptable, the value of "average $r_m{}^2$" should be >0.5 and "delta $r_m{}^2$" should be <0.2 (Mitra et al., 2011, Roy et al., 2012). We used RmSquare Calculator (http://aptsoftware.co.in/rmsquare/) to calculate these metrics for the training, test, and whole dataset. Our results indicated that although delta $r_m{}^2$ values for training and whole set are slightly higher than recommended, all other values are within the specified range: average $r_m{}^2$(LOO) = 0.54 and delta $r_m{}^2$(LOO) = 0.30; average $r_m{}^2$(test) = 0.86 and delta $r_m{}^2$(test) = 0.04; average $r_m{}^2$ (overall = 0.60 and delta $r_m{}^2$ (overall) = 0.29.

3.3. y-Randomization test

To ensure the developed QSAR model is robust and not derived due to chance, the y-randomization test was performed on the training set data as recommended (Roy and Mitra, 2011; Tropsha et al., 2003). In this test, MLR models are generated by randomly scrambling the dependent variable (activity data) while keeping the independent variable (descriptors) unchanged. The resulting models are expected to have significantly low r^2 and cross validated q^2 values for several trials, which confirm that the developed models are robust. We performed 100-y-randomization tests and observed that for all the models except one, the values of r^2 and q^2_{LOO} were <0.5 (Fig. 3). This test confirms that the developed model is robust and not derived merely due to chance.

3.4. Evaluation of the applicability domain of the model

Evaluation of the applicability domain of the QSAR model is considered an important step to establish that the model is reliable to make predictions within the chemical space for which it was developed (Eriksson et al., 2003; Roy and Mitra 2011; Tropsha et al., 2003; Tropsha 2010; Tropsha and Golbraikh, 2007). There are several methods for defining the applicability domain of a QSAR model, but we used the most commonly used leverage approach in this study (Gramatica, 2007). Leverage of a given chemical compound h_i is defined as: $h_i = x_i^T(X^T X)^{-1}x_i$, where x_i is the descriptor row of the query compound and X is the descriptor matrix of the training set compounds used to develop the model. As a prediction tool, the warning leverage h^* is defined as: $h^* = 3(p+1)/n$, where n is the number of training compounds, and p is the number of descriptors in the model. The test compounds with leverages $h_i < h^*$ are considered to be reliably predicted by the model. The Williams plot, a plot of standardized residuals vs. leverage values, is used to interpret the applicability domain of the model. The domain of reliable prediction for external test set compounds is defined as compounds which have leverage values within the threshold ($h_i < h^*$) and standardized residuals no greater than 3 units ($\pm\sigma$). Test set compounds where ($h_i > h^*$) are considered to be unreliably predicted by the model due to substantial extrapolation. For the training set, the Williams plot is used to identify compounds with the greatest structural influence ($h_i > h^*$) in developing the model.

The Williams plot for the training set shown in Fig. 4, establishes applicability domain of the model within $\pm 3\sigma$ and a leverage

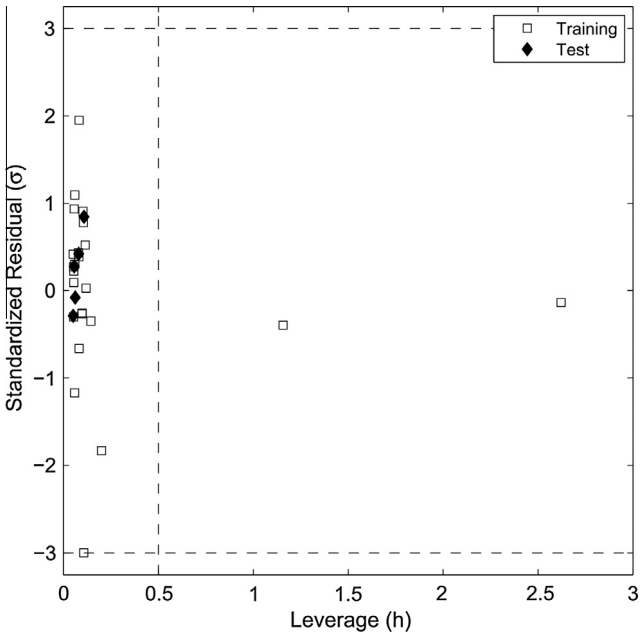

Fig. 4. Williams plot to evaluate the applicability domain of QSAR model.

threshold h^* = 0.5. It is clear from Fig. 4 that all the compounds in the dataset are within the applicability domain of the model except two training set compounds (#26 and #28). Both of these compounds have their leverage values greater than the warning h^* value and could be high leverage compounds influencing the performance of the model. However, their standard residual values are very low and within the established limit. As a result, these two compounds could be considered as influential in fitting the model performance but not necessarily outliers to be deleted from the training dataset, and thus the model can be applied with confidence within the defined applicability domain.

3.5. Effect of highly hydrophobic compounds on the model

Fig. 2 shows that one ClogP value (11.27) (compound #28 in Table 1) in the dataset is far from the other chemicals. We removed this compound from the whole dataset and derived Eq. (6) that shows this chemical did not force Eqs. (2) and (3) to assume a parabolic shape. Note that the optimum value of ClogP in Eq. (6) is similar to the one observed in Eq. (3). Hexachlorocyclopentadiene was a misfit here also and omitted in deriving Eq. (6).

$$\begin{aligned}&\text{Log BCF} = 3.183\ \text{ClogP} - 0.210\ \text{ClogP}^2 - 0.805\ \text{MgVol} - 5.592\\&n = 27,\ r^2 = 0.802,\ q^2 = 0.729,\ s = 0.600\\&\text{Optimum ClogP } (\log P_0) = 7.570\end{aligned} \tag{6}$$

In addition, Fig. 2 shows that three compounds (#5, 8, and 15 in Table 1) with >7.7 ClogP values are clustered together. To study the effect of these highly hydrophobic chemicals on the parabolic model, we removed all four (#5, 8, 15 and 28 in Table 1) from the whole dataset and derived Eqs. (7) and (8).

$$\begin{aligned}&\text{Log BCF} = 0.845\ \text{ClogP} - 1.022\\&n = 25,\ r^2 = 0.613,\ q^2 = 0.552,\ s = 0.843\end{aligned} \tag{7}$$

$$\begin{aligned}&\text{Log BCF} = 2.698\ \text{ClogP} - 0.176\ \text{ClogP}^2 - 5.639\\&n = 25,\ r^2 = 0.644,\ q^2 = 0.568,\ s = 0.825\\&\text{Optimum ClogP } (\log P_0) = 7.656\end{aligned} \tag{8}$$

Looking at these two equations, it is evident that when highly hydrophobic chemicals are removed from the dataset, the parabolic model (Eq. (8)) is not superior to the linear model (Eq. (7)). Further addition of the MgVol parameter did not improve Eq. (8). This clearly shows that QSAR modeling is unable to reveal the parabolic relationship between log BCF and hydrophobicity of highly hydrophobic chemicals if the dataset under study does not have sufficient spread in the hydrophobic parameter value.

3.6. Comparison of log*P* and ClogP parameters

We developed a QSAR model for the whole dataset, similar to Eq. (3), using log*P* values measured by Veith and Morris (1978) listed in Table 1, to compare two hydrophobic parameters, ClogP and log*P*. It was noticed that the parabolic hydrophobic term and MgVol parameter are no longer significant in the model developed using log*P* values, and the model has low predictability as evident by its low q^2 and high s values (n = 28, r^2 = 0.561, q^2 = 0.194, s = 0.913). Note that log*P* values were measured by using reverse phase HPLC method (Veith and Morris, 1978), and it is possible that reverse phase HPLC method may not be effective in the measurements of partition coefficients >7.0 for highly hydrophobic chemicals. Looking at the range (low–high) of hydrophobic parameter values in Table 1, (experimental log*P* = 3.00–7.13, log*P* = 2.97–7.05, and ClogP = 3.16–11.27), shows that experimental log*P* and log*P* values have insufficient spread in the parameter values to reveal the optimum value of log*P* (7.716), whereas 21% of the chemicals (6 out of 29) have ClogP > 7.0 (Table 1). Because of the wide spread in ClogP parameter values, optimum value is revealed in Eq. (3) and a parabolic relationship between log BCF and ClogP value of highly hydrophobic chemicals in the dataset under study could be observed. As mentioned earlier in Section 2.2, ClogP has been widely used in QSAR studies, and a very high correlation (r = 0.98) between experimental log*P* and ClogP gives confidence in using ClogP values whenever experimental log*P* values are not available. It would be interesting to compare the range of hydrophobic parameter values in a few larger datasets used for deriving BCF QSAR models.

4. Conclusion

In this study we developed QSAR models for the whole dataset, training set, and test set. Cross validation (LOO and LMO), Tropsha's metrics, and r_m^2 metrics validate the internal and external predictabilities of the model developed using training and test set. Results of y-randomization test ensure that the developed QSAR model is robust and not derived merely due to chance. Evaluation of the applicability domain establishes that the developed model is reliable to make predictions within the chemical space for which it is developed.

We found that Eq. (3) developed using the whole dataset is similar to Eq. (5) developed using the training set which is validated by cross validation, y-randomization and applicability domain analysis. This suggests that for the dataset in Table 1, the model developed using whole dataset (Eq. (3)) could be used for prediction of the new compounds not included in developing the model. Our results are supported by observations made in other studies (Mitra et al., 2011; Tropsha, 2010) where it is suggested that if a dataset is small (like the dataset used in this study), one should use the whole dataset for model development as dividing it into training and test sets could result in loss of an appreciable amount of chemical information.

It is demonstrated that a parabolic QSAR model is better able to explain the low log BCF values of highly hydrophobic chemicals in comparison to a linear model. The parabolic equations formulated

in this study underline the presence of an optimum log P value in the range of 7–8. This is in agreement with values reported in other QSAR studies for predicting BCF. Our results suggest that BCF prediction using a linear QSAR equation may introduce error into the prediction of highly hydrophobic chemicals. Linear models are unable to explain the parabolic scatter observed between log BCF and hydrophobic parameter for highly hydrophobic chemicals. To summarize, if the dataset under study has enough data points and the hydrophobic parameter values vary over a wide range, a non-linear relationship between BCF and the hydrophobic parameter, along with inclusion of additional molecular size, weight and/or volume related parameters, should be considered while developing a QSAR model for more reliable prediction of the BCF of highly hydrophobic chemicals.

Conflict of Interest

The authors declare that there are no conflicts of interest.

Transparency Document

The Transparency document associated with this article can be found in the online version.

Acknowledgements

This work was supported by a contract to Rajni Garg by American Chemistry Council (Contract number 5466). Authors are grateful to Gene Ko and Srinivas Reddy for their help in conducting validation studies.

References

Abraham, M.H., 1993. Scales of solute hydrogen-bonding – their construction and application to physicochemical and biochemical processes. Chem. Soc. Rev. 22, 73–83.

Arnot, J.A., Gobas, F.A.P.C., 2003. A generic QSAR for assessing the bioaccumulation potential of organic chemicals in aquatic food webs. QSAR Comb. Sci. 22, 337–345.

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

ASTM (American Society for Testing and Materials), 1993. ASTM Standards on Aquatic Toxicology and Hazard Evaluation. ASTM Committee E-47 on Biological Effects and Environmental Fate. ASTM Publication Code Number (PCN): 03-547093-16, p. 538.

Bintein, S., Devillers, J., Karcher, W., 1993. Non-linear dependence of fish bioconcentration on n-octanol/water partition coefficients. SAR QSAR Environ. Res. 1, 29–39.

C-QSAR Program, 2006. BioByte Corp. 201 W. 4th St. Suite 204, Claremont, CA 91711. (<http://biobyte.com/bb/prod/cqsarad.html>).

Cramer, R.D., Bunce, J.D., Patterson, D.E., Frank, I.E., 1988. Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. Quant. Struct.-Act. Relat. 7, 18–25.

Dearden, J.C., 2004. Improved prediction of fish bioconcentration factor of hydrophobic chemicals. SAR QSAR Environ. Res. 15, 449–455.

Dearden, J.C., Hewitt, M., 2010. QSAR modeling of bioconcentration factor using hydrophobicity, hydrogen bonding and topological descriptor. SAR QSAR Environ. Res. 21, 671–680.

Devillers, J., Domine, D., Bintein, S., Karcher, W., 1998. Comparison of fish bioconcentration models. In: Devillers, J. (Ed.), Comparative QSAR. Taylor and Francis, Washington, DC, pp. 1–50.

Dimitrov, S.D., Dimitrova, N.C., Walker, J.D., Veith, G.D., Mekenyan, O.G., 2002. Predicting bioconcentration factors of highly hydrophobic chemicals. Effects of molecular size. Pure Appl. Chem. 74, 1823–1830.

DTC cheminformatics tools. (<http://dtclab.webs.com/software-tools>).

ECETOC (European Centre for Ecotoxicology and Toxicology of Chemicals), 1995. The role of bioaccumulation in environmental risk assessment: the aquatic environment and related food webs. In: Technical Report 67, Brussels, Belgium.

ECHA (European Chemical Agencies), 2012. Guidance on Information Requirements and Chemical Safety Assessment, Chapter R.11: PBT Assessment, (Version 1.1). (<http://www.echa.europa.eu/web/guest/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment>).

Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M., Gramatica, P., 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs. Environ. Health Perspect. 111, 1361–1375.

Garg, R., Gupta, S.P., Gao, H., Babu, M.S., Debnath, A.K., Hansch, C., 1999. Comparative quantitative structure–activity relationship studies on anti-HIV drugs. Chem. Rev. 99, 3525–3602.

Golbraikh, A., Tropsha, A., 2002. Beware of q2! J. Mol. Graph. Model. 20, 269–276.

Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.D., Lee, K.H., Tropsha, A., 2003. Rational selection of training and test set for the development of validated QSAR models. J. Comp.-Aided Mol. Des. 17, 241–253.

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

Gramatica, P., Chirico, N., Papa, E., Cassani, S., Kovarich, S., 2013. QSARINS: new software for the development, analysis, and validation of QSAR MLR models. J. Comp. Chem. 34, 2121–2132.

Hansch, C., Leo, A., 1995. Exploring QSAR: Fundamentals and Applications in Chemistry and Biology. In: Heller, S.R. (Ed.). American Chemical Society Professional Reference Book, Washington, DC.

Hansch, C., Kim, D., Leo, A.J., Novellino, E., Silipo, C., Vittoria, A., 1989. Toward a quantitative comparative toxicology of organic compounds. CRC Crit. Rev. Toxicol. 19, 185–226.

Hansch, C., Leo, A., Hoekman, D., 1995. Exploring QSAR: Hydrophobic, Electronic, and Steric Constants. In: Heller, S.R. (Ed.). American Chemical Society Professional Reference Book, Washington, DC.

Jonker, M., van der Haijden, S., 2007. Bioconcentration factor hydrophobicity cut-off: An artificial phenomenon reconstructed. Environ. Sci. Technol. 41, 7363–7369.

Kenaga, E.E., Goring, C.A.I., 1980. Relationship Between Water Solubility and Soil Sorption, Octanol–Water Partitioning and Bioconcentration of Chemicals in Biota. In: Eaton, J.G., Parrish, P.R.P., Hendricks, A.C. (Eds.). Special Technical Publication 707, American Society for Testing and Materials, Philadelphia, PA, pp. 78–115.

Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11, 137–148.

Kiralj, R., Ferreira, M.M.C., 2009. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. J. Braz. Chem. Soc. 20, 770–787.

Kubinyi, H., 1976. Quantitative structure–activity relationships. IV. Non-linear dependence of biological activity on hydrophobic character: a new model. Arzneim-Forschung 26, 1991–1997.

Leo, A., 1993. Calculating log Poct from structures. Chem. Rev. 93, 1281–1306.

Leo, A., Hansch, C., 1999. Role of hydrophobic effects in mechanistic QSAR. Pers. Drug Discov. Des. 17, 1–25.

Leo, A., Hansch, C., Elkins, D., 1971. Partition coefficients and their use. Chem. Rev. 71, 525–616.

Lomax, R.G., Hahs-Vaughn, D.L., 2013. An Introduction to Statistical Concepts, third ed. Taylor and Francis.

Mackay, D., Fraser, A., 2000. Bioaccumulation of persistent organic chemicals: mechanisms and models. Environ. Pollut. 110, 375–391.

Mannhold, R., Poda, G.I., Ostermann, C., Tetko, I.V., 2009. Calculation of molecular lipophilicity: state-of-the-art and comparison of log P methods on more than 96,000 compounds. J. Pharm. Sci. 98, 861–893.

Mitra, I., Saha, A., Roy, K., 2011. Chemometric QSAR modeling and in silico design of antioxidant NO donor phenols. Sci. Pharm. 79, 31–57.

Müller, M., Nendza, M., 2009. Literature study: comparative analysis of estimated and measured BCF data (OECD 305) with a special focus on differential accumulation of (mixtures of) stereoisomers. (<http://www.uba.de/uba-info-medien-e/4088.html>).

Organization for Economic Co-operation and Development, 2007. Guidance document on the validation of (quantitative) structure–activity relationships [(Q)SAR] models. OECD Document ENV/JM/MONO 2.

Pavan, M., Worth, A.P., Netzeva, T.I., 2006. Review of QSAR models for bioconcentration. European Commission Publication Code Number EUR 22327 EN.

RmSquare Calculator. (<http://www.aptsoftware.co.in/rmsquare/>).

Roy, K., Mitra, I., 2011. On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. Comb. Chem. High. Throughput Screen. 14, 450–474.

Roy, K., Mitra, I., Kar, S., Ojha, P., Das, R.N., Kabir, H., 2012. Comparative studies on some metrics for external validation of QSPR models. J. Chem. Inf. Model. 52, 396–408.

Selassie, C.D., Garg, R., Mekapati, S.B., 2003. A mechanism-based approach to the study of the toxicity of endocrine disruptive agents. Pure Appl. Chem. 75, 2363–2374.

Smith, C.J., Perfetti, T.A., Morton, M.J., Rodgman, A., Garg, R., Selassie, C.D., Hansch, C., 2002. The relative toxicity of substituted phenols reported in cigarette mainstream smoke. Toxicol. Sci. 69, 265–278.

Smith, C.J., Perfetti, Garg, R., Hansch, C., 2003. IARC carcinogens reported in cigarette mainstream smoke and their calculated log P values. Food Chem. Toxicol. 41, 807–817.

Smith, C.J., Perfetti, T.A., Garg, R., Martin, P., Hansch, C., 2004. Percutaneous penetration enhancers in cigarette mainstream smoke. Food Chem. Toxicol. 42, 9–15.

Smith, C.J., Perfetti, T.A., Garg, R., Hansch, C., 2006. Utility of the mouse dermal promotion assay in comparing the tumorigenic potential of cigarette mainstream smoke. Food Chem. Toxicol. 44, 1699–1706.

Tropsha, A., 2010. Best practices for QSAR model development, validation, and exploitation. Mol. Inf. 29, 476–488.

Tropsha, A., Golbraikh, A., 2007. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Curr. Pharm. Des. 13, 3494–3504.
Tropsha, L., Gramatica, P., Gombar, V.K., 2003. The importance of being ernest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb. Sci. 22, 69–77.
USEPA (United States Environmental Protection Agency), 2012. Estimation Programs Interface Suite™ (EPI Suite) for Microsoft® Windows, v 4.10., Washington, DC, USA. (<http://www.epa.gov/opptintr/exposure/pubs/episuitedl.htm>).
Veith, G.D., Morris, R.T., 1978. A rapid method for estimating log P for organic chemicals. U.S. Environ. Prot. Agency, Ecol. Res. Ser. EPA-600/3-78-049, p. 15.
Veith, G.D., DeFoe, D.L., Bergstedt, B.V., 1979. Measuring and estimating the bioconcentration factor of chemicals in fish. J. Fish. Res. Bd. Can. 36, 1040–1048.
Zhao, Y.H., Abraham, A.H., Zissimos, A.M., 2003. Determination of McGowan volumes for ions and correlation with van der Waals volumes. J. Chem. Inf. Comput. Sci. 43, 1848–1854.