molecular informatics

DOI: 10.1002/minf.201200069

CORAL: Monte Carlo Method as a Tool for the Prediction of the Bioconcentration Factor of Industrial Pollutants

A. P. Toropova,[a] A. A. Toropov,*[a] S. E. Martyanov,[b] E. Benfenati,[a] G. Gini,[c] D. Leszczynska,[d] and J. Leszczynski[e]

Abstract: The CORAL software (http://www.insilico.eu/coral/) has been evaluated for application in QSAR modeling of the bioconcentration factor in fish (log*BCF*). The data used include 237 organic substances (industrial pollutants). Six random splits of the data into sub-training (30–50%), calibration (20–30%), test (13–30%), and validation sets (7–25%) have been carried out. The following numbers display the average statistical characteristics of the models for the external validation set: correlation coefficient $r^2 = 0.880 \pm 0.017$ and standard error of estimation $s = 0.559 \pm 0.131$. The best models were obtained with a combined representation of the molecular structure by SMILES together with hydrogen suppressed graph.

Keywords: QSAR · SMILES · Molecular graph · CORAL software · Bioconcentration factor

1 Introduction

Quantitative Structure–Property/Activity Relationship (QSPR/QSAR) models, which are based on structural descriptors, are often classified as theory. However, they make it possible to formulate a new type of experiments. Instead of experimental work with chemical compounds, one can employ computational treatment of available experimental data to gain novel insight and supplement it by information on compounds not studied experimentally.[1–9]

The choice of the representation of the molecular structure is an important component of the QSPR/QSAR analyses. CORAL software[10,11] is a tool that could be used to build up a QSPR/QSAR model. The Simplified Molecular Input Line Entry System (SMILES) has been tested as representation of the molecular structure for models generated by the CORAL software. However, there are various approaches that could be applied as representations of molecular structures. The molecular graph is the "classic" alternative to SMILES in QSPR/QSAR studies. It should be pointed out that there are endpoints for which a preferable model can be calculated with representation of the molecular structure by SMILES,[13] but there are also endpoints for which a preferable model can be calculated with "hybrid" representation (i.e. taking into account both representations by SMILES and by molecular graph).[14]

The Bioconcentration Factor (*BCF*) represents an important ecological characteristic of substances which can be considered as industrial pollutants.[15] Recently, the CORAL models calculated with SMILES for *BCF* were examined.[16–18] The aim of the present study is to compare the models for *BCF* calculated with: (i) SMILES, (ii) molecular graph, and (iii) the "hybrid" model which is calculated with representation of the molecular structure by SMILES together with molecular graph.[14]

2 Method

2.1 Data

We used bioconcentration factor (fish) data of 239 compounds taken from Lu et al.[15] The CAS numbers of the considered compounds are defined in the US Medicinal Laboratory.[19] We found that two substances in the data set are ambiguous. These are acenaphthalene and ace-

[a] *A. P. Toropova, A. A. Toropov, E. Benfenati*
Istituto di Ricerche Farmacologiche Mario Negri
20156, Via La Masa 19, Milano, Italy
**e-mail: andrey.toropov@marionegri.it*

[b] *S. E. Martyanov*
Teleca OOO
603093, 23, Rodionova st, Nizhny Novgorod, Russia

[c] *G. Gini*
Department of Electronics and Information, Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133 Milano, Italy

[d] *D. Leszczynska*
Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental Engineering, Jackson State University
1325 Lynch St, Jackson, MS 39217-0510, USA

[e] *J. Leszczynski*
Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry, Jackson State University
400 J.R. Lynch Street, P. O. Box 17910, Jackson, MS 39217, USA

Supporting Information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201100069

naphthylene. The analysis of literature data[19] has shown that both substances have the same CAS number 208-96-8. Under these circumstances we have decided to consider the remaining 237 compounds, without the above-mentioned two compounds. We have split the data of these substances six times into sub-training set, calibration set, test set, and validation set. The validation set is not involved in building up the model. These splits are random, but we have tried to obtain the same ranges of endpoint for these three sets. SMILES for calculations with the CORAL software were generated with ACD/ChemSketch.[20]

2.2 Optimal Descriptors

We have formulated the following principles of building up a model of an endpoint with CORAL software:

- The molecular structure of each compound can be represented by molecular features which are extracted from (i) SMILES, (ii) graph, (iii) SMILES together with graph.
- There are local and global molecular features which can be extracted in the above-mentioned cases (i), (ii), and (iii).
- The building up of a QSPR/QSAR model for an arbitrary split into the training and test sets can be examined as a random event.
- The statistical quality of each QSPR/QSAR model is a mathematical function of the split into training and test sets.
- The average statistical quality of QSPR/QSAR models that is obtained for several splits into training and test sets is a more robust criterion for the estimation of an approach than the statistical quality for solely one split.
- The average statistical quality of models for external test sets is a more significant attribute than the average statistical quality for training sets.

The correlation weights for molecular features (which are calculated with SMILES) can be used for classification of the above-mentioned features according to their values for several models into three categories: features with stable positive values of correlation weights (promoters of increase for an endpoint); features with stable negative values of correlation weights (promoters of decrease of an endpoint); and unstable features which have positive values of correlation weights together with negative correlation weight values for several models.

The graph based optimal descriptors are calculated as the following:

$$^{\text{Graph}}DCW(Threshold, N_{\text{epoch}}) = \sum CW(A_k) + \alpha \sum CW(EC0_k) + \beta \sum CW(EC1_k) + \gamma \sum CW(EC2_k) + \delta \sum CW(EC3_k) \quad (1)$$

where A_k is a chemical element (C, O, N, etc.). The chemical elements represent vertexes in hydrogen-suppressed molecular graphs (HSG), covalent bonds are edges in HSG. The extended connectivity of *j*-th order (*ECJ*) is an integer characteristic of a vertex in HSG calculated with the recurrent formula (Figure 1). The extended connectivity can be also calculated with the hydrogen-filled graph and with graph of atomic orbitals.[21,22]

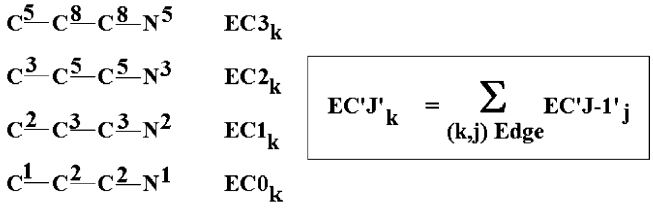

Figure 1. Example of calculation of extended connectivity for vertexes of HSG by the recurrent formula.

The extended connectivity of zero order is the number of vertexes (atoms) connected with a given vertex (atom). The adjacency matrix is the representation of a molecular graph used for computational operations (Figure 2).

The SMILES based optimal descriptors are calculated as the following:

$$^{\text{SMILES}}DCW(Threshold, N_{\text{epoch}}) = a \sum CW(S_k) + \beta \sum CW(SS_k) + \gamma \sum CW(SSS_k) + \delta \cdot CW(PAIR) + x \cdot CW(NOSP) + y \cdot CW(HALO) + z \cdot CW(BOND) \quad (2)$$

where S_k, SS_k, SSS_k are local SMILES attributes which are extracted from SMILES; If SMILES are represented by "ABCDE", the definitions of S_k, SS_k, SSS_k are the following:

S_k : A, B, C, D, E

SS_k : AB, BC, CD, DE

SSS_k : ABC, BCD, CDE

It should be noted that: (i) S_k can be represented by two characters e.g. 'Cl', 'Br', '@@', etc.; (ii) SS_k, SSS_k are ordered according their ASCII code in order to avoid a situation where the same molecular fragment is represented twice: AB and BA, or ABC and CBA. PAIR, NOSP, HALO, and BOND are global SMILES attributes which are extracted from SMILES.[10,23,24] Table 1 contains definition for PAIR. Table 2 contains definitions for NOSP, HALO, and BOND.

The CORAL software gives the possibility to define the "hybrid" optimal descriptors which are calculated as the following:

The calculation of $^{hybrid}DCW(4,60)$ for Acrylonitrile (CAS 107-13-1)

HSG attribute (SA)	Correlation Weight
EC0-C...1...	0.6992890
EC0-C...2...	1.5509784
EC0-C...2...	1.5509784
EC0-N...1...	0.0845414
EC1-C...2...	-2.1337970
EC1-C...3...	1.5904429
EC1-C...3...	1.5904429
EC1-N...2...	0.0

The HSG-representation of the molecular structure

Adjacency matrix

	C	C	C	N	EC0
C	0	2	0	0:	1
C	2	0	1	0:	2
C	0	1	0	3:	2
N	0	0	3	0:	1

SMILES attribute (SA)	Correlation Weight
C.....	-0.5476522
=.....	2.1201360
C.....	-0.5476522
C.....	-0.5476522
#.....	0.0
N.....	-4.5476904
C...=.....	1.6506864
C...=.....	1.6506864
C...C.....	1.9396520
C...#.....	0.0
N...#.....	0.0
HALO00000000	3.8735236
BOND11000000	0.0
++++N---B2==	-0.8658837
++++N---B3==	0.0
++++B2--B3==	0.0

The SMILES-representation of the molecular structure

C=CC#N

$^{hybrid}DCW(4,60) = 9.1110297$

Figure 2. Example of calculation of the hybrid descriptor with correlation weights obtained by the Monte Carlo method

Table 1. The definition of PAIR descriptors indicates simultaneous presence of two molecular features. B2 and B3 are indicators of presence of double and triple bonds, respectively.

	Cl	Br	N	O	S	P	B2	B3
F	++++F— Cl==	++++F— Br==	++++F— N===	++++F— O===	++++F— S===	++++F— P===	++++F— B2==	++++F— B3==
Cl		++++Cl– Br==	++++Cl– N===	++++Cl– O===	++++Cl– S===	++++Cl– P===	++++Cl— B2==	++++Cl– B3==
Br			++++Br– N===	++++Br– O===	++++Br– S===	++++Br– P===	++++Br— B2==	++++Br– B3==
N				++++N— O===	++++N— S===	++++N— P===	++++N— B2==	++++N— B3==
O					++++O— S===	++++O— P===	++++O— B2==	++++O— B3==
S						++++S— P===	++++S— B2==	++++S— B3==
P							++++P— B2==	++++P— B3==
B2								++++B2– B3==

$$^{Hybrid}DCW(Threshold, N_{epoch}) = {}^{SMILES}DCW(Threshold, N_{epoch}) + {}^{Graph}DCW(Threshold, N_{epoch}) \quad (3)$$

Threshold and N_{epoch} represent parameters of the Monte Carlo optimization. *Threshold* is a tool that defines two classes of molecular features (i.e. graph invariants and/or SMILES attributes): rare (noise) and not rare, i.e. active. The optimal descriptors are calculated with the correlation weights of active molecular features (attributes). Correlation weights for rare attributes are fixed equal to zero, i.e. these are not involved in the modeling process. Figure 2 shows the architecture of the hybrid representation of the molecular structure together with correlation weights for various molecular features extracted from HSG and SMILES.

N_{epoch} is the number of iterations of the Monte Carlo optimization. The target function (TF) of the optimization is defined as the following:

Table 2. Definitions of the BOND, NOSP, and HALO attributes.

Calculation of the BOND index			
=	#	@	Comments
0	0	0	There are no double, triple, or stereo chemical bonds
0	0	1	The molecule only contains stereo chemical bonds
0	1	0	The molecule only contains triple bonds
0	1	1	The molecule contains triple and stereo chemical bonds
1	0	0	The molecule only contains double bonds
1	0	1	The molecule contains double bonds and stereo chemical bonds
1	1	0	The molecule contains double and triple bonds
1	1	1	The molecule contains double, triple, and stereo chemical bonds

Calculation of the NOSP index				
N	O	S	P	Comments
0	0	0	0	Nitrogen, oxygen, sulfur, and phosphorus are absent
0	0	0	1	The molecule only contains phosphorus
0	0	1	0	The molecule only contains sulfur
0	0	1	1	The molecule contains sulfur and phosphorus
0	1	0	0	The molecule only contains oxygen
0	1	0	1	The molecule contains oxygen and phosphorus
0	1	1	0	The molecule contains oxygen and sulfur
0	1	1	1	The molecule contains oxygen, sulfur, and phosphorus
1	0	0	0	The molecule only contains nitrogen
1	0	0	1	The molecule contains nitrogen and phosphorus
1	0	1	0	The molecule contains nitrogen and sulfur
1	0	1	1	The molecule contains nitrogen, sulfur, and phosphorus
1	1	0	0	The molecule contains nitrogen and oxygen
1	1	0	1	The molecule contains nitrogen, oxygen and phosphorus
1	1	1	0	The molecule contains nitrogen, oxygen, and sulfur
1	1	1	1	The molecule contains nitrogen, oxygen, sulfur, and phosphorus

Calculation of the HALO index			
F	Cl	Br	Comments
0	0	0	Fluorine, chlorine and bromine are absent
0	0	1	The molecule only contains bromine
0	1	0	The molecule only contains chlorine
0	1	1	The molecule contains chlorine and bromine
1	0	0	The molecule only contains fluorine
1	0	1	The molecule contains fluorine and bromine
1	1	0	The molecule contains fluorine and chlorine
1	1	1	The molecule contains fluorine, chlorine, and bromine

$$TF = R + R' - W_R \cdot |R - R'| - W_C \cdot (|C_0| + |C_0'| + |C_1 - C_1'|) \quad (4)$$

where R and R' are correlation coefficients between the optimal descriptor and an endpoint (*EP*) for sub-training and calibration sets, respectively; C_0, C_1, $C_{0'}$,and $C_{1'}$ are coefficients from equations obtained by the Least squares method:

$$EP = C_0 + C_1 \cdot^{\text{Choice}} DCW(Threshold, N_{\text{epoch}}) \quad (5)$$

or sub − training set

$$EP = C_0' + C_1' \cdot^{\text{Choice}} DCW(Threshold, N_{\text{epoch}}) \quad (6)$$

or calibration set

$W_R = 0.1$ and $W_C = 0.01$ are empirical parameters; 'Choice' includes SMILES, or Graph, or Hybrid. Coefficients α, β, γ, δ, x, y, and z can be 0 or 1: it gives possibility to select different versions for the optimal descriptors.

The increase of the threshold leads to decrease of correlation coefficient (between experimental and calculated values of endpoint) for the sub-training and calibration sets, but as the rule, there is a maximum of the correlation coefficient for the test set. The increase of the number of epochs of the Monte Carlo optimization leads to increase of the correlation coefficient for sub-training and calibration sets, but again, as the rule, there is the maximum of the correlation coefficient for the test set. Thus, it is necessary to define preferable values of the threshold (T*) and

the number of epochs (N*) which provide maximum of correlation coefficient for the test set (Figure 1).

The method that has been used for the HSG-based models is defined as $\alpha=1$, $\beta=1$, $\gamma=0$, $\delta=0$ (in Equation 1). The method that has been used for the SMILES-based models is defined as $\alpha=1$, $\beta=1$, $\gamma=0$, $\delta=1$, $x=0$, $y=0$, $z=1$ (in Equation 2). The hybrid method is the unification of the two described methods (HSG-based and SMILES-based).

3 Results and Discussion

Table 3 contains statistical characteristics of one-variable models calculated with Equation 5 for the sub-training, calibration, and test sets, for six random splits. One can see that preferable models are obtained in the case of the hybrid representation of the molecular structure (i.e. by SMILES together with hydrogen suppressed graph). Unfortunately, the best models are revealed for quite different values of the threshold (*T**) and values of the number of epochs (*N**). It indicates that a split into sub-training, calibration, and test sets influences the statistical quality of the models.

Table 4 contains various criteria of predictability[2,25,26] for the above-mentioned best models. One can see (Table 4) that all six models are quite acceptable according to those criteria. Figure 4 contains the graphical representation of models for log*BCF* (for six splits) which are calculated with the CORAL software.

According to OECD principles,[27] a QSPR/QSAR model must be associated with the following information:

- a defined endpoint;
- an unambiguous algorithm;
- a defined domain of applicability;
- appropriate measures of goodness-of-fit, robustness and predictability;
- a mechanistic interpretation, if possible.

The approach described above has been applied in building up the log*BCF* model using data taken from the literature.[15] Thus the endpoint should be classified as a quite "defined". The algorithm of the Monte Carlo optimization has been described and checked up in a few previous studies.[10,13,14,16] The ideal applicability domain for CORAL models involves substances without blocked attributes. In reality, however, one should use some compromise, e.g. select substances with less than 10% of blocked attributes. The correlation coefficient between experimental and calculated values of an endpoint for test set can be used as a measure of statistical quality of a CORAL model. Stable positive values of correlation weight in series runs of the Monte Carlo optimization are indicators of molecular features which are promoters of increase for an endpoint. Contrary, stable negative values of correlation weights are indicator of molecular features which are promoters of de-

Table 3. QSAR models for the bioconcentration factor log*BCF* calculated according to scheme shown in Figure 3. The best statistical characteristics are indicated in bold.

Split	*T**	*N**	Sub-training set			Calibration set			Test set		
			n	r^2	*s*	*n*	r^2	s	*n*	r^2	*s*
Optimal descriptors calculated with SMILES											
1	10	50	118	0.8303	0.561	50	0.8303	0.634	45	0.8436	0.517
2	2	30	122	0.8270	0.577	50	0.8561	0.525	39	0.8482	0.465
3	3	24	87	0.7728	0.602	63	0.8384	0.669	70	0.8507	0.475
4	5	49	84	0.8119	0.599	73	0.8811	0.545	62	0.8514	0.514
5	9	22	72	0.7047	0.763	73	0.7158	0.778	31	0.8348	0.512
6	6	29	82	0.7822	0.563	65	0.8243	0.490	69	0.8364	0.631
Optimal descriptors calculated with hydrogen suppressed molecular graph											
1	4	33	118	0.7916	0.622	50	0.8509	0.603	45	0.8432	0.516
2	1	30	122	0.8129	0.600	50	0.8386	0.618	39	0.8896	0.512
3	2	25	87	0.7732	0.601	63	0.8224	0.706	70	0.8625	0.576
4	3	50	84	0.7953	0.624	73	0.8860	0.404	62	0.8424	0.534
5	7	26	72	0.7079	0.759	73	0.7348	0.759	31	0.8573	0.485
6	1	35	82	0.8326	0.603	65	0.8325	0.453	69	0.8204	0.637
Optimal descriptors calculated with SMILES together with hydrogen suppressed molecular graph											
1	4	49	118	0.8841	0.464	50	0.8897	0.529	45	**0.8629**	**0.467**
2	1	38	122	0.8813	0.478	50	0.9012	0.446	39	**0.9158**	**0.440**
3	1	27	87	0.8334	0.515	63	0.8804	0.603	70	**0.8858**	**0.529**
4	2	42	84	0.8741	0.490	73	0.8996	0.443	62	**0.8930**	**0.520**
5	1	32	72	0.8691	0.508	73	0.8684	0.576	31	**0.8788**	**0.438**
6	5	32	82	0.8264	0.614	65	0.8691	0.413	69	**0.8836**	**0.541**

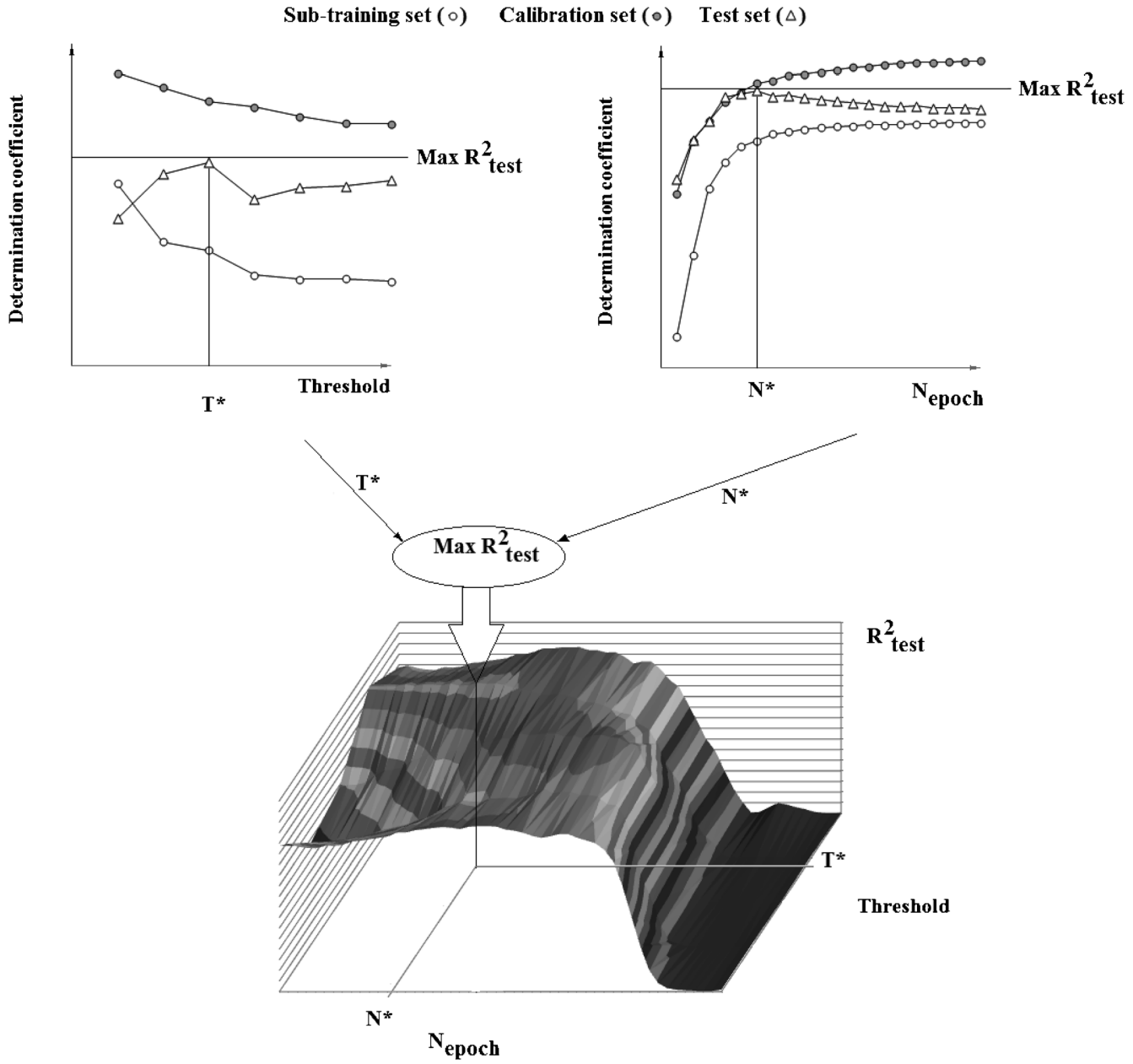

Figure 3. General scheme of the building up of CORAL models.

crease for an endpoint. This data can be an indication for a mechanistic interpretation related to a given endpoint. For a given data set and for all six splits, the carbon vertex in HSG with Morgan's connectivity of zero order (i.e. $^{0}EC_{k}=2$), branching in an aromatic system (i.e. fragment of SMILES = 'c('), and an aromatic ring (i.e. fragment of SMILES = 'c1') are promoters of increase for log*BCF*. On the other hand, $^{0}EC_{k}=3$; $^{1}EC_{k}=4$ (carbon vertex in HSG); and the presence of oxygen together with chlorine (*PAIR*) are promoters of decrease for log*BCF*. Thus, the developed CORAL model follows the OECD principles.

Two models for log*BCF* described by Lu et al.[15] have been done for two groups of substances according to the range of octanol water partition coefficient ($\log K_{ow}$). The first model is characterized by $n=214$, $r^2=0.781$, $s=0.614$ ($1<\log K_{ow}<7$). The second model is characterized by $n=20$, $r^2=0.795$, $s=0.617$ ($\log K_{ow}>7$). The statistical characteristics of the log*BCF* model from Toropov et al.[18] are $n=105$, $r^2=0.805$, $s=0.528$. The statistical characteristics of a model for log*BCF* from Sahu and Singh[28] are $n=131$, $r^2=0.871$, $s=0.978$. The model for log*BCF* from Jacksonet al.[29] gives $n=93$, $r^2=0.88$. The model for log*BCF* from Dimitrov et al.[30] is characterized by $n=511$, $r^2=0.84$. According to

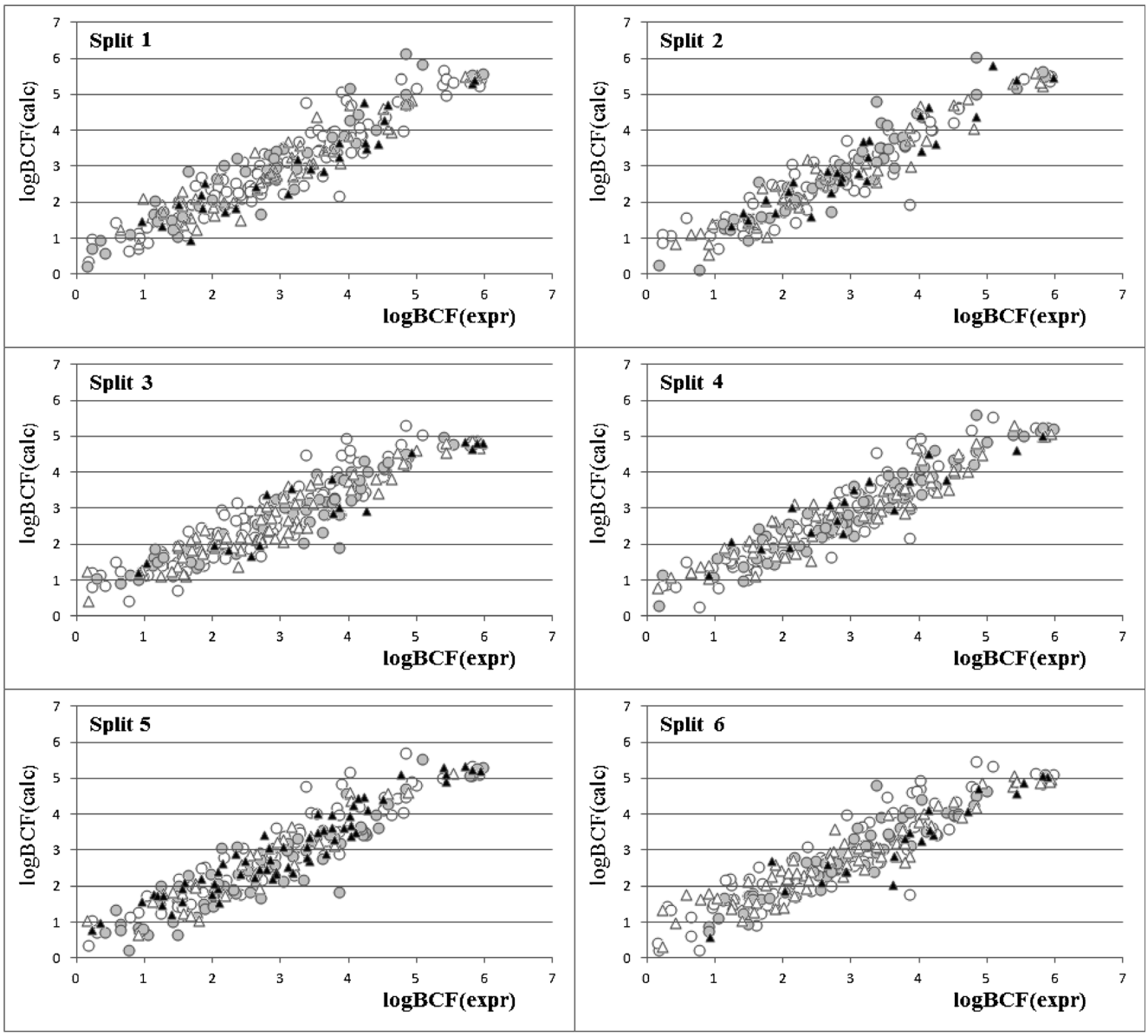

Figure 4. Graphical representation of the best CORAL models for log*BCF*.

Lombardo et al.,[31] BCFBAF v3.00 and CAESAR give models of log*BCF* which are characterized by $n = 527$, $r^2 = 0.75$, $s = 0.68$ and $n = 527$, $r^2 = 0.81$, $s = 0.57$, respectively.

The comparison of the statistical quality of the above-mentioned models described in the literature[15,18,28–31] with the statistical quality of the models represented in Tables 3 and 4 shows that the CORAL software gives quite good models for log*BCF*.

The CORAL software is available (freely) on the Internet together with instructions how to use this software. It should be noted that the CORAL software has been tested as a tool of QSAR analysis of various endpoints (not only log*BCF*).[8–11,13,14,32–34]

Supporting Information

The Supporting Information contains details of six splits into the sub-training, calibration, and test sets which were examined in this study together with data on the log*BCF* and octanol/water partition coefficient.

4 Conclusions

The models for the bioconcentration factor (log*BCF*) developed here, by means of the CORAL software, are confirmed to comply with the OECD principles. The statistical quality

Table 4. Criteria of predictability[2,25,26] for models calculated with the CORAL software.

Split	The statistical characteristics
1	log*BCF* = 0.0037 (±0.0100) + 0.0840 (±0.0003) * DCW(4,49)
	Test set
	$n = 45$
	$r^2 = 0.8629$
	$r_0^2 = 0.8455$
	$r'_0{}^2 = 0.8629$
	$\frac{r^2 - r_0^2}{r^2} = 0.0000 < 0.1$
	$\frac{r^2 - r'^2_0}{r^2} = 0.0201 < 0.1$
	$k = 1.0035 (0.85 < k < 1.15)$
	$k' = 0.9753$ $(0.85 < k' < 1.15)$
	$r_m^2 = 0.8599 > 0.5$
	$\overline{r_m^2} = 0.8045 > 0.5, \Delta r_m^2 = 0.1108 < 0.2$
	Validation set
	$n = 24$
	$r^2 = 0.8850$
	$r_0^2 = 0.8839$
	$r_{0'}^2 = 0.8791$
	$\frac{r^2 - r_0^2}{r^2} = 0.0012 < 0.1$
	$\frac{r^2 - r'^2_0}{r^2} = 0.0067 < 0.1$
	$k = 1.0702 (0.85 < k < 1.15)$
	$k' = 0.9177$ $(0.85 < k' < 1.15)$
	$r_m^2 = 0.8560 > 0.5$
	$\overline{r_m^2} = 0.8364 > 0.5, \Delta r_m^2 = 0.0391 < 0.2$
2	log*BCF* = 0.3806 (±0.0120) + 0.0789 (±0.0003) * DCW(1,38)
	Test set
	$n = 39$
	$r^2 = 0.9158$
	$r_0^2 = 0.9056$
	$r_{0'}^2 = 0.9153$
	$\frac{r^2 - r_0^2}{r^2} = 0.0111 < 0.1$
	$\frac{r^2 - r'^2_0}{r^2} = 0.0005 < 0.1$
	$k = 0.9706$ $(0.85 < k < 1.15)$
	$k' = 1.0115$ $(0.85 < k' < 1.15)$
	$r_m^2 = 0.8967 > 0.5$
	$\overline{r_m^2} = 0.8600 > 0.5, \Delta r_m^2 = 0.0734 < 0.2$
	Validation set
	$n = 26$
	$r^2 = 0.8901$
	$r_0^2 = 0.8857$
	$r_{0'}^2 = 0.8885$
	$\frac{r^2 - r_0^2}{r^2} = 0.0049 < 0.1$
	$\frac{r^2 - r'^2_0}{r^2} = 0.0017 < 0.1$
	$k = 0.9995$ $(0.85 < k < 1.15)$
	$k' = 0.9851$ $(0.85 < k' < 1.15)$
	$r_m^2 = 0.8552 > 0.5$
	$\overline{r_m^2} = 0.8432 > 0.5, \Delta r_m^2 = -0.0239 < 0.2$

Table 4. (Continued)

Split	The statistical characteristics
3	log*BCF* = 0.0004 (±0.0165) + 0.0507 (±0.0003) * DCW(1,27)
	Test set
	$n = 70$
	$r^2 = 0.8858$
	$r_0^2 = 0.8566$
	$r_{0'}^2 = 0.8831$
	$\frac{r^2 - r_0^2}{r^2} = 0.0330 < 0.1$
	$\frac{r^2 - r'^2_0}{r^2} = 0.0030 < 0.1$
	$k = 0.8917$ $(0.85 < k < 1.15)$
	$k' = 1.0982$ $(0.85 < k' < 1.15)$
	$r_m^2 = 0.8400 > 0.5$
	$\overline{r_m^2} = 0.7872 > 0.5, \Delta r_m^2 = 0.1056 < 0.2$
	Validation set
	$n = 17$
	$r^2 = 0.8700$
	$r_0^2 = 0.8691$
	$r_{0'}^2 = 0.8432$
	$\frac{r^2 - r_0^2}{r^2} = 0.0011 < 0.1$
	$\frac{r^2 - r'^2_0}{r^2} = 0.0308 < 0.1$
	$k = 1.1552$ $(0.85 < k < 1.15)$
	$k' = 0.8471$ $(0.85 < k' < 1.15)$
	$r_m^2 = 0.8431 > 0.5$
	$\overline{r_m^2} = 0.7854 > 0.5, \Delta r_m^2 = 0.1155 < 0.2$
4	log*BCF* = 0.0132 (±0.0148) + 0.0700 (±0.0004) * DCW(2,42)
	Test set
	$n = 62$
	$r^2 = 0.8930$
	$r_0^2 = 0.8267$
	$r_{0'}^2 = 0.8776$
	$\frac{r^2 - r_0^2}{r^2} = 0.0742 < 0.1$
	$\frac{r^2 - r'^2_0}{r^2} = 0.0171 < 0.1$
	$k = 0.9409$ $(0.85 < k < 1.15)$
	$k' = 1.0374$ $(0.85 < k' < 1.15)$
	$r_m^2 = 0.7825 > 0.5$
	$\overline{r_m^2} = 0.7228 > 0.5, \Delta r_m^2 = 0.1194 < 0.2$
	Validation set
	$n = 18$
	$r^2 = 0.8476$
	$r_0^2 = 0.8330$
	$r_{0'}^2 = 0.7493$
	$\frac{r^2 - r_0^2}{r^2} = 0.0172 < 0.1$
	$\frac{r^2 - r'^2_0}{r^2} = 0.1160 < 0.1$
	$k = 1.0120$ $(0.85 < k < 1.15)$
	$k' = 0.9637$ $(0.85 < k' < 1.15)$
	$r_m^2 = 0.7452 > 0.5$
	$\overline{r_m^2} = 0.6635 > 0.5, \Delta r_m^2 = 0.1634 < 0.2$

Table 4. (Continued)

Split	The statistical characteristics
5	$\log BCF = 0.0052\ (\pm 0.0133) + 0.0900\ (\pm 0.0005)$ * DCW(1,47)
	Test set
	$n = 31$
	$r^2 = 0.8788$
	$r_0^2 = 0.8707$
	$r_{0'}^2 = 0.8782$
	$\frac{r^2 - r_0^2}{r^2} = 0.0007 < 0.1$
	$\frac{r^2 - r'^2_0}{r^2} = 0.0093 < 0.1$
	$k = 0.9928\ (0.85 < k < 1.15)$
	$k' = 0.9859\ (0.85 < k' < 1.15)$
	$r_m^2 = 0.8573 > 0.5$
	$\overline{r_m^2} = 0.8283 > 0.5, \Delta r_m^2 = 0.0579 < 0.2$
	Validation set
	$n = 61$
	$r^2 = 0.9043$
	$r_0^2 = 0.9011$
	$r_{0'}^2 = 0.8794$
	$\frac{r^2 - r_0^2}{r^2} = 0.0036 < 0.1$
	$\frac{r^2 - r'^2_0}{r^2} = 0.0276 < 0.1$
	$k = 1.0308\ (0.85 < k < 1.15)$
	$k' = 0.9548\ (0.85 < k' < 1.15)$
	$r_m^2 = 0.8530 > 0.5$
	$\overline{r_m^2} = 0.8072 > 0.5, \Delta r_m^2 = 0.0917 < 0.2$
6	$\log BCF = -0.0967\ (\pm 0.0179) + 0.0630\ (\pm 0.0004)$ * DCW(5,32)
	Test set
	$n = 69$
	$r^2 = 0.8836$
	$r_0^2 = 0.8125$
	$r_{0'}^2 = 0.8685$
	$\frac{r^2 - r_0^2}{r^2} = 0.0805 < 0.1$
	$\frac{r^2 - r'^2_0}{r^2} = 0.0171 < 0.1$
	$k = 0.9255\ (0.85 < k < 1.15)$
	$k' = 1.0519\ (0.85 < k' < 1.15)$
	$r_m^2 = 0.7750 > 0.5$
	$\overline{r_m^2} = 0.7115 > 0.5, \Delta r_m^2 = 0.1271 < 0.2$
	Validation set
	$n = 21$
	$r^2 = 0.8838$
	$r_0^2 = 0.8766$
	$r_{0'}^2 = 0.8829$
	$\frac{r^2 - r_0^2}{r^2} = 0.0081 < 0.1$
	$\frac{r^2 - r'^2_0}{r^2} = 0.0010 < 0.1$
	$k = 1.1406\ (0.85 < k < 1.15)$
	$k' = 0.8651\ (0.85 < k' < 1.15)$
	$r_m^2 = 0.8580 > 0.5$
	$\overline{r_m^2} = 0.8335 > 0.5, \Delta r_m^2 = -0.0490 < 0.2$

of these models is a mathematical function of the split into the sub-training, calibration, and test sets. These models have been checked up with external validation sets (i.e. with substances which were not involved in building up the model). Thus, for each split, the CORAL software gives a quite good model.

Acknowledgements

The authors express their gratitude to *ANTARES* (Project Number LIFE08-ENV/IT/00435) and the *NSF CREST Interdisciplinary Nanotoxicity Center NSF-CREST*, Grant # HRD-0833178 for financial support, and to Dr. *L. Cappellini*, Dr. *G. Bianchi* and Dr. *R. Bagnati* for valuable consultations on the computer science.

References

[1] J. T. Leonard, K. Roy, *Eur. J. Med. Chem.* **2008**, *43*, 81–92.
[2] P. P. Roy, K. Roy, *QSAR Comb. Sci.* **2008**, *27*, 302–313.
[3] G. Melagraki, A. Afantitis, H. Sarimveis, P. A. Koutentis, G. Kollias, O. Igglessi-Markopoulou, *Mol. Divers.* **2009**, *13*, 301–311.
[4] A. Afantitis, G. Melagraki, H. Sarimveis, P. A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, *QSAR Comb. Sci.* **2008**, *27*, 432–436.
[5] E. Vicente, P. R. Duchowicz, E. A. Castro, A. Monge, *J. Mol. Graphics. Model.* **2009**, *28*, 28–36.
[6] P. R. Duchowicz, E. A. Castro, *Int. J. Mol. Sci.* **2009**, *10*, 2558–2577.
[7] B. Furtula, I. Gutman, *J. Chemometr.* **2011**, *25*, 87–91.
[8] J. García, P. R. Duchowicz, M. F. Rozas, J. A. Caram, M. V. Mirífico, F. M. Fernández, E. A. Castro, *J. Mol. Graph. Model.* **2011**, *31*, 10–19.
[9] L. M.A. Mullen, P. R. Duchowicz, E. A. Castro, *Chemometr. Intell. Lab.* **2011**, *107*, 269–275.
[10] A. A. Toropov, A. P. Toropova, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, *Anti-Cancer Agents Med. Chem.* **2011**, *11*, 974–982.
[11] E. Benfenati, A. A. Toropov, A. P. Toropova, A. Manganaro, R. Gonella Diaza, *Chem. Biol. Drug Des.* **2011**, *77*, 471–476.
[12] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
[13] A. A. Toropov, A. P. Toropova, S. E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, *Chemometr. Intell. Lab.* **2011**, *109*, 94–100.
[14] A. P. Toropova, A. A. Toropova, S. E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, *Chemometr. Intell. Lab.* **2012**, *110*, 177–181.
[15] X. Lu, S. Tao, H. Hu, R. W. Dawson, *Chemosphere* **2000**, *41*, 1675–1688.
[16] A. A. Toropov, A. P. Toropova, A. Lombardo, A. Roncaglioni, E. Benfenati, G. Gini , *Eur. J. Med. Chem.* **2011**, *46*, 1400–1403.
[17] A. P. Toropova, A. A. Toropov, A. Lombardo, A. Roncaglioni, E. Benfenati, G. Gini, *Eur. J. Med. Chem.***2010**, *45*, 4399–4402.
[18] A. A. Toropov, A. P. Toropova, E. Benfenati, *Eur. J. Med. Chem.* **2009**, *44*, 2544–2551.
[19] *US National Medicinal library*, http://toxnet.nlm.nih.gov/, accessed December **2011.**
[20] *ACD/ChemSketch Freeware*, Version 11.00, Advanced Chemistry Development, Inc., Toronto, ON, Canada, www.acdlabs.com, **2007**, accessed February **2009**
[21] A. A. Toropov, A. P. Toropova, *J. Mol. Struct. (THEOCHEM)* **2003**, *637*, 1–10.
[22] A. A. Toropov, A. P. Toropova, *J. Mol. Struct. (THEOCHEM)* **2002**, *578*, 129–134.
[23] A. P. Toropova, A. A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, *J. Comput. Chem.* **2011**, *32*, 2727–2733.
[24] *The CORAL software*, http://www.insilico.eu/coral/, accessed January, **2012**

[25] A. Golbraikh, A. Tropsha, *J. Mol. Graph. Model.* **2002**, *20*, 269–276.
[26] P. K. Ojha, I. Mitra, R. N. Das, K. Roy, *Chemometr. Intell. Lab.* **2011**, *107*, 194–205.
[27] Organization for Economic Co-operation and Development http://www.oecd.org/dataoecd/55/35/38130292.pdf, accessed February **2012**
[28] V. K. Sahu, R. K. Singh, *Clean Soil, Air, Water* **2009**, *37*, 850–857.
[29] S. H. Jackson, C. E. Cowan-Ellsberry, G. Thomas, *J. Agric. Food Chem.* **2009**, *57*, 958–967.
[30] S. Dimitrov, N. Dimitrova, T. Parkerton, M. Comber, M. Bonnell, O. Mekenyan, *SAR QSAR Environ. Res.* **2005**, *16*, 531–554.
[31] A. Lombardo, A. Roncaglioni, E. Boriani, C. Milan, E. Benfenati, *Chem. Centr. J.* **2010**, *4* (SUPPL. 1), S1 (http://www.journal.chemistrycentral.com/content/4/S1/S1).
[32] J. C. Garro Martinez, P. R. Duchowicz, M. R. Estrada, G. N. Zamarbide, E. A. Castro, *Int. J. Mol. Sci.* **2011**, *12*, 9354–9368.
[33] E. Ibezim, P. R. Duchowicz, E. V. Ortiz, E. A. Castro, *Chemometr. Intell. Lab.* **2012**, *110*, 81–88.
[34] A. A. Toropov, A. P. Toropova, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, *Anti-Cancer Agents Med. Chem.* **2012**, *12*, 807–817.

Received: July 5, 2012
Accepted: November 16, 2012
Published online: November 30, 2012

