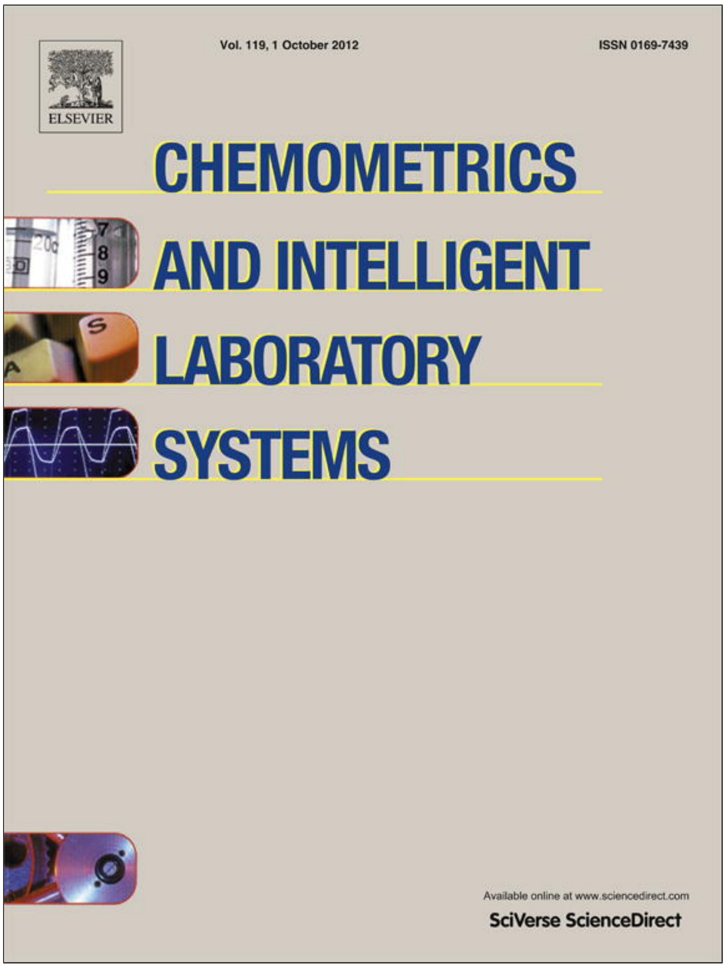


Vol. 119, 1 October 2012

ISSN 0169-7439

ELSEVIER

CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems 119 (2012) 39–43

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Short Communication

CORAL: Models of toxicity of binary mixtures

Alla P. Toropova [a], Andrey A. Toropov [a,*], Emilio Benfenati [a], Giuseppina Gini [b], Danuta Leszczynska [c], Jerzy Leszczynski [d]

[a] *Istituto di Ricerche Farmacologiche Mario Negri, 20156, Via La Masa 19, Milano, Italy*
[b] *Department of Electronics and Information, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy*
[c] *Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental Engineering, Jackson State University, 1325 Lynch St, Jackson, MS 39217-0510, USA*
[d] *Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry, Jackson State University, 1400 J. R. Lynch Street, P.O. Box 17910, Jackson, MS 39217, USA*

ARTICLE INFO

Article history:
Received 15 August 2012
Received in revised form 1 October 2012
Accepted 5 October 2012
Available online 13 October 2012

Keywords:
QSAR
Toxicity
Binary mixture
CORAL software

ABSTRACT

Quantitative structure–activity relationships (QSAR) for toxicity of binary mixtures (expressed as pEC50 (i.e. log [1/EC50], logarithm of the inverse of the effective concentration required to bring about a 50% decrease in light emission), for *Photobacterium phosphoreum*) have been developed. The simplified molecular input-line entry system (SMILES) was used as the representation of the molecular structure of components of binary mixtures. Using the Monte Carlo technique the SMILES-based optimal descriptors were calculated. One-variable correlations between the optimal descriptors and toxicity of the binary mixtures were analyzed to develop a predictive model. Six random splits of the data into sub-training, calibration, and test sets were tested. A satisfactory statistical quality of the model was achieved for each above-mentioned split.

© 2012 Published by Elsevier B.V.

1. Introduction

Toxicity represents a complex phenomenon, investigated by both experimental techniques and computational methods. Quantitative structure–activity relationships (QSARs) are a tool for prediction of various endpoints, in general [1–8], and for toxicity, in particular [9,10]. Prediction of toxicity becomes even more complicated when toxicity is caused by a number of factors, not a single chemical compound. Data on toxicity of mixtures, in general, and on toxicity of binary mixtures, in particular, is important from ecological point of view. There are several studies related to this issue [11–14]. However, due to its importance and complexity, novel approaches that could generate larger pool of data are needed. In this study we tested the CORAL software [15–19] as a possible tool to model the toxicity of binary mixtures.

2. Method

2.1. Data

The numerical data on the toxicity of binary mixtures was taken from the literature [11]. The toxicities are expressed as pEC_{50} (i.e. negative decimal logarithm $\log[1/EC_{50}]$), logarithm of the inverse of the effective concentration required to bring about a 50% decrease in light emission, for *Photobacterium phosphoreum* (T3 mutation). Table 1 contains the list of substances which are components of the binary mixtures. The SMILES used for the representation of the binary mixtures are displayed in the Table 2. In this study six splits into the sub-training set, calibration set, test set, and validation set were examined. These splits were carried out according to the following principles: (i) the range of the endpoint should be similar for each set; and (ii) the distribution of data into above-mentioned sets should be different for each split. The validation set represents a list of substances which are not involved in the process of the building up a model.

2.2. Optimal SMILES-based descriptor

The optimal descriptor used in this study is calculated as follows:

$$\mathrm{DCW(T,N)} = \Sigma\,\mathrm{CW}(S_k) + \Sigma\,\mathrm{CW}(\mathrm{SS}_k) + \Sigma\,\mathrm{CW}(\mathrm{SSS}_k) \qquad (1)$$

where S_k, SS_k, SSS_k are attributes of SMILES notation [20]. The S_k, SS_k, and SSS_k contain one, two, and three SMILES elements, respectively; the element of SMILES often is one character (e.g. 'c', 'C', '=', etc.) but also it can be more than one character (e.g. 'Cl', 'Br', etc); the $CW(S_k)$, $CW(SS_k)$, $CW(SSS_k)$ are correlation weights of SMILES attributes which represent various molecular features extracted from SMILES. By means of the Monte Carlo method optimization procedure [15–19] one can calculate correlation weights which yield the maximum for target function calculated as:

$$\mathrm{TF} = \mathrm{R} + \mathrm{R}' - |\mathrm{R} - \mathrm{R}'| * \mathrm{C} \qquad (2)$$

where R and R' are correlation coefficients between DCW(T,N) and pEC_{50} for sub-training set and calibration set; C is an empirical constant equal to 0.1. Table 3 contains an example of the representation of the

[*] Corresponding author.
E-mail address: andrey.toropov@marionegri.it (A.A. Toropov).

0169-7439/$ – see front matter © 2012 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.chemolab.2012.10.001

Table 1
Structure of components of the binary mixtures.

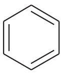
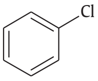
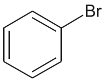
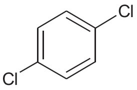
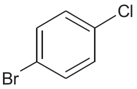
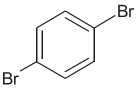
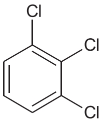
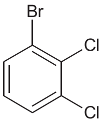
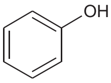
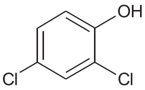
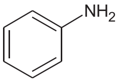
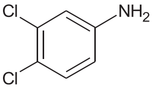
1	71-43-2		c1ccccc1
2	108-90-7		Clc1ccccc1
3	108-86-1		Brc1ccccc1
4	106-46-7		Clc1ccc(Cl)cc1
5	106-39-8		Clc1ccc(Br)cc1
6	106-37-6		Brc1ccc(Br)cc1
7	87-61-6		Clc1cccc(Cl)c1Cl
8	56961-77-4		Clc1cccc(Br)c1Cl
9	108-95-2		Oc1ccccc1
10	120-83-2		Clc1cc(Cl)c(O)cc1
11	62-53-3		Nc1ccccc1
12	95-76-1		Nc1cc(Cl)c(Cl)cc1

Table 2
SMILES which have been used for representation of the binary mixtures and numerical data on the pEC50.

No.	Comp1 + Comp2	SMILES	pEC_{50}
1	1 + 2	c1ccccc1.Clc1ccccc1	2.85
2	1 + 3	c1ccccc1.Brc1ccccc1	2.99
3	1 + 4	c1ccccc1.Clc1ccc(Cl)cc1	2.94
4	1 + 5	c1ccccc1.Clc1ccc(Br)cc1	3.03
5	1 + 6	c1ccccc1.Brc1ccc(Br)cc1	2.96
6	1 + 7	c1ccccc1.Clc1cccc(Cl)c1Cl	3.02
7	1 + 8	c1ccccc1.Clc1cccc(Br)c1Cl	2.98
8	2 + 3	Clc1ccccc1.Brc1ccccc1	3.73
9	2 + 4	Clc1ccccc1.Clc1ccc(Cl)cc1	3.88
10	2 + 5	Clc1ccccc1.Clc1ccc(Br)cc1	3.97
11	2 + 6	Clc1ccccc1.Brc1ccc(Br)cc1	3.96
12	2 + 7	Clc1ccccc1.Clc1cccc(Cl)c1Cl	3.89
13	2 + 8	Clc1ccccc1.Clc1cccc(Br)c1Cl	3.90
14	3 + 4	Brc1ccccc1.Clc1ccc(Cl)cc1	3.98
15	3 + 5	Brc1ccccc1.Clc1ccc(Br)cc1	4.09
16	3 + 6	Brc1ccccc1.Brc1ccc(Br)cc1	4.02
17	3 + 7	Brc1ccccc1.Clc1cccc(Cl)c1C	4.06
18	3 + 8	Brc1ccccc1.Clc1cccc(Br)c1C	4.04
19	4 + 5	Clc1ccc(Cl)cc1. Clc1ccc(Br)cc1	4.36
20	4 + 6	Clc1ccc(Cl)cc1. Brc1ccc(Br)cc1	4.34
21	4 + 7	Clc1ccc(Cl)cc1. Clc1cccc(Cl)c1Cl	4.38
22	4 + 8	Clc1ccc(Cl)cc1. Clc1cccc(Br)c1Cl	4.39
23	5 + 6	Clc1ccc(Br)cc1. Brc1ccc(Br)cc1	4.49
24	5 + 7	Clc1ccc(Br)cc1. Clc1cccc(Cl)c1Cl	4.65
25	5 + 8	Clc1ccc(Br)cc1. Clc1cccc(Br)c1Cl	4.55
26	6 + 7	Brc1ccc(Br)cc1. Clc1cccc(Cl)c1Cl	4.62
27	6 + 8	Brc1ccc(Br)cc1. Clc1cccc(Br)c1Cl	4.45
28	7 + 8	Clc1cccc(Cl)c1Cl. Clc1cccc(Br)c1Cl	4.70
29	2 + 9	Clc1ccccc1. Oc1ccccc1	3.03
30	2 + 10	Clc1ccccc1. Clc1cc(Cl)c(O)cc1	3.42
31	2 + 11	Clc1ccccc1. Nc1ccccc1	2.45
32	2 + 12	Clc1ccccc1. Nc1cc(Cl)c(Cl)cc1	3.67
33	3 + 9	Brc1ccccc1. Oc1ccccc1	3.28
34	3 + 10	Brc1ccccc1. Clc1cc(Cl)c(O)cc1	3.77
35	3 + 11	Brc1ccccc1. Nc1ccccc1	2.68
36	3 + 12	Brc1ccccc1. Nc1cc(Cl)c(Cl)cc1	3.91
37	7 + 9	Clc1cccc(Cl)c1Cl. Oc1ccccc1	3.39
38	7 + 10	Clc1cccc(Cl)c1Cl. Clc1cc(Cl)c(O)cc1	4.27
39	7 + 11	Clc1cccc(Cl)c1Cl. Nc1ccccc1	2.63
40	7 + 12	Clc1cccc(Cl)c1Cl. Nc1cc(Cl)c(Cl)cc1	4.29
41	8 + 9	Clc1cccc(Br)c1Cl. Oc1ccccc1	3.42
42	8 + 10	Clc1cccc(Br)c1Cl. Clc1cc(Cl)c(O)cc1	4.66
43	8 + 11	Clc1cccc(Br)c1Cl. Nc1ccccc1	2.91
44	8 + 12	Clc1cccc(Br)c1Cl. Nc1cc(Cl)c(Cl)cc1	4.52
45	9 + 10	Oc1ccccc1. Clc1cc(Cl)c(O)cc1	3.11
46	9 + 11	Oc1ccccc1. Nc1ccccc1	2.50
47	9 + 12	Oc1ccccc1. Nc1cc(Cl)c(Cl)cc1	3.16
48	10 + 11	Clc1cc(Cl)c(O)cc1. Nc1ccccc1	2.60
49	10 + 12	Clc1cc(Cl)c(O)cc1. Nc1cc(Cl)c(Cl)cc1	4.44
50	11 + 12	Nc1ccccc1. Nc1cc(Cl)c(Cl)cc1	2.50

SMILES attributes and their correlation weights. An example of the DCW(T,N) calculation for binary mixture is shown in Table 4.

In general, the correlation coefficients between experimental values of an endpoint and the value calculated with the optimal descriptor are mathematical functions of the threshold. The threshold represents a coefficient for division of the SMILES attributes into two categories: rare and active (Fig. 1), and the number of the epochs of the optimization (Fig. 2). The SMILES for representation of the binary mixtures were combinations of two SMILES of pure components of the mixture separated by '.' [20].

3. Results and discussion

Table 5 shows the statistical quality of the model for the toxicity of binary mixtures obtained by means of six different splits of the data into the sub-training, calibration, and test sets. The threshold and the number of epochs of the Monte Carlo optimization were selected in order to obtain the best statistical quality for the test set. One can see that preferable threshold and the number of epochs are not identical for the examined splits.

It should be noted that the balance of correlations [21] (i.e., the split into the sub-training, calibration and test set) provides for all six random splits considerably better statistical quality of the prediction, in comparison to the "classic scheme" (i.e. the split into training and test sets without calibration). In the case of the balance of correlations the calibration set plays the role of a "preliminary test set". The preliminary test of a model gives possibility to avoid, or at least to decrease the probability of the overtraining [21–26].

The statistical characteristics of the models for six splits calculated with the preferable threshold (T^*) and the number of epochs (N^*_{ep}) are the following:

$$\begin{aligned}
&\text{Split 1}\\
&pEC_{50} = -0.0090\ (\pm 0.0678) + 0.1148\ (\pm 0.0022) * \ DCW(2,10)\\
&n = 14,\ r^2 = 0.9584,\ q^2 = 0.9426,\ s = 0.167,\ F = 277\ (\text{Sub}-\text{training set})\\
&n = 14,\ r^2 = 0.9566,\ s = 0.125\ (\text{calibration set})\\
&n = 10,\ r^2 = 0.9362,\ s = 0.200, \overline{R^2_m} = 0.7164, \Delta R^2_m = 0.1108\ {}^cR^2_p\\
&\quad = 0.7006\ (\text{test set})\\
&n = 12,\ r^2 = 0.9454,\ s = 0.404, \overline{R^2_m} = 0.6043, \Delta R^2_m\\
&\quad = -0.1671\ (\text{validation set})
\end{aligned} \tag{3}$$

Table 3
List of molecular features extracted from SMILES and their correlation weights (Split 1).

SA_k	$CW(SA_k)$	N_{TRN}	N_{CLB}	N_{TST}
(..........	0.0	12	14	9
(…Br..(…	1.19150	6	6	6
(…Cl..(…	0.68950	9	9	6
(…O…(…	−0.25100	2	4	1
(…c…(…	1.00100	3	6	3
1..........	−0.50000	14	14	10
1…c…(…	1.25100	7	7	5
C..........	0.0	0	1	1
C…1.......	0.0	0	1	1
Br..(.......	1.25300	6	6	6
Br..........	2.12500	7	7	9
Br..^…1…	1.87100	4	1	0
Br..c…1…	3.43550	5	2	5
Cl..(.......	1.31350	9	9	6
Cl..........	1.18750	13	13	10
Cl..1.......	−1.00300	7	6	4
Cl..^…1…	−0.25100	3	8	7
Cl..^…Cl..	0.0	1	1	0
Cl..c…1…	3.62700	12	13	9
N..........	0.31050	4	4	2
N…^…1…	1.87300	3	2	1
N…^…Cl..	0.0	1	2	1
N…c…1…	−1.12900	4	4	2
O…(.......	−0.18650	2	4	1
O..........	0.62600	4	4	2
O…^…1…	0.0	0	0	1
O…^…Cl..	1.12600	2	0	0
O…c…1…	1.05750	2	1	1
^..........	-1.31250	14	14	10
^…1.......	0.68350	10	11	9
^…Br......	2.12800	4	1	0
^…Cl......	1.50300	7	11	8
^…Cl..1…	1.50100	4	3	1
^…N.......	1.49900	4	4	2
^…O.......	1.00300	2	0	1
c…(.......	0.19050	12	14	9
c…(…Br..	1.49800	6	6	6
c…(…Cl..	1.99800	9	9	6
c…(…O…	−0.25000	2	4	1
c..........	−0.24800	14	14	10
c…1.......	−0.87300	14	14	10
c…1…C…	0.0	0	1	1
c…1…Cl..	−1.43850	7	6	4
c…1…^…	1.99900	10	11	9
c…1…c…	0.49600	14	14	10
c…Br......	3.50300	5	2	5
c…Br..^…	0.18550	4	1	0
c…Cl......	3.55850	12	13	9
c…Cl..^…	2.19150	4	9	7
c…N.......	−0.06750	4	4	2
c…N…^…	1.19250	4	4	2
c…O.......	1.43950	2	1	1
c…O…^…	1.87100	2	0	1
c…c…(…	−0.31750	12	14	9
c…c.......	0.50200	14	14	10
c…c…1…	0.87700	14	14	10
c…c…c…	1.37200	14	13	10

Split 2

$$\begin{aligned}
&pEC_{50} = -0.0014\ (\pm 0.1508947) + 0.1630\ (\pm 0.0063) * \ DCW(3,11)\\
&n = 16,\ r^2 = 0.9400,\ q^2 = 0.8931,\ s = 0.188,\ F = 219\ (\text{Sub-training set})\\
&n = 12,\ r^2 = 0.9606,\ s = 0.137\ (\text{calibration set})\\
&n = 11,\ r^2 = 0.9124,\ s = 0.248, \overline{R_m^2} = 0.8183, \Delta R_m^2 = 0.0903\ {}^cR_p^2 = 0.7344\ (\text{test set})\\
&n = 11,\ r^2 = 0.9616,\ s = 0.191, \overline{R_m^2} = 0.8469, \Delta R_m^2 = 0.0500\ (\text{validation set})
\end{aligned} \tag{4}$$

Split 3

$$\begin{aligned}
&pEC_{50} = \ 1.2380\ (\pm 0.1016) + 0.1296\ (\pm 0.0049) * \ DCW(3,17)\\
&n = 12,\ r^2 = 0.9664,\ q^2 = 0.9357,\ s = 0.108,\ F = 288\ (\text{Sub-training set})\\
&n = 15,\ r^2 = 0.9560,\ s = 0.318\ (\text{calibration set})\\
&n = 10,\ r^2 = 0.9451,\ s = 0.158, \overline{R_m^2} = 0.8507, \Delta R_m^2 = 0.0612\ {}^cR_p^2 = 0.7335\ (\text{test set})\\
&n = 13,\ r^2 = 0.9815,\ s = 0.175, \overline{R_m^2} = 0.7574, \Delta R_m^2 = 0.0677\ (\text{validation set})
\end{aligned} \tag{5}$$

Split 4

$$\begin{aligned}
&pEC_{50} = 0.0012\ (\pm 0.1285) + 0.1205\ (\pm 0.0037) * \ DCW(2,11)\\
&n = 10,\ r^2 = 0.9745,\ q^2 = 0.9529,\ s = 0.103,\ F = 306\ (\text{Sub-training set})\\
&n = 13,\ r^2 = 0.9556,\ s = 0.199\ (\text{calibration set})\\
&n = 10,\ r^2 = 0.9369,\ s = 0.282, \overline{R_m^2} = 0.6158, \Delta R_m^2 = 0.1646\ {}^cR_p^2\\
&\quad = 0.8053\ (\text{test set})\\
&n = 17,\ r^2 = 0.8649,\ s = 0.263, \overline{R_m^2} = 0.7299, \Delta R_m^2 = -0.1399\ (\text{validation set})
\end{aligned} \tag{6}$$

Split 5

$$\begin{aligned}
&pEC_{50} = -0.0144\ (\pm 0.1269) + 0.0969\ (\pm 0.0029) * \ DCW(1,5)\\
&n = 12,\ r^2 = 0.9409,\ q^2 = 0.9105,\ s = 0.197,\ F = 159\ (\text{Sub-training set})\\
&n = 10,\ r^2 = 0.9549,\ s = 0.155\ (\text{calibration set})\\
&n = 14,\ r^2 = 0.8602,\ s = 0.355, \overline{R_m^2} = 0.4693, \Delta R_m^2 = 0.2911\ {}^cR_p^2\\
&\quad = 0.7840\ (\text{test set})\\
&n = 17,\ r^2 = 0.8649,\ s = 0.263, \overline{R_m^2} = 0.7299, \Delta R_m^2 = -0.1399\ (\text{validation set})
\end{aligned} \tag{7}$$

Split 6

$$\begin{aligned}
&pEC_{50} = -0.0053\ (\pm 0.0619) + 0.1306\ (\pm 0.0021) * \ DCW(1,10)\\
&n = 14,\ r^2 = 0.9586,\ q^2 = 0.9454,\ s = 0.129,\ F = 278\ (\text{Sub-training set})\\
&n = 14,\ r^2 = 0.9359,\ s = 0.219\ (\text{calibration set})\\
&n = 11,\ r^2 = 0.9374,\ s = 0.204, \overline{R_m^2} = 0.8834, \Delta R_m^2 = 0.0619\ {}^cR_p^2\\
&\quad = 0.7705\ (\text{test set})\\
&n = 11,\ r^2 = 0.9592,\ s = 0.174, \overline{R_m^2} = 0.7595, \Delta R_m^2 = -0.0798\ (\text{validation set})
\end{aligned} \tag{8}$$

The predictability of models calculated with Eqs. (3)–(8) has been checked with: (i) $\overline{R_m^2}$ (a model has desired predictability if $\overline{R_m^2} > 0.5$ [27–29]); (ii) ΔR_m^2 (a model has desired predictability if $\Delta R_m^2 < 0.2$ [27–29]); and (iii) ${}^cR_p^2$ (this characteristic should be larger than 0.5 [30]) metrics. The only model developed here for split 5 is unsatisfactory, according to these criteria ($\overline{R_m^2}$ and ΔR_m^2 in Eq. (7)). In many cases a QSPR/QSAR analyses are based on one split into the training and test sets. We believe that consideration of a group of splits represents a more informative approach.

Having results of three runs of the Monte Carlo optimization, one can divide the SMILES attributes (which are representation of various molecular features) into three categories: (i) stable promoters of pEC50 increase (correlation weights are positive in the three runs of the Monte Carlo optimization); (ii)) stable promoters of pEC_{50} decrease (correlation weights are negative in the three runs of the optimization); and (iii) attributes which possess an unclear role, since there are both positive and negative correlation weights [31,32]. Our computational experiments show that the presence of chlorine, bromine and oxygen is the promoter of pEC_{50} increase. On the other hand, the presence of nitrogen is the promoter of pEC50 decrease. Thus, the models calculated using Eqs. (3)–(8) have the mechanistic interpretation.

The statistical quality of four-variables model (calculated with involvement of the quantum mechanics descriptors) suggested in the literature [11] for the toxicity of the same 50 binary mixtures is the following: $n = 50$, $r^2 = 0.85$, $s = 0.270$. The models calculated by Eqs (3)–(8) for sets which involve sub-training, calibration, and test set, but without validation set, are characterized by $n = 38$, $r^2 = 0.9498$, $s = 0.156$ (split 1); $n = 39$, $r^2 = 0.9296$, $s = 0.186$ (split 2); $n = 37$, $r^2 = 0.9225$, $s = 0.218$ (split 3); $n = 33$, $r^2 = 0.9369$, $s = 0.197$ (split 4); $n = 36$, $r^2 = 0.8920$, $s = 0.241$ (split 5); $n = 39$, $r^2 = 0.9350$, $s = 0.179$ (split 6). Thus for all cases the CORAL software gives models which are better than the above-mentioned model [11].

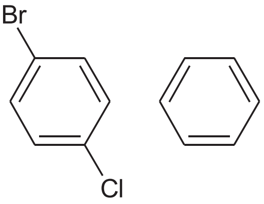
The *supplementary material* section contains details of six splits into the sub-training, calibration, and test sets which are analyzed in this study.

4. Conclusions

We concluded that CORAL can be efficiently used for modeling of the toxicity of binary mixtures. The split into the sub-training, calibration,

Table 4
An example of the DCW(T,N) calculation.
SMILES c1ccccc1.Clc1ccc(Br)cc1.

The molecular structure



SA_k	$CW(SA_k)$
S_k	
c..........	−0.2480
1..........	−0.5000
c..........	−0.2480
c..........	−0.2480
c..........	−0.2480
c..........	−0.2480
c..........	−0.2480
1..........	−0.5000
^..........[a]	−1.3125
Cl.........	1.1875
c..........	−0.2480
1..........	−0.5000
c..........	−0.2480
c..........	−0.2480
c..........	−0.2480
(..........	0.0
Br.........	2.1250
(..........	0.0
c..........	−0.2480
c..........	−0.2480
1..........	−0.5000
SSk	
c…1.......	−0.8730
c…1.......	−0.8730
c…c.......	0.5020
c…c.......	0.5020
c…c.......	0.5020
c…c.......	0.5020
c…1.......	−0.8730
^…1.......	0.6835
^…Cl......	1.5030
c…Cl......	3.5585
c…1.......	−0.8730
c…1.......	−0.8730
c…c.......	0.5020
c…c.......	0.5020
c…(.......	0.1905
Br..(.......	1.2530
Br..(.......	1.2530
c…(.......	0.1905
c…c.......	0.5020
c…1.......	−0.8730
SSS_k	
c…1…c…	0.4960
c…c…1…	0.8770
c…c…c…	1.3720
c…c…c…	1.3720
c…c…c…	1.3720
c…c…1…	0.8770
c…1…^…	1.9990
Cl..^…1…	−0.2510
c…Cl..^…	2.1915
Cl..c…1…	3.6270
c…1…c…	0.4960
c…c…1…	0.8770
c…c…c…	1.3720
c…c…(…	−0.3175
c…(…Br..	1.4980
(…Br..(…	1.1915
c…(…Br..	1.4980
c…c…(…	−0.3175
c…c…1…	0.8770
	$\sum CW(SA_k) = 25.0390$

[a] The dot in SMILES is changed by '^'.

and test sets has apparent influence upon the statistical quality of models calculated with the CORAL software. The CORAL models developed here for toxicity of the binary mixtures have mechanistic interpretations: presence of chlorine, bromine, and oxygen is the promoter of pEC50 increase, whereas the presence of nitrogen is the promoter of pEC50 decrease.

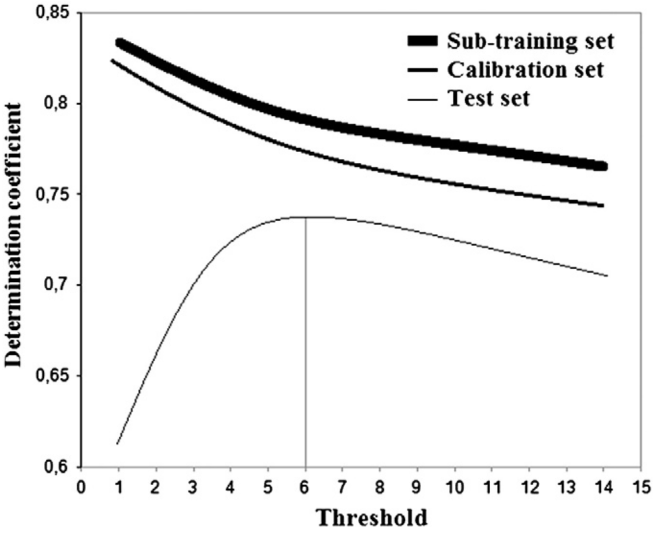


Fig. 1. Determination coefficients of sub-training, calibration, and test sets represented by mathematical functions of the threshold. There is the maximum of the determination coefficient for the external test set.

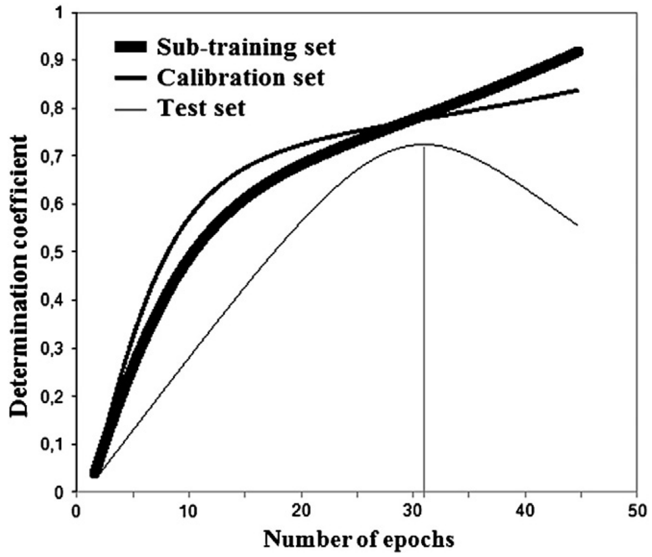


Fig. 2. Determination coefficients of sub-training, calibration, and test sets represented by functions of the number of epochs of the Monte Carlo optimization. There is the maximum of the determination coefficient for the external test set.

Table 5
Statistical characteristics of the model for toxicity of binary mixtures obtained in three runs of the Monte Carlo method optimization for six random splits with preferable threshold (T*) and the number of epochs (N^*_{ep}) which give the best statistical quality for the test set.

			Run 1	Run 2	Run 3	Average
Split	T^a	N^a_{ep}	R^2_{test}[a]	R^2_{test}	R^2_{test}	$\overline{R^2_{test}}$
1	2	10	0.9302	0.9306	0.9157	0.9255 ± 0.0069
2	3	11	0.9397	0.9254	0.9134	0.9262 ± 0.0107
3	3	17	0.9509	0.9497	0.9504	0.9503 ± 0.0005
4	2	11	0.9200	0.8868	0.9325	0.9131 ± 0.0193
5	1	5	0.7489	0.8681	0.8188	0.8119 ± 0.0489
6	1	10	0.9397	0.9350	0.9466	0.9404 ± 0.0048
						0.9112 ± 0.0460

[a] R^2_{test} is the correlation coefficient between DCW(T*,N*) and pEC_{50} for the test set.

Acknowledgments

The authors express their gratitude to ANTARES (the project number LIFE08-ENV/IT/00435) and the NSF CREST Interdisciplinary Nanotoxicity Center NSF-CREST — Grant # HRD-0833178 for financial support. Also, authors express their gratitude to Dr. L. Cappellini, Dr. G. Bianchi and Dr. R. Bagnati for their valuable consultations on computer sciences, and to Dr. J. Baggot for the English revisions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.chemolab.2012.10.001.

References

[1] I. Mitra, A. Saha, K. Roy, Chemometric modeling of free radical scavenging activity of flavone derivatives, Europena Journal of Medicinal Chemistry 45 (2010) 5071–5079.

[2] P.K. Ojha, K. Roy, Chemometric modeling, docking and in silico design of triazolopyrimidine-based dihydroorotate dehydrogenase inhibitors as antimalarials, Europena Journal of Medicinal Chemistry 45 (2010) 4645–4656.

[3] G. Melagraki, A. Afantitis, H. Sarimveis, P.A. Koutentis, G. Kollias, O. Igglessi-Markopoulou, Predictive QSAR workflow for the in silico identification and screening of novel HDAC inhibitors, Molecular Diversity 13 (2009) 301–311.

[4] A. Afantitis, G. Melagraki, P.A. Koutentis, H. Sarimveis, G. Kollias, Ligand-based virtual screening procedure for the prediction and the identification of novel β-amyloid aggregation inhibitors using Kohonen maps and Counterpropagation Artificial Neural Networks, Europena Journal of Medicinal Chemistry 46 (2011) 497–508.

[5] J. García, P.R. Duchowicz, M.F. Rozas, J.A. Caram, M.V. Mirífico, F.M. Fernández, E.A. Castro, A comparative QSAR on 1,2,5-thiadiazolidin-3-one 1,1-dioxide compounds as selective inhibitors of human serine proteinases, Journal of Molecular Graphics & Modelling 31 (2011) 10–19.

[6] J.C. Garro Martinez, P.R. Duchowicz, M.R. Estrada, G.N. Zamarbide, E.A. Castro, QSAR Study and molecular design of open-chain enaminones as anticonvulsant agents, International Journal of Molecular Sciences 12 (2011) 9354–9368.

[7] E. Ibezim, P.R. Duchowicz, E.V. Ortiz, E.A. Castro, QSAR on aryl-piperazine derivatives with activity on malaria, Chemometrics and Intelligent Laboratory Systems 110 (2012) 81–88.

[8] L.M.A. Mullen, P.R. Duchowicz, E.A. Castro, QSAR treatment on a new class of triphenylmethyl-containing compounds as potent anticancer agents, Chemometrics and Intelligent Laboratory Systems 107 (2011) 269–275.

[9] K. Roy, I. Mitra, On the use of the metric r^2_m as an effective tool for validation of QSAR models in computational drug design and predictive toxicology, Mini-Review in Medicinal Chemistry 12 (2012) 491–504.

[10] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, QSAR modelling toxicity toward rats of inorganic substances by means of CORAL, Central European Journal of Chemistry 9 (2011) 75–85.

[11] L. Zhang, P. Zhou, F. Yang, Z. Wang, Computer-based QSARs for predicting mixture toxicity of benzene and its derivatives, Chemosphere 67 (2007) 396–401.

[12] P.R. Duchowicz, M.G. Vitale, E.A. Castro, Partial Order Ranking for the aqueous toxicity of aromatic mixtures, Journal of Mathematical Chemistry 44 (2008) 541–549.

[13] R. Altenburger, M. Nendza, G. Schuurmann, Mixture toxicity and its modeling by quantitative structure–activity relationships, Environmental Toxicology and Chemistry 22 (2003) 1900–1915.

[14] M. Góral, A. Bok, T. Kasprzycka-Gutman, P. Oracz, Recommended vapor-liquid equilibrium data. Part 4. Binary alkanol-alkene/alkyne systems, Journal of Physical and Chemical Reference Data 35 (2006) 1577–1596.

[15] A.P. Toropova, A.A. Toropov, R. Gonella Diaza, E. Benfenati, G. Gini, Analysis of the co-evolutions of correlations as a tool for QSAR-modeling carcinogenicity: an unexpected good prediction based on a model that seems untrustworthy, Central European Journal of Chemistry 9 (2011) 165–174.

[16] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, Co-evolutions of correlations for QSAR of toxicity of organometallic and inorganic substances: an unexpected good prediction based on a model that seems untrustworthy, Chemometrics and Intelligent Laboratory Systems 105 (2011) 215–219.

[17] A.A. Toropov, A.P. Toropova, A. Lombardo, A. Roncaglioni, E. Benfenati, G. Gini, CORAL: building up the model for bioconcentration factor and defining it's applicability domain, European Journal of Medicinal Chemistry 46 (2011) 1400–1403.

[18] A.A. Toropov, A.P. Toropova, E. Benfenati, QSAR modelling of the toxicity to *Tetrahymena pyriformis* by balance of correlations, Molecular Diversity 14 (2010) 821–827.

[19] http://www.insilico.eu/coral.

[20] http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html.

[21] A.A. Toropov, B.F. Rasulev, J. Leszczynski, QSAR modeling of acute toxicity by balance of correlations, Bioorganic & Medicinal Chemistry 16 (2008) 5999–6008.

[22] A.A. Toropov, A.P. Toropova, I. Gutman, Comparison of QSPR models based on hydrogen-filled graphs and on graphs of atomic orbitals, Croatica Chemica Acta 78 (2005) 503–509.

[23] A.A. Toropov, A.P. Toropova, E. Benfenati, D. Leszczynska, J. Leszczynski, QSAR analysis of 1,4-dihydro-4-oxo-1-(2-thiazolyl)-1,8-naphthyridines exhibiting anticancer activity by optimal SMILES-based descriptors, Journal of Mathematical Chemistry 47 (2010) 647–666.

[24] A.A. Toropov, A.P. Toropova, E. Benfenati, D. Leszczynska, J. Leszczynski, Additive InChI-based optimal descriptors: QSPR modeling of fullerene C 60 solubility in organic solvents, Journal of Mathematical Chemistry 46 (2009) 1232–1251.

[25] A.A. Toropov, A.P. Toropova, E. Benfenati, D. Leszczynska, J. Leszczynski, Use of the international chemical identifier for constructing QSPR-model of normal boiling points of acyclic carbonyl substances, Journal of Mathematical Chemistry 47 (2009) 355–369.

[26] A.A. Toropov, B.F. Rasulev, D. Leszczynska, J. Leszczynski, Multiplicative SMILES-based optimal descriptors: QSPR modeling of fullerene C60 solubility in organic solvents, Chemical Physics Letters 457 (2008) 332–336.

[27] K. Roy, I. Mitra, S. Kar, P.K. Ojha, R.N. Das, H. Kabir, Comparative studies on some metrics for external validation of QSPR models, Journal of Chemical Information and Modeling 52 (2012) 396–408.

[28] P.K. Ojha, I. Mitra, R.N. Das, K. Roy, Further exploring rm 2 metrics for validation of QSPR models, Chemometrics and Intelligent Laboratory Systems 107 (2011) 194–205.

[29] http://203.200.173.43:8080/rmsquare/.

[30] P.K. Ojha, K. Roy, Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection, Chemometrics and Intelligent Laboratory Systems 109 (2011) 146–161.

[31] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, QSAR models for toxicity of organic substances to *Daphnia magna* built up by using the CORAL freeware, Chemical Biology & Drug Design 79 (2012) 332–338.

[32] A.P. Toropova, A.A. Toropov, S.E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, CORAL: QSAR modeling of toxicity of organic chemicals towards *Daphnia magna*, Chemometrics and Intelligent Laboratory Systems 110 (2012) 177–181.