Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 16 (2008) 4801–4809
Bioorganic & Medicinal Chemistry

Additive SMILES-based optimal descriptors in QSAR modelling bee toxicity: Using rare SMILES attributes to define the applicability domain

A. A. Toropov[a,*] and E. Benfenati[b]

[a]*Institute of Geology and Geophysics, Khodzhibaev St. 49, Tashkent 100041, Uzbekistan*
[b]*Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, Milano 20156, Italy*

Received 25 February 2008; accepted 20 March 2008
Available online 23 March 2008

Abstract—The additive SMILES-based optimal descriptors have been used for modelling the bee toxicity. The influence of relative prevalence of the SMILES attributes in a training and test sets to the models for bee toxicity has been analysed. Avoiding the use of rare attributes improves statistical characteristics of the model on the external test set. The possibility of using the probability of the presence of SMILES attributes in training and test sets for rational definition of the applicability domain is discussed.

1. Introduction

Quantitative structure-property/activity relationship (QSPR/QSAR) is a tool for the estimation of unavailable numerical data on endpoints of interest by means of correlations 'descriptor-property/activity'.[1] The descriptor is a numerical index of the molecular structure that can be calculated using information on the molecular architecture, for instance, represented by molecular graphs.[2–5] As an alternative to the graph, the simplified molecular input line entry system (SMILES) can be used for the elucidation of the molecular structure in the QSPR/QSAR analyses.[6–8]

The toxicity of pesticides towards bee is an important ecologic indicator, since the bees have influence to many natural processes related to fruit trees.[9] The experimental definition of the numerical values of the toxicity towards bee involves considerable time and resources.[10] Thus there are motivations to search for robust models of the toxicity towards bee.[9–12]

The number of SMILES-oriented databases in the internet is gradually increasing. SMILES-based optimal descriptors gave reasonable prediction for the bee toxicity.[12] The aim of the present study is an attempt to use more transparent version of the SMILES-based descriptor. In fact, each SMILES attribute is a representation of molecular fragment. The additive scheme[13] is a well-known approach of modelling properties by the selection of special contributions for molecular fragments. The new version of the SMILES-based optimal descriptors is a SMILES-based realization of the additive scheme.

2. Results and discussion

Figure 1 shows the plot of ARP versus Lim S. One can see from Figure 1 that the ARP and Lim S are increasing. However for Lim $S = 4$ the ARP is unexpectedly decreased. This an interesting point probably it can be an indicator for a robust selection of the Lim S. The hypothesis is that the ARP can be used for a preliminary estimation of the split into training and test sets as well as for estimation of the Lim S.

Figure 2 shows the correlation coefficients for the training and test sets for different Lim S. One can see from Figure 2 that the best statistical characteristics of the model for the external test set take place for Lim $S = 4$. Table 1 contains the statistical characteristics of the models for the bee toxicity over Lim S from 1 to 10 (see Figs. 1 and 2).

The first probe of the Monte Carlo optimization with Lim $S = 4$ gave the model (see Figs. 3 and 4).

Keywords: QSAR; SMILES; Toxicity towards bee; Applicability domain.
* Corresponding author. E-mail: aatoropov@yahoo.com

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.03.048

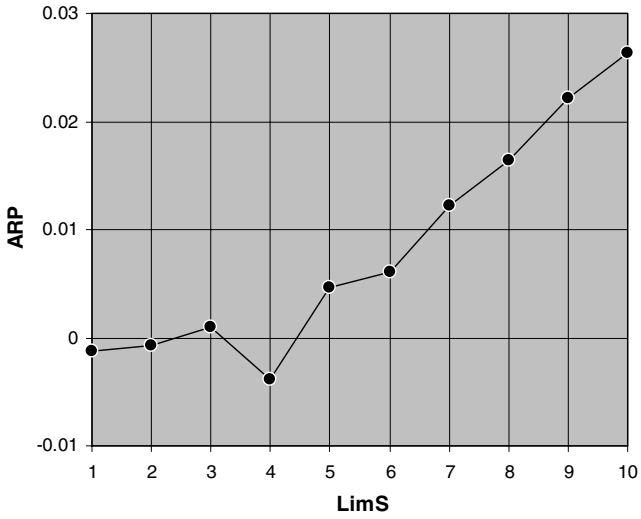

Figure 1. Plot of Lim S versus average relative prevalence (ARP).

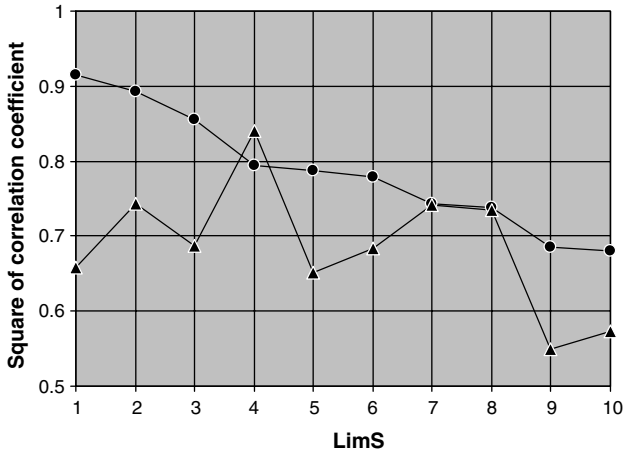

Figure 2. Plot of the correlation coefficients R^2 for training (circles) and test (triangles) sets versus the Lim S.

$$\mathrm{Log}(1/C) = -3.460(\pm 0.027) + 0.201(\pm 0.001)\,\mathrm{DCW} \quad (1)$$

$$n = 85,\ r^2 = 0.793,\ s = 0.65,\ F = 318\ (\text{training set})$$

$$n = 20,\ r^2 = 0.838,\ s = 0.61,\ F = 93\ (\text{test set})$$

Table 2 contains the correlation weights for the DCW calculation obtained in three probes of the Monte Carlo optimization. Table 3 shows an example of the DCW calculation. Table 4 contains experimental and calculated with Eq. 1 values of the toxicity towards bee.

Correlation weights for the SMILES attributes (Table 2) have similar numerical values over the three probes of the Monte Carlo optimization. These have a transparent interpretation: $\mathrm{CW}(x) < 0$ means that the x is related to the decrease of the bee toxicity, and vice versa $\mathrm{CW}(x) > 0$ means that the x is related to the increase of the bee toxicity.

According to Ref. 14 each version of the SMILES software produces different SMILES notations. Thus using of a mixture of SMILES generated by different software can produce results not reproducible. However using SMILES generated by a single software (e.g., ChemSketch[15]) can provide reasonable QSPR/QSAR prediction in general and the QSAR prediction of the bee toxicity in particular.

In a previous study[12] we obtained the follows statistical characteristics of the bee toxicity: $n = 85$, $r^2 = 0.68$, $s = 0.82$, $F = 180$ (training set) and $n = 20$, $r^2 = 0.72$, $s = 0.68$, $F = 46$. Thus the statistical characteristics of the model calculated with Eq. 1 are better.

Instead of the SMILES attributes which containing two components, one can use attributes which are containing one or three components. However, our experiments have shown that both these models are worst in comparison with the model of two components, at least for the case of the bee toxicity. In the first case (i.e., one component) correlations are low. In the second case (i.e., three components) overtraining takes place. These results have not been included in Table 1.

Also, one can attempt to take into account a numbers of other chemical elements in the molecular structure (e.g., oxygen, phosphorus, sulfur, etc.). However, our experiments have shown that taking into account a number of other chemical elements does not improve this model. More exactly, taking into account the numbers of the oxygen, phosphorus and sulfur improves the model for the training set, but the statistical characteristics for the test set are considerably lower (overtraining). Most probably this situation is exclusive for the structures of these pesticides and for the bee toxicity. It is probable that for other substances and other endpoints, the numbers of the oxygen, phosphorus and sulfur may become more useful.

The difference between the present study and what described in Ref. 12 is on the use of SMILES attributes with two components (and not one as in the previous one) and the Lim S index instead of the Lim N. It is to be also noted we have used the additive scheme instead of the multiplicative scheme. For bee toxicity these modifications gave significative improvements of the model. Currently, it is not apparent whether these modifications are useful for other endpoints and other substances. We are planning to study this matter in the near future.

Thus, the proposed method improves the previous model and allows user to assess if the method is appropriate for a certain new compound or not; for a new compound what has to be checked is if, using Eqs. 2 and 1, the SMILES of the substance (prepared by ChemSketch[15]) does not contain rare SMILES attributes.

3. Conclusion

The additive SMILES-based optimal descriptors can be used as a tool for predicting the values of the bee toxicity. The suggested Lim S index can be used as a tool for the selection of the list of the SMILES attributes for the robust SMILES-based model.

Table 1. Statistical characteristics of the models for the bee toxicity with different Lim S for the training (n = 85) and test (n = 20) sets

Lim S	Nact	Probe	r^2_t	s_t	F_t	r^2_v	s_v	F_v	ARP
1	127	1	0.9162	0.4117	434	0.6768	0.9793	15	−0.00118
		2	0.9129	0.4197	415	0.6581	0.9943	14	
		3	0.9153	0.4140	429	0.6420	1.0001	13	
		Average	0.9148	0.4151	426	0.6590	0.9913	14	
2	96	1	0.8960	0.4587	338	0.7300	0.9242	21	−0.00067
		2	0.8879	0.4762	309	0.7453	0.9408	22	
		3	0.8933	0.4646	328	0.7552	0.9650	24	
		Average	0.8924	0.4665	325	0.7435	0.9433	22	
3	83	1	0.8551	0.5414	226	0.6987	0.8029	17	0.00096
		2	0.8561	0.5395	228	0.6911	0.8506	16	
		3	0.8563	0.5391	228	0.6718	0.8618	15	
		Average	0.8559	0.5400	227	0.6872	0.8384	16	
4	**69**	**1**	**0.7928**	**0.6475**	**140**	**0.8380**	**0.6111**	**42**	**−0.00384**
		2	0.7962	0.6422	144	**0.8426**	0.5976	44	
		3	0.7951	0.6439	143	**0.8384**	0.6208	43	
		Average	0.7947	0.6445	142	**0.8397**	0.6098	43	
5	64	1	0.7883	0.6545	136	0.6571	1.0816	14	0.00469
		2	0.7873	0.6561	135	0.6655	1.0683	14	
		3	0.7871	0.6563	135	0.6320	1.0701	12	
		Average	0.7875	0.6556	136	0.6516	1.0733	13	
6	61	1	0.7760	0.6732	126	0.6819	1.0253	16	0.00612
		2	0.7830	0.6626	132	0.6742	1.0646	15	
		3	0.7752	0.6745	125	0.6943	1.0141	17	
		Average	0.7781	0.6701	127	0.6835	1.0347	16	
7	58	1	0.7442	0.7194	103	0.7414	0.7278	22	0.01227
		2	0.7436	0.7203	103	0.7391	0.7323	22	
		3	0.7423	0.7221	102	0.7430	0.7266	22	
		Average	0.7433	0.7206	103	0.7412	0.7289	22	
8	54	1	0.7398	0.7256	100	0.7316	0.7321	21	0.01634
		2	0.7391	0.7265	100	0.7377	0.7190	21	
		3	0.7368	0.7297	99	0.7361	0.7303	21	
		Average	0.7386	0.7273	100	0.7351	0.7271	21	
9	50	1	0.6865	0.7965	74	0.5500	0.9434	8	0.02212
		2	0.6845	0.7990	73	0.5268	0.9642	7	
		3	0.6833	0.8005	73	0.5723	0.9148	9	
		Average	0.6847	0.7987	73	0.5497	0.9408	8	
10	47	1	0.6816	0.8026	72	0.5552	0.9405	8	0.02635
		2	0.6789	0.8060	71	0.5990	0.8985	10	
		3	0.6812	0.8032	72	0.5645	0.9308	8	
		Average	0.6806	0.8039	72	0.5729	0.9233	9	

Nact is number of not blocked SMILES attributes. R, s, and F are correlation coefficient, standard error of estimation, and F-ratio Fischer, respectively. Letters t and v are used to indicate training (t) and test (v) sets. Total number of the attributes is 133. Values in bold are for the best predictions.

4. Method

As endpoint we used the decimal logarithm log(1/C), where C is the concentration of the pesticides expressed in mmol/bee, which kills 50% of the bees.[12] These 105 pesticides have been split into training set (n = 85) and test set (n = 20).

Optimal descriptors examined in the present study have been defined as

$$\begin{aligned} \mathrm{DCW} = {} & \mathrm{CW(b)} + \mathrm{CW(db)} + \mathrm{CW(tb)} + \mathrm{CW(N)} \\ & + \mathrm{CW(Cl)} + \mathrm{CW(Br)} + \mathrm{CW(F)} \\ & + \sum \mathrm{CW(SS_k)} \end{aligned} \quad (2)$$

where CW(b) is the correlation weight of the given number of branching (i.e., number of brackets in the SMILES); CW(db) is the correlation weight of the given number of double bonds (i.e., number of '='); CW(tb) is the correlation weight of the number of triple bonds (i.e., number of '#'); CW(N), CW(Cl), CW(Br), and CW(F) are correlation weights for given numbers of nitrogen, chlorine, bromine, and fluorine atoms, respectively; CW(SS_k) is the correlation weight of two components SMILES fragment. The following SMILES components contain two characters: 'Cl', 'Br', 'C=', 'N=', 'O=', 'S=', 'N+', and 'n+'. All other SMILES components contain one character. The sequence of the two components SMILES fragment is an association of the components by the scheme

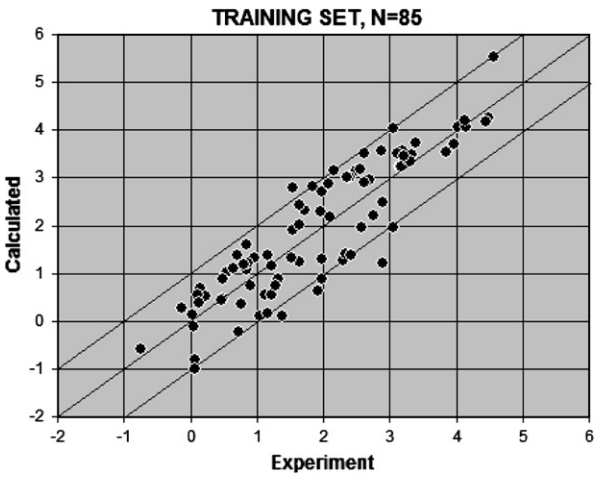

Figure 3. Correlation between experimental and calculated with Eq. 1 values of the bee toxicity for the training set.

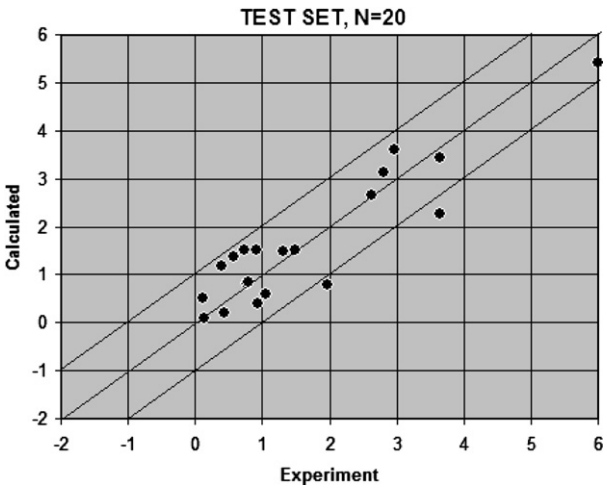

Figure 4. Correlation between experimental and calculated with Eq. 1 values of the bee toxicity for the test set.

Table 2. Correlation weights over three probes of the Monte Carlo optimization and other characteristics of the SMILES attributes

SMILES attributes	CW in probe 1	CW in probe 2	CW in probe 3	NT	NV	NTS	NVS
tb							
#000_____	4.002	3.898	4.004	73	19	73	19
#001_____	5.673	5.560	5.974	11	1	11	1
#002_____	0.0	0.0	0.0	1	0	1	0
b							
(000_____	0.0	0.0	0.0	1	0	1	0
(001_____	0.0	0.0	0.0	1	1	1	1
(002_____	2.499	2.451	2.309	9	0	9	0
(003_____	0.397	0.498	0.411	16	3	16	3
(004_____	0.566	0.578	0.503	22	6	22	6
(005_____	1.748	1.791	1.425	14	2	14	2
(006_____	1.004	1.085	0.693	11	2	11	2
(007_____	1.297	1.303	1.300	4	4	4	4
(008_____	0.0	0.0	0.0	3	1	3	1
(009_____	0.0	0.0	0.0	2	0	2	0
(010_____	0.0	0.0	0.0	1	0	1	0
(012_____	0.0	0.0	0.0	0	1	0	1
(016_____	0.0	0.0	0.0	1	0	1	0
db							
=000_____	3.097	3.424	3.090	10	4	10	4
=001_____	2.454	2.690	2.397	29	6	29	6
=002_____	1.948	2.102	1.914	29	3	29	3
=003_____	4.352	4.753	4.573	11	2	11	2
=004_____	3.521	3.995	3.653	4	5	4	5
=005_____	0.0	0.0	0.0	2	0	2	0
F							
F000_____	2.458	1.754	2.323	77	19	77	19
F001_____	0.0	0.0	0.0	1	0	1	0
F003_____	4.800	4.598	5.074	5	1	5	1
F005_____	0.0	0.0	0.0	1	0	1	0
F006_____	0.0	0.0	0.0	1	0	1	0
Br							
Br00_____	1.379	1.324	1.199	79	19	79	19
Br01_____	0.0	0.0	0.0	2	1	2	1
Br02_____	0.0	0.0	0.0	3	0	3	0
Br04_____	0.0	0.0	0.0	1	0	1	0
Cl							
Cl00_____	4.429	4.573	4.348	58	11	58	11
Cl01_____	3.526	3.689	3.584	13	3	13	3
Cl02_____	4.304	4.003	4.050	5	3	5	3

Table 2 (*continued*)

SMILES attributes	CW in probe 1	CW in probe 2	CW in probe 3	NT	NV	NTS	NVS
Cl03_____	0.0	0.0	0.0	3	1	3	1
Cl04_____	0.0	0.0	0.0	3	0	3	0
Cl05_____	0.0	0.0	0.0	1	0	1	0
Cl06_____	0.0	0.0	0.0	2	2	2	2
N							
N000_____	2.154	2.249	2.324	28	6	28	6
N001_____	4.674	4.826	4.875	34	3	34	3
N002_____	4.703	4.787	4.703	12	8	12	8
N003_____	2.172	2.276	2.127	8	3	8	3
N004_____	0.0	0.0	0.0	3	0	3	0
SS_k							
(___(_____	−0.696	−0.683	−0.896	66	15	49	13
1___(_____	0.815	0.747	0.725	66	19	44	12
2___(_____	−0.801	−0.802	−0.797	12	2	12	2
2___1_____	0.0	0.0	0.0	1	2	1	2
3___(_____	0.0	0.0	0.0	3	0	3	0
4___(_____	0.0	0.0	0.0	0	2	0	2
C=___(_____	−2.002	−2.000	−2.001	16	7	11	5
C=___1_____	4.996	5.183	5.177	12	2	11	2
C=___2_____	0.0	0.0	0.0	2	1	2	1
C=___3_____	0.0	0.0	0.0	1	0	1	0
C___#_____	−0.196	−0.200	−0.196	15	1	12	1
C___(_____	0.112	0.103	0.153	545	143	80	19
C___1_____	0.091	0.099	0.146	73	17	31	5
C___2_____	1.275	1.224	1.260	14	15	6	3
C___3_____	0.0	0.0	0.0	1	12	1	3
C___4_____	0.0	0.0	0.0	0	4	0	2
C___C=_____	3.077	3.199	3.101	14	3	11	2
C___C_____	−0.204	−0.200	−0.185	112	34	48	8
Br___(_____	1.490	1.537	1.511	21	2	6	1
Br___1_____	0.0	0.0	0.0	1	0	1	0
Cl___(_____	1.076	1.105	1.126	103	36	25	9
Cl___1_____	4.800	4.688	4.497	4	3	4	3
Cl___3_____	0.0	0.0	0.0	1	0	1	0
Cl___C=_____	0.0	0.0	0.0	1	0	1	0
Cl___C_____	−1.153	−1.133	−1.478	4	3	4	3
F___(_____	0.062	−0.002	0.039	50	5	8	1
F___C_____	0.0	0.0	0.0	1	0	1	0
N=___(_____	1.598	1.124	1.328	7	1	7	1
N=___1_____	0.0	0.0	0.0	3	0	3	0
N=___C_____	6.400	6.525	6.538	8	1	8	1
N=___N=_____	0.0	0.0	0.0	2	0	2	0
O=___(_____	−0.045	−0.040	−0.113	91	32	37	7
O=___1_____	0.0	0.0	0.0	3	0	3	0
O=___C_____	−0.376	−0.312	−0.368	35	5	35	5
N___#_____	0.097	−0.149	−0.200	11	1	10	1
N___(_____	−0.976	−0.913	−0.965	67	27	40	12
N___1_____	−1.411	−1.413	−1.334	14	1	11	1
N___C_____	0.750	0.797	1.001	23	10	18	8
N___N=_____	0.0	0.0	0.0	2	0	2	0
N___N_____	0.0	0.0	0.0	2	0	2	0
O___(_____	0.202	0.104	0.202	118	30	56	13
O___1_____	1.840	1.898	1.796	5	2	4	1
O___2_____	0.0	0.0	0.0	1	0	1	0
O___4_____	0.0	0.0	0.0	0	2	0	2
O___C=_____	0.0	0.0	0.0	0	1	0	1
O___C_____	1.489	1.517	1.521	86	24	48	11
O___N=_____	0.0	0.0	0.0	3	1	3	1
S=___(_____	0.0	0.0	0.0	4	2	2	1
S=___C_____	0.0	0.0	0.0	1	1	1	1
P___(_____	0.097	0.102	0.145	23	3	22	3
P___O=_____	3.099	3.202	3.403	9	2	9	2
P___S=_____	4.925	5.087	5.104	13	1	13	1
S___(_____	0.704	0.718	0.638	29	7	24	7

(*continued on next page*)

Table 2 (*continued*)

SMILES attributes	CW in probe 1	CW in probe 2	CW in probe 3	NT	NV	NTS	NVS
S___1_____	0.0	0.0	0.0	4	0	3	0
S___C_____	0.479	0.399	0.453	29	5	23	5
S___O=_____	0.0	0.0	0.0	2	3	2	3
S___N_____	0.0	0.0	0.0	1	0	1	0
S___O_____	0.0	0.0	0.0	1	0	1	0
S___P_____	0.0	0.0	0.0	1	0	1	0
S___S_____	0.0	0.0	0.0	0	1	0	1
[___(_____	0.252	0.252	0.303	29	9	10	2
[___1_____	0.0	0.0	0.0	1	0	1	0
[___2_____	0.0	0.0	0.0	1	0	1	0
[___3_____	0.0	0.0	0.0	1	0	1	0
[___C_____	0.0	0.0	0.0	1	0	1	0
[___N+____	0.016	0.216	−0.111	20	6	7	2
[___O-____	0.565	0.360	0.578	20	6	7	2
[___O_____	0.0	0.0	0.0	3	0	3	0
[___S_____	0.0	0.0	0.0	4	0	3	0
[___[_____	4.303	4.280	4.599	5	1	5	1
c___(_____	−0.297	−0.253	−0.270	219	46	55	12
c___1_____	0.254	0.297	0.363	220	40	52	12
c___2_____	1.325	1.392	1.371	33	0	12	0
c___3_____	0.0	0.0	0.0	5	0	2	0
c___C_____	2.204	2.199	2.200	10	0	10	0
c___N=____	0.0	0.0	0.0	1	0	1	0
c___N_____	−1.510	−1.510	−1.374	7	5	7	5
c___O_____	4.080	4.075	3.979	25	3	24	3
c___S_____	0.0	0.0	0.0	2	1	2	1
c___[_____	0.0	0.0	0.0	1	0	1	0
c___c_____	−0.387	−0.435	−0.450	225	33	49	9
n+___[_____	0.0	0.0	0.0	4	0	1	0
n___(_____	3.471	3.338	3.255	12	12	9	7
n___1_____	−2.005	−2.004	−2.004	7	14	6	7
n___2_____	0.0	0.0	0.0	2	0	2	0
n___S_____	0.0	0.0	0.0	4	0	3	0
n___[_____	0.0	0.0	0.0	4	0	3	0
n___c_____	−1.590	−1.592	−1.504	13	12	8	7
s___(_____	0.0	0.0	0.0	1	1	1	1
s___c_____	0.0	0.0	0.0	1	1	1	1

NT and NV are numbers of a given attribute in the training and test sets, respectively; NTS and NVS are numbers of the SMILES containing a given attribute in the training and test sets, respectively.

$$\mathrm{ABCDE}\cdots = \text{'AB'} + \text{'BC'} + \text{'CD'} + \text{'DE'} + \cdots$$

It is possible that a molecular fragment can be represented by two different forms: that is, 'AB' and 'BA'. It can lead to the existence of two different correlation weights, CW('AB') ≠ CW('BA') for the same molecular fragment. In order to avoid this situation each two components SMILES fragment is ranged according to ASCII codes of its characters.

The SS_k are local SMILES attributes (i.e., these are characterizing a fragment of molecule). Other SMILES attributes are global, because they are characterizing molecules in whole.

Numerical data for the correlation weights can be obtained by the Monte Carlo method optimization[2,3,12] with the correlation coefficient between the DCW and toxicity as target function (in other words it is the calculation of the correlation weights which produce a maximum of the correlation coefficient).

The numbers of the local and global SMILES attributes in a training set can be from zero till thousands. Using correlation weights of rare attributes can lead to overtraining. One can block the influence of rare attributes[12] by fixing their correlation weights to zero if the total number of the analysed attribute in the training set is less than an assigned number ($\mathrm{Lim}\,N$).

There is another possibility to block a SMILES attribute if the number of SMILES which are containing this attribute is less than an assigned number ($\mathrm{Lim}\,S$). For example, if the total number of rare attribute in the training set is three, it is possible that these are present in a single structure, and the $\mathrm{Lim}\,S$ keeps into account this situation. The $\mathrm{Lim}\,S$ is used in this study.

The probability of finding a SMILES attribute in a given SMILES of training (test) set is an informative characteristic of the attribute. The preferable situation if this probability is not zero and one: zero indicates the ab-

Table 3. Example of the DCW calculation with correlation weights of the first probe (Lim $S = 4$)

SMILES attributes	CW in probe 1 Lim $S = 4$
O=___C_____	−0.376
C___(_____	0.112
N___(_____	−0.976
N___(_____	−0.976
c___(_____	−0.297
c___1_____	0.254
c___1_____	0.254
c___(_____	−0.297
c___(_____	−0.297
c___c_____	−0.387
c___c_____	−0.387
c___c_____	−0.387
c___1_____	0.254
C___1_____	0.091
C___C_____	−0.204
C___(_____	0.112
C___(_____	0.112
C___(_____	0.112
C___(_____	0.112
O___C_____	1.489
O___C_____	1.489
C___C_____	−0.204
C___(_____	0.112
C___(_____	0.112
Cl___C_____	−1.153
(003_____	0.397
=001_____	2.454
#000_____	4.002
N001_____	4.674
Cl01_____	3.526
Br00_____	1.379
F000_____	2.458

SMILES = 'O=C(N(c1c(cccc1CC)C)COCC)CCl'; CAS = 34256-82-1 DCW = 17.564.

sence of the attribute; one indicates that the attribute takes place in each SMILES.

The difference between the mentioned probabilities (for the training set and the test set) is the measure of relative prevalence of the attribute in training and test sets. The average relative prevalence (ARP) overall the active (not blocked) attributes may be a criterion for the estimation of split into a training and a test: the ideal situation is when the difference is zero:

$$\mathrm{ARP} = (1/\mathrm{Nact}) \sum_{k} [P_{\mathrm{t}}(A_{\mathrm{k}}) - P_{\mathrm{v}}(A_{\mathrm{k}})]$$

where Nact is the number of active (i.e., not blocked) SMILES attributes; $P_{\mathrm{t}}(A_{\mathrm{k}})$ and $P_{\mathrm{v}}(A_{\mathrm{k}})$ are the probability of the kth SMILES attribute presence in SMILES notations in the training and a test sets, respectively.

If this difference is more than zero then the majority of attributes is in the training set and thus overtraining becomes possible. If this difference is less than zero then deficit of information for this model becomes possible. This difference is a mathematical function of the Lim S.

Under such circumstances, one can suppose, that the test set is adequate (i.e., the test set is a subset of the domain of applicability for this SMILES-based model) if the test set is satisfying the following conditions: (i) the test set does not contain attributes which are rare in the training set; (ii) the value of average relative prevalence is approximately equals to zero.

Table 4. Bee toxicity values, log(1/C): experimental and calculated with Eq. 2

CAS	SMILES	DCW	Expr	Calc	Expr − Calc
	Training set				
34256-82-1	O=C(N(c1c(cccc1CC)C)COCC)CCl	17.564	−0.803	0.070	−0.873
116-06-3	O=C(ON=CC(SC)(C)C)NC	26.383	2.825	1.843	0.982
834-12-8	CSc1nc(nc(n1)NC(C)C)NCC	21.075	0.357	0.776	−0.419
741-58-2	S=P(OC(C)C)(OC(C)C)SCCNS(=O)(=O)c1ccccc1	31.600	1.219	2.891	−1.672
68359-37-5	N#CC(c1cc(c(cc1)F)Oc1ccccc1)OC(=O)C1C(C1C=C(Cl)Cl)(C)C	37.288	4.070	4.034	0.036
314-40-9	O=C1N(C(=O)C(=C(N1)C)Br)C(CC)C	17.362	0.130	0.029	0.101
1689-84-5	N#Cc1cc(c(c(c1)Br)O)Br	27.115	1.281	1.990	−0.709
1689-99-2	N#Cc1cc(c(c(c1)Br)OC(=O)CCCCCCC)Br	26.950	2.304	1.957	0.347
63-25-2	O=C(Oc1c2c(ccc1)cccc2)NC	27.612	2.190	2.090	0.100
1563-66-2	O=C(Oc1cccc2c1OC(C2)(C)C)NC	29.615	3.141	2.492	0.649
5234-68-4	O=C(C1=C(OCCS1)C)Nc1ccccc1	24.055	0.114	1.375	−1.261
2439-01-2	O=C1Sc2c(nc3c(n2)ccc(c3)C)S1	22.858	0.547	1.134	−0.587
54593-83-8	S=P(OC(C(Cl)(Cl)Cl)Cl)(OCC)OCC	31.546	3.572	2.880	0.692
13121-70-5	O[Sn](C1CCCCC1)(C1CCCCC1)C1CCCCC1	19.911	1.031	0.542	0.489
1596-84-5	O=C(NN(C)C)CCC(=O)O	17.476	−0.108	0.052	−0.160
52918-63-5	N#CC(c1cc(ccc1)Oc1ccccc1)OC(=O)C1C(C1C=C(Br)Br)(C)C	39.938	5.527	4.567	0.960
87674-68-8	O=C(N(c1c(scc1C)C)C(COC)C)CCl	19.660	0.468	0.491	−0.023
60-51-5	S=P(SCC(=O)NC)(OC)OC	27.994	3.156	2.166	0.990
97886-45-8	O=C(c1c(nc(c(c1CC(C)C)C(=O)SC)C(F)(F)F)C(F)F)SC	17.940	0.695	0.146	0.549
563-12-2	S=P(SCSP(=S)(OCC)OCC)(OCC)OCC	28.609	1.272	2.290	−1.018
13356-08-6	CC(c1ccccc1)(C[Sn](O[Sn](CC(c1ccccc1)(C)C)(CC(c1ccccc1)(C)C)CC(c1ccccc1)(C)C)(CC(c1ccccc1)(C)C)CC(c1ccccc1)(C)C)C	13.521	−0.578	−0.743	0.165
66441-23-4	O=C(C(Oc1ccc(cc1)OC1=Nc2ccc(cc2O1)Cl)C)OCC	33.137	3.558	3.200	0.358

(*continued on next page*)

Table 4 (*continued*)

CAS	SMILES	DCW	Expr	Calc	Expr − Calc
55-38-9	S=P(Oc1cc(c(cc1)SC)C)(OC)OC	30.678	2.956	2.706	0.250
76-87-9	O[Sn](c1ccccc1)(c1ccccc1)c1ccccc1	17.955	0.505	0.149	0.356
133-07-3	O=C1N(C(=O)c2c1cccc2)SC(Cl)(Cl)Cl	20.692	1.389	0.699	0.690
23422-53-9	O=C(Oc1cc(ccc1)N=CN(C)C)NC	28.619	1.257	2.292	−1.035
58-89-9	ClC1C(C(C(C(C1Cl)Cl)Cl)Cl)Cl	27.080	2.715	1.983	0.732
67485-29-4	FC(c1ccc(cc1)C=CC(=NN=C1NCC(CN1)(C)C)C=Cc1ccc(cc1)C(F)(F)F)(F)F	23.796	0.868	1.323	−0.455
121-75-5	S=P(SC(CC(=O)OCC)C(=O)OCC)(OC)OC	32.899	3.218	3.152	0.066
16752-77-5	O=C(ON=C(SC)C)NC	29.161	3.006	2.401	0.605
21087-64-9	O=C1N(C(=NN=C1C(C)(C)C)SC)N	23.325	0.550	1.228	−0.678
132-66-1	O=C(c1c(cccc1)C(=O)O)Nc1c2c(ccc1)cccc2	19.495	0.442	0.458	−0.016
105827-78-9	[O−][N+](=O)N=C1N(Cc2ccc(nc2)Cl)CCN1	32.738	3.516	3.120	0.396
23135-22-0	O=C(C(=NOC(=O)NC)SC)N(C)C	21.992	1.327	0.960	0.367
301-12-2	O=P(SCCS(=O)CC)(OC)OC	24.853	1.914	1.535	0.379
56-38-2	S=P(Oc1ccc(cc1)[N+](=O)[O−])(OCC)OCC	33.052	3.221	3.183	0.038
298-00-0	S=P(Oc1ccc(cc1)[N+](=O)[O−])(OC)OC	33.460	3.375	3.265	0.110
40487-42-1	[O−][N+](=O)c1c(c(cc(c1C)C)[N+](=O)[O−])NC(CC)CC	23.636	0.752	1.290	−0.538
13684-63-4	O=C(Nc1cc(ccc1)C)Oc1cccc(c1)NC(=O)OC	22.372	0.094	1.036	−0.942
298-02-2	S=P(SCSCC)(OCC)OCC	28.836	1.413	2.336	−0.923
732-11-6	S=P(SCN1C(=O)c2c(cccc2)C1=O)(OC)OC	31.595	2.476	2.890	−0.414
2312-35-8	C#CCOS(=O)OC1C(Oc2ccc(cc2)C(C)(C)C)CCCC1	29.279	1.369	2.425	−1.056
66841-25-6	N#CC(c1cc(ccc1)Oc1ccccc1)OC(=O)C1C(C1C(C(Br)(Br)Br)Br)(C)C	36.946	3.712	3.966	−0.254
1582-09-8	[O−][N+](=O)c1cc(cc(c1N(CCC)CCC)[N+](=O)[O−])C(F)(F)F	23.320	1.142	1.227	−0.085
3547-33-9	OCCSCCCCCCCC	18.334	0.525	0.225	0.300
1861-40-1	[O−][N+](=O)c1c(c(cc(c1)C(F)(F)F)[N+](=O)[O−])N(CC)CCCC	23.016	1.364	1.166	0.198
52315-07-8	N#CC(c1cc(ccc1)Oc1ccccc1)OC(=O)C1C(C1C=C(Cl)Cl)(C)C	39.536	4.258	4.486	−0.228
149877-41-8	O=C(NNc1cc(ccc1OC)c1ccccc1)OC(C)C	21.391	1.586	0.839	0.747
122453-73-0	N#Cc1c(n(c(c1Br)C(F)(F)F)COCC)c1ccc(cc1)Cl	36.384	3.531	3.853	−0.322
95266-40-3	O=C1C(=C(C2CC2)O)C(=O)CC(C1)C(=O)OCC	21.734	0.730	0.908	−0.178
91465-08-6	N#CC(c1cccc(c1)Oc1ccccc1)OC(=O)C1C(C1C=C(C(F)(F)F)Cl)(C)C	37.844	4.073	4.146	−0.073
82-68-8	[O−][N+](=O)c1c(c(c(c(c1Cl)Cl)Cl)Cl)Cl	33.839	3.470	3.341	0.129
23103-98-2	O=C(Oc1nc(nc(c1C)C)N(C)C)N(C)C	20.401	1.105	0.640	0.465
29232-93-7	S=P(Oc1nc(nc(c1)C)N(CC)CC)(OC)OC	30.276	2.894	2.625	0.269
23031-36-9	C#CCC1=C(C(OC(=O)C2C(C2C=C(C)C)(C)C)CC1=O)C	32.456	4.031	3.063	0.968
96489-71-3	O=C1N(N=CC(=C1Cl)SCc1ccc(cc1)C(C)(C)C)C(C)(C)C	39.318	4.182	4.443	−0.261
94-82-6	O=C(O)CCCOc1c(cc(cc1)Cl)Cl	25.404	1.235	1.646	−0.411
86-50-0	S=P(SCN1N=Nc2c(cccc2)C1=O)(OC)OC	27.598	2.878	2.087	0.791
1897-45-6	N#Cc1c(c(c(c(c1Cl)C#N)Cl)Cl)Cl	22.987	0.166	1.160	−0.994
333-41-5	S=P(Oc1nc(nc(c1)C)C(C)C)(OCC)OCC	29.932	3.182	2.556	0.626
957-51-7	O=C(C(c1ccccc1)c1ccccc1)N(C)C	17.601	−1.007	0.077	−1.084
115-29-7	O=S1OCC2C(C3(C(C2(C(=C3Cl)Cl)Cl)(Cl)Cl)Cl)CO1	32.457	1.956	3.063	−1.107
16672-87-0	O=P(CCCl)(O)O	21.409	1.077	0.843	0.234
22224-92-6	O=P(NC(C)C)(Oc1cc(c(cc1)SC)C)OCC	30.881	2.210	2.747	−0.537
77182-82-2	O=P(CCC(C(=O)O)N)(O)C	20.874	−0.241	0.735	−0.976
2032-65-7	O=C(Oc1cc(c(c(c1)C)SC)C)NC	24.905	2.779	1.546	1.233
13171-21-6	O=P(OC(=C(C(=O)N(CC)CC)Cl)C)(OC)OC	25.786	2.312	1.723	0.589
1918-02-1	O=C(c1c(c(c(c(n1)Cl)Cl)N)Cl)O	21.570	1.221	0.875	0.346
22248-79-9	O=P(OC(=CCl)c1c(cc(c(c1)Cl)Cl)Cl)(OC)OC	25.383	2.427	1.642	0.785
7696-12-0	O=C1N(C(=O)C2=C1CCCC2)COC(=O)C1C(C1C=C(C)C)(C)C	33.738	3.330	3.321	0.009
51630-58-1	N#CC(c1cc(ccc1)Oc1ccccc1)OC(=O)C(c1ccc(cc1)Cl)C(C)C	28.987	3.010	2.366	0.644
10265-92-6	O=P(SC)(N)OC	25.429	2.013	1.651	0.362
52-68-6	O=P(C(C(Cl)(Cl)Cl)O)(OC)OC	26.801	0.634	1.927	−1.293
2008-41-5	O=C(N(CC(C)C)CC(C)C)SCC	19.661	0.875	0.491	0.384
7786-34-7	O=P(OC(=CC(=O)OC)C)(OC)OC	30.261	3.505	2.622	0.883
584-79-2	O=C(C1C(C1C=C(C)C)(C)C)OC1C(=C(C(=O)C1)CC=C)C	30.072	1.949	2.584	−0.635
1076-46-6	O=C(c1c(c(cc(c1)Cl)N)Cl)O	21.238	1.187	0.808	0.379
85-00-7	c1c[n+]2c(c3[n+]1cccc3)cccc2	17.718	0.537	0.101	0.436
26002-80-2	O=C(C1C(C1C=C(C)C)(C)C)OCc1cccc(c1)Oc1ccccc1	34.061	3.719	3.386	0.333
140-56-7	O=S(=O)(N=Nc1ccc(cc1)N(C)C)O	17.884	0.391	0.134	0.257
173584-44-6	O=C(N1N=C2C(OC1)(Cc1cc(ccc12)Cl)C(=O)OC)N(c1ccc(cc1)OC(F)(F)F)C(=O)OC	33.200	3.467	3.213	0.254
150-68-5	O=C(Nc1ccc(cc1)Cl)N(C)C	16.527	0.257	−0.138	0.395
142-59-6	S=C(S)NCCNC(=S)S	24.824	1.326	1.529	−0.203
52645-53-1	O=C(C1C(C1C=C(Cl)Cl)(C)C)OCc1cccc(c1)Oc1ccccc1	37.792	4.212	4.136	0.076
76578-14-8	O=C(C(Oc1ccc(cc1)Oc1nc2ccc(cc2nc1)Cl)C)OCC	27.106	0.873	1.988	−1.115

Table 4 (*continued*)

CAS	SMILES	DCW	Expr	Calc	Expr − Calc
	Test set				
90982-32-4	O=S(=O)(c1c(cccc1)C(=O)OCC)NC(=O)Nc1nc(cc(n1)Cl)OC	20.877	1.521	0.736	0.785
62-73-7	O=P(OC=C(Cl)Cl)(OC)OC	30.331	2.645	2.636	0.009
60-57-1	ClC12C(C(C3C2C2C4C(O4)C3C2)(C(=C1Cl)Cl)Cl)(Cl)Cl	35.346	3.438	3.644	−0.206
330-54-1	O=C(Nc1cc(c(cc1)Cl)Cl)N(C)C	19.419	0.206	0.443	−0.237
2439-10-3	N=C(NCCCCCCCCCCCC)N	20.072	1.376	0.574	0.802
72-20-8	ClC12C(C(C3C2C2C4C(O4)C3C2)(C(=C1Cl)Cl)Cl)(Cl)Cl	35.346	2.275	3.644	−1.369
2104-64-5	S=P(c1ccccc1)(Oc1ccc(cc1)[N+](=O)[O−])OCC	31.185	3.121	2.808	0.313
55283-68-6	[O−][N+](=O)c1cc(cc(c1N(CC(=C)C)CC)[N+](=O)[O−])C(F)(F)F	21.055	0.815	0.772	0.043
41198-08-7	O=P(Oc1ccc(cc1Cl)Br)(SCCC)OCC	32.010	3.595	2.974	0.621
1610-18-0	COc1nc(nc(n1)NC(C)C)NC(C)C	26.986	0.797	1.964	−1.167
7287-19-6	CSc1nc(nc(n1)NC(C)C)NC(C)C	21.896	0.397	0.941	−0.544
74051-80-2	O=C1C(=C(CC(C1)CC(SCC)C)O)C(=NOCC)CCC	24.656	1.515	1.495	0.020
131929-60-7	O=C1C2=CC3C(C2CC(=O)OC(CCCC(C1C)OC1OC(C(CC1)N(C)C)C)CC)C= CC1C3CC(C1)OC1C(C(C(C(O1)C)OC)OC)OC	47.049	5.402	5.996	−0.594
137-26-8	S=C(SSC(=S)N(C)C)N(C)C	17.772	0.512	0.112	0.400
21725-46-2	N#CC(Nc1nc(nc(n1)NCC)Cl)(C)C	17.895	0.095	0.136	−0.041
127-20-8	O=C(C(Cl)(Cl)C)O	21.142	0.834	0.789	0.045
79277-27-3	O=S(=O)(c1c(scc1)C(=O)OC)NC(=O)Nc1nc(nc(n1)OC)C	23.692	1.491	1.302	0.189
101200-48-0	O=S(=O)(c1ccccc1C(=O)OC)NC(=O)N(c1nc(nc(n1)C)OC)C	22.453	0.597	1.053	−0.456
83055-99-6	O=C(c1c(cccc1)CS(=O)(=O)NC(=O)Nc1nc(cc(n1)OC)OC)OC	21.738	1.516	0.909	0.607
72-43-5	ClC(C(c1ccc(cc1)OC)c1ccc(cc1)OC)(Cl)Cl	19.209	1.166	0.401	0.765

Acknowledgement

The authors thank the Marie Curie fellowship for financial support (the contract ID 39036, CHEMPREDICT).

References and notes

1. Hu, Q. N.; Liang, Y. Z.; Fang, K. T. *J. Data Sci.* **2003**, *1*, 361.
2. Toropov, A. A.; Schultz, T. W. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 560.
3. Toropov, A. A.; Roy, K. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 179.
4. Gutman, I.; Toropov, A. A.; Toropova, A. P. *MATCH Commun. Math. Comput. Chem.* **2005**, *53*, 215.
5. Gutman, I.; Furtula, B.; Toropov, A. A.; Toropova, A. P. *MATCH Commun. Math. Comput. Chem.* **2005**, *53*, 225.
6. Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
7. Weininger, D.; Weininger, A.; Weininger, J. L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97.
8. Weininger, D. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 237.
9. Devillers, J.; Pham-Delègue, M. H.; Decourtye, A.; Budzinski, H.; Cluzeau, S.; Maurin, G. *SAR QSAR Environ. Res.* **2002**, *13*, 641.
10. Devillers, J.; Doré, J. C.; Tisseur, M.; Cluzeau, S.; Maurin, G. *Comput. Electron. Agric.* **2004**, *42*, 87.
11. Vighi, M.; Masoero, M.; Calamari, G.; Galamari, D. *Sci. Total Environ.* **1991**, *109–110*, 605.
12. Toropov, A. A.; Benfenati, E. *Comput. Biol. Chem.* **2007**, *31*, 57.
13. Zenkevich, I. G.; Moeder, M.; Koeller, G.; Schrader, S. *J. Chromatogr., A* **2004**, *1025*, 227.
14. Coles, S. J.; Day, N. E.; Murray-Rust, P.; Rzepa, H. S.; Zhang, Y. *Org. Biomol. Chem.* **2005**, *3*, 1832.
15. http://www.acdlabs.com/.