Mol Divers (2010) 14:183–192
DOI 10.1007/s11030-009-9156-6

SHORT COMMUNICATION

QSAR-modeling of toxicity of organometallic compounds by means of the balance of correlations for InChI-based optimal descriptors

A. A. Toropov · A. P. Toropova · E. Benfenati

Received: 19 January 2009 / Accepted: 27 April 2009 / Published online: 19 May 2009

Abstract Quantitative structure–activity relationships (QSAR) for toxicity toward rats (pLD50) have been built by means of optimal descriptors. Comparison of the optimal descriptors calculated using the International Chemical Identifier (InChI) with the optimal descriptors calculated using the simplified molecular input line entry system (SMILES) has shown that the InChI-based models give more accurate prediction for the abovementioned toxicity of organometallic compounds. These models were obtained by means of the balance of correlation: one subset of the training set (subtraining set) plays role of the training; the second subset (calibration set) plays role of the preliminary check of the models. It has been shown that the balance of correlations is a more robust predictor for the toxicity than the classic scheme (training set—test set: without the calibration set). Three splits into the subtraining set, calibration set, and test set were examined.

Keywords QSAR · InChI · SMILES · Balance of correlations · Organometallic compound · Toxicity toward rats

Electronic supplementary material The online version of this article (doi:10.1007/s11030-009-9156-6) contains supplementary material, which is available to authorized users.

A. A. Toropov · A. P. Toropova
Institute of Geology and Geophysics, Khodzhibaev St. 49, 100041 Tashkent, Uzbekistan

A. A. Toropov (✉) · A. P. Toropova · E. Benfenati
Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy
e-mail: aatoropov@yahoo.com

Abbreviations

QSPR	Quantitative structure–property relationships
QSAR	Quantitative structure–activity relationships
SMILES	Simplified molecular input line entry system
InChI	International Chemical Identifier
DCW	Descriptor of the correlation weights

Introduction

Quantitative structure–property/activity relationships (QSPR /QSAR) are tools for the prediction of physicochemical and biological parameters of substances. There is a series of QSPR/QSAR approaches described in literature [1–10].

Computational models of different kinds of toxicity are necessary for ecology, biology, and medicine, owing to at least two main reasons: (1) experimental determinations are often costly and time-consuming while accurate predictive models offer acceptable values easily and quickly; (2) models can be useful for developing further knowledge on the toxicity phenomena [11–21].

When using molecular fragments in the QSPR/QSAR, two approaches can be adopted: (1) the additive contributions of molecular fragments—the Free-Wilson model [22], and (2) the nonlinear contributions—the Fujita-Ban calculations [23]. The optimal descriptors calculated from molecular graphs have been used in the QSPR/QSAR analyses. In fact, the abovementioned optimal descriptors are a redefinition of the schemes based on the additive contributions and nonlinear contributions of the molecular fragments [24–34].

The simplified molecular input line entry system (SMILES) is a representation of a molecular structure by a sequence of symbols [35–37]. In a SMILES notation each symbol or group of symbols can represent a molecular fragment. Consequently, one can organize a modification of the abovementioned schemes based on SMILES [6,13–16,38–40]. The number of internet databases on physicochemical properties and biological activity of substances with representation of the molecular structure by means of SMILES gradually increases. Thus, the construction of QSPR/QSAR models based on SMILES becomes convenient and useful.

The International Chemical Identifier notation system (InChI) [41–44] is an alternative to the SMILES molecular representation system. The aim of this study is to estimate the robustness of the InChI notation system as the molecular structure representation system for the QSAR toxicity modeling of organometallic compounds by means of optimal descriptors.

Methods

Dataset

Rat toxicity data (LD50, in mg/kg, oral exposure) were taken from the U.S. Library of Medicine [45]. We have used compounds which could be correctly involved in the modeling process with our software: molecular nomenclature strings are limited to 130 symbols. The modeled endpoint is pLD50. These compounds contain Ni, Na, Cu, As, Fe, Ag, Mg, Mn, Co, K, Au, and Sb. There are compounds (in subtraining set, in calibration set, and in test set) which are containing one metal. There are compounds which are containing more than one metal. Finally, metals with different oxidation/ionic states take place in the compounds examined.

The chemicals contain heterocyclic and aromatic ring and carboxylic groups, oxygen, nitrogen, and sulfur atoms. Thus, the set is very diverse. We are not aware of other studies dedicated to QSAR models for a heterogeneous set of organometallic compounds.

The SMILES notations have been generated with the ChemSketch software [46]. The work set ($n = 56$) was randomly split into a subtraining ($n = 23$), calibration ($n = 23$), and a test set ($n = 10$). Three different splits have been examined. These splits are random, but their pLD50 ranges are similar for all the subtraining sets, all the calibration sets, and all the test sets. *Supplementary materials* section contains these splits.

SMILES-based optimal descriptors

The descriptors, which were used in this study, were calculated with SMILES attributes. The SMILES attribute is a combination of SMILES elements. The SMILES element is a group, that contains four, two, or one symbol of a SMILES notation.

In this study, 17 SMILES elements have been used: three elements of four symbols: "[N+]", "NH4+", and "[O−]"; 16 elements of two symbols: "+2", "+3", "@@", "Ag", "As", "Au", "Cl", "Co", "Cu", "Fe", "O=", "Mg", "Na", "Mn", "Ni", "Sb"; and 46 elements of one symbol: "#", " (", " +", "–", ".", "/", "1", "2", "3", "4", "5", "=", "@", "C", "F", "H", "K", "N", "O", "P", "S", "[", "\", "c", "n", "o", and "s".

Table 1 Example of the preparation SMILES attributes for copper 3-phenylsalicylate SMILES = "[Cu+2].O=C([O−])c2cccc(c1ccccc1)c2[O−]"; CAS=5328-04 − 1 The symbol "x" indicates unused positions

^1SA_k	^2SA_k	^3SA_k
[xxxxxxxxxxx		
Cuxxxxxxxxxx	[xxxCuxxxxxx	
+2xxxxxxxxxx	Cuxx + 2xxxxxx	[xxxCuxx + 2xx
[xxxxxxxxxxx	[xxx + 2xxxxxx	[xxx + 2xxCuxx
.xxxxxxxxxxx	[xxx.xxxxxxx	.xxx [xxx + 2xx
O = xxxxxxxxxx	O = xx.xxxxxxx	[xxx.xxxO = xx
Cxxxxxxxxxxx	O = xxCxxxxxxx	CxxxO = xx.xxx
(xxxxxxxxxxx	Cxxx (xxxxxxx	O = xxCxxx(xxx
[O−] xxxxxxxx	[O−] (xxxxxxx	[O−] (xxxCxxx
(xxxxxxxxxxx	[O−] (xxxxxxx	(xxx[O−] (xxx
cxxxxxxxxxxx	cxxx(xxxxxxx	cxxx(xxx[O−]
2xxxxxxxxxxx	cxxx2xxxxxxx	2xxxcxxx(xxx
cxxxxxxxxxxx	cxxx2xxxxxxx	cxxx2xxxcxxx
cxxxxxxxxxxx	cxxxcxxxxxxx	cxxxcxxx2xxx
cxxxxxxxxxxx	cxxxcxxxxxxx	cxxxcxxxcxxx
cxxxxxxxxxxx	cxxxcxxxxxxx	cxxxcxxxcxxx
(xxxxxxxxxxx	cxxx (xxxxxxx	cxxxcxxx (xxx
cxxxxxxxxxxx	cxxx (xxxxxxx	cxxx (xxxcxxx
1xxxxxxxxxxx	cxxx1xxxxxxx	1xxxcxxx (xxx
cxxxxxxxxxxx	cxxx1xxxxxxx	cxxx1xxxcxxx
cxxxxxxxxxxx	cxxxcxxxxxxx	cxxxcxxx1xxx
cxxxxxxxxxxx	cxxxcxxxxxxx	cxxxcxxxcxxx
cxxxxxxxxxxx	cxxxcxxxxxxx	cxxxcxxxcxxx
cxxxxxxxxxxx	cxxxcxxxxxxx	cxxxcxxxcxxx
1xxxxxxxxxxx	cxxx1xxxxxxx	cxxxcxxx1xxx
(xxxxxxxxxxx	1xxx(xxxxxxx	cxxx1xxx(xxx
cxxxxxxxxxxx	cxxx(xxxxxxx	cxxx (xxx1xxx
2xxxxxxxxxxx	cxxx2xxxxxxx	2xxxcxxx (xxx
[O−] xxxxxxxx	[O−] 2xxxxxxx	cxxx2xxx[O−]

Our modeling studies were conducted in three steps.

Step 1. Preparation of list of SMILES attributes for every SMILES notation. Each SMILES attribute is a string of 12 symbols. This string is separated into three zones: the first four symbols are the zone-1; the second four symbols are the zone-2; the third four symbols are the zone-3.

There are three categories of the SMILES attributes. The first category includes attributes (^1SA_k) which only contain SMILES element positioned in the zone-1; the second category includes attributes (^2SA_k) which are containing

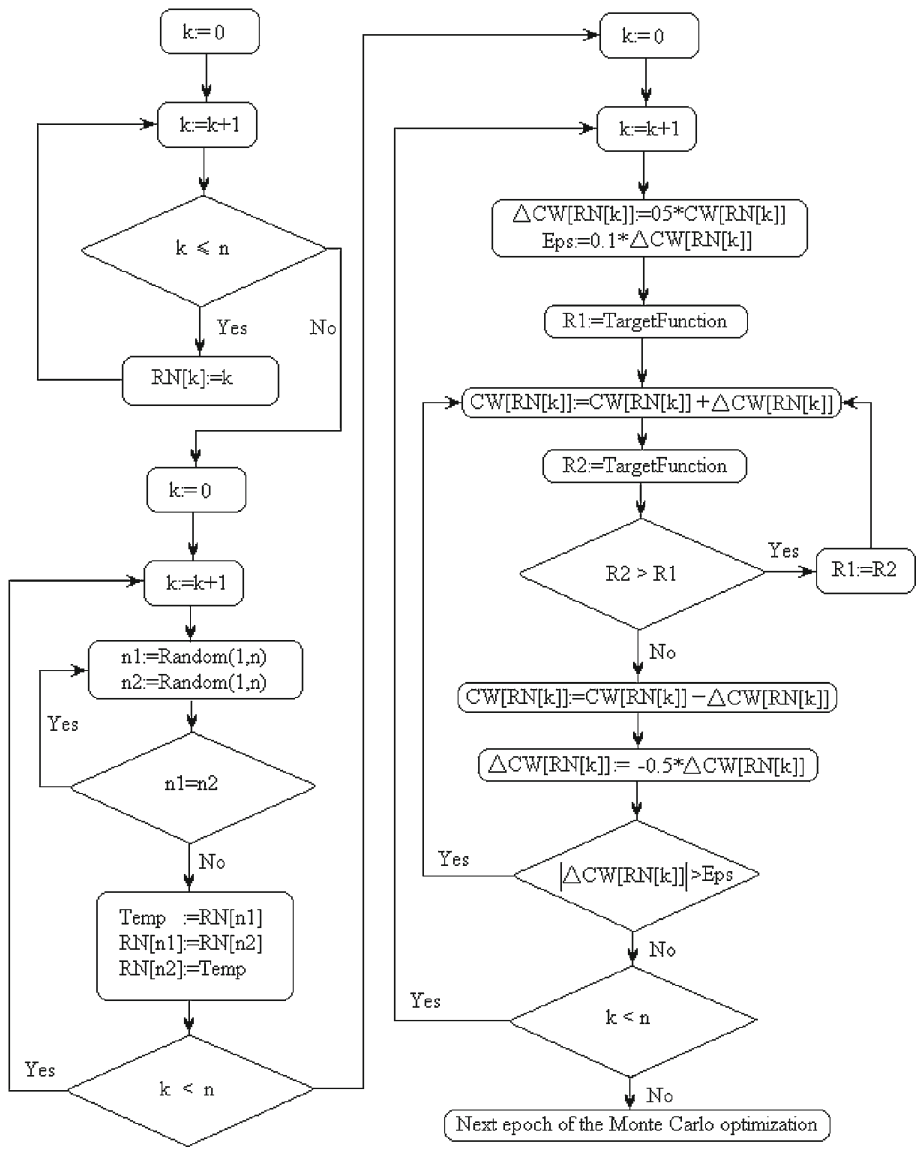


Fig. 1 The flowchart of the Monte Carlo optimization: CW[k] is the correlation weight of k-th attribute; Random(1,n) is the random integer from (1,n); n is the number of attributes; the RN is the array for the random sequence of the numbers

two SMILES elements which are positioned in zone-1 and zone-2; the third category includes attributes (^3SA_k) which contain three SMILES elements which are positioned in zone-1, zone-2, and zone-3.

Table 1 contains an example of the preparation of a list of the attributes for a SMILES notation.

In order to avoid a situation where two different SMILES attributes are representing the same molecular fragment, for instance, the '(N' and the 'N(', the elements for the ^2SA_k and ^3SA_k are ranged according to their ASCII codes. In addition, the symbol ')' is replaced by '(', because these are the representations of the same phenomenon (i.e., branch in molecular skeleton). The same takes place for the "[" and "]".

Step 2. Preparation of a complete list of the SMILES attributes which take place in the work set (i.e., in both the training and test sets). Every SMILES attribute is provided by a correlation weight equal to 1.

Step 3. Optimization of the correlation weights by the Monte Carlo method: Two systems of the modeling were examined: (1) the maximization of the correlation coefficient between the DCW(LimN) and log(LD_{50}) for the subtraining set and calibration set (the classic scheme); (2) the maximization of the criterion calculated as follows:

$$BC = R_s + R_c - ABS(R_s - R_c)^*0.1 \quad (1)$$

where R_s and R_c are the correlation coefficients between the DCW(LimN) and pLD50 for the subtraining set and calibration set, respectively (balance of correlations) [14].

The SMILES-based optimal descriptor is calculated as follows:

$$DCW(LimN) = \Sigma CW(^1SA_k) + \Sigma CW(^2SA_k) + \Sigma CW(^3SA_k) \quad (2)$$

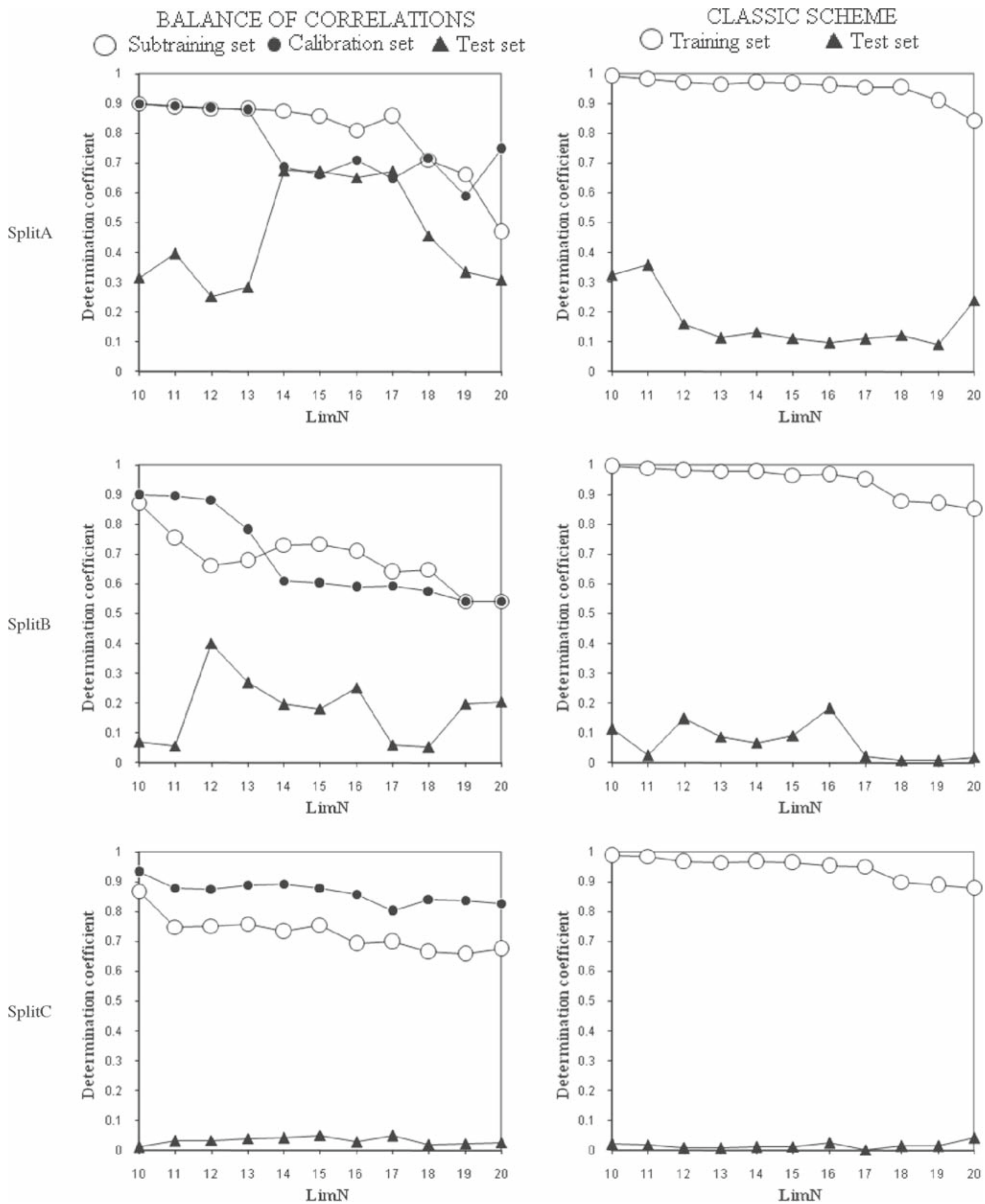


Fig. 2 QSAR-models for toxicity which were obtained with the SMILES-based optimal descriptors

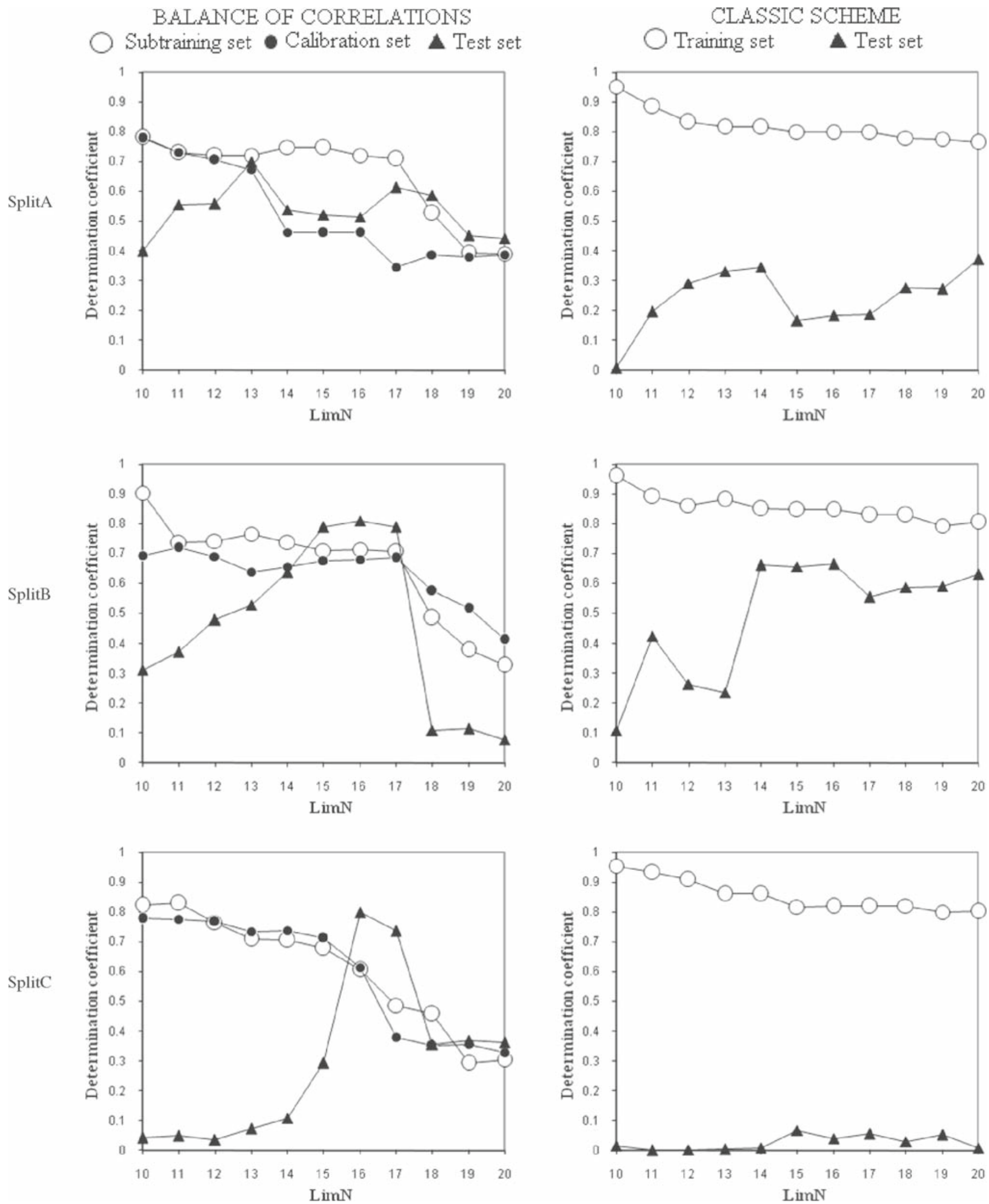


Fig. 3 QSAR-models for toxicity which were obtained with the InChI-based optimal descriptors

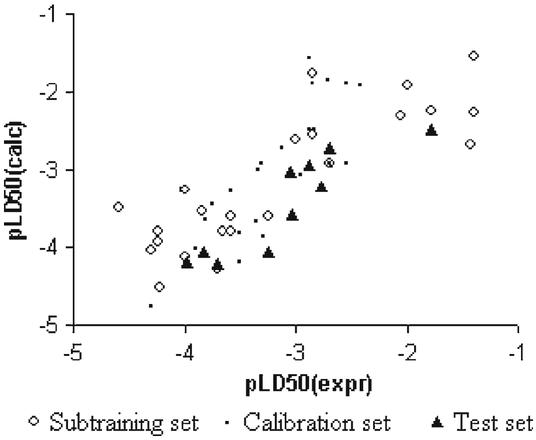


Fig. 4 Experimental and calculated values of the toxicity toward rats using Eq. 4

where $CW(^1SA_k)$, $CW(^2SA_k)$, and $CW(^3SA_k)$ are the correlation weights for the abovementioned SMILES attributes. The LimN is a parameter of the model, which gives possibility to classify the SMILES attributes into two categories: rare or not rare. Our hypothesis is that rare SMILES attributes can lead to overtraining. The influence of the rare attributes may be blocked, if correlation weights of the rare attributes are fixed equal to zero. The physical meaning of the LimN is the minimal number of the attribute ^mSA_k ($m = 1, 2, 3$) in the subtraining set which defines the ^mSA_k as "not rare". For instance, if LimN = 10, then ^mSA_k is not rare if the number of the ^mSA_k in the subtraining set is more than nine. The training set for the classic scheme is the united set that contains both the subtraining set and calibration set.

The flowchart of one epoch of the Monte Carlo optimization is shown in Fig. 1. In this study, 30 epochs have been used for each model, because 20–25 epochs give poorer statistical quality, whereas 35–40 do not improve the models. In the case of the classic scheme, the target function is correlation coefficients for the united set that contains both the subtraining and calibration set. In the case of the correlation balance, the target function is the values of the BC calculated with Eq. 1.

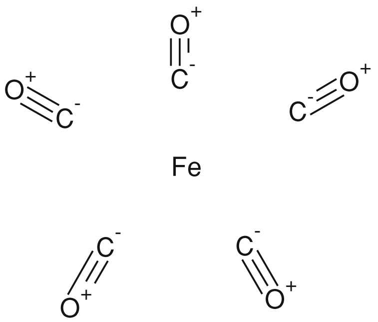
Table 2 Correlation weights of the InChI attributes which have been obtained in the three probes of the Monte Carlo optimization (Split B, limN = 16); The N(Train), N(Calib), and N(Test) are the number of the given IA_k in the subtraining set, calibration set, and test set, respectively

IA_k	CW(IAk) Probe 1	CW(IAk) probe 2*	CW(IAk) probe 3	N(Train)	N(Calib)	N(Test)
(	0.5510924	**0.5985947**	0.6529911	118	116	62
*	−1.4001338	**−1.3243332**	−1.3267216	31	34	20
+	−2.2153219	**−2.2027432**	−2.2541177	18	14	6
,	2.4982076	**2.8041291**	2.7988417	18	17	6
-1	1.3019853	**1.3029165**	1.2951617	17	21	8
-2	−2.2997618	**−2.3023527**	−2.2959676	29	22	11
-3	1.4028696	**1.2995441**	1.2970604	17	14	5
-4	1.4294410	**1.4477678**	1.6672913	17	14	11
-5	−0.4127315	**−0.2022803**	−0.2000199	19	22	5
-7	2.9497926	**3.0503589**	3.2740513	17	13	4
.	−1.8272951	**−1.8148697**	−2.0249550	38	44	15
/	0.7131912	**0.8280859**	0.8245566	116	112	47
1	−1.6892054	**−1.7135685**	−1.8723899	102	100	42
2	2.0724838	**2.0030188**	2.0674595	50	50	28
;;	1.7674783	**1.7011293**	1.7956587	19	23	7
;	−0.5281883	**−0.5527381**	−0.6004048	60	50	25
=	2.7098726	**2.3249858**	2.9350858	23	23	10
H2	−0.9541424	**−1.0544814**	−1.0544744	23	33	10
H3	−0.2369197	**−0.3222115**	−0.4230117	17	13	3
H	0.7224586	**0.5263852**	0.5721372	36	30	18
I	4.9042302	**5.4479441**	5.3998817	23	22	10
c	2.9461705	**3.3220270**	2.1136173	46	46	20
h	7.5019033	**7.6015628**	7.6041812	24	24	11
i	0.0960859	**0.0003001**	−0.3022785	23	24	10
n	3.2919846	**2.1641790**	2.9614935	23	23	10
q	−2.2996951	**−2.3008030**	−2.2533275	21	21	10

Data on the blocked attributes are omitted: full list of attributes is represented in Supplementary materials

* Correlation weights used for Eq. 4 are indicated by bold

Table 3 Example of the DCW(16) calculation for iron pentacarbonyl (CAS 13463-40-6) InChI=1/5CO.Fe/c5*1−2; DCW= 16.4177669



IA_k	$CW(IA_k)$
I	5.4479441
n	2.1641790
c	3.3220270
h	7.6015628
i	0.0003001
=	2.3249858
1	−1.7135685
/	0.8280859
5	0.0
C	0.0
O	0.0
.	−1.8148697
Fe	0.0
/	0.8280859
c	3.3220270
5	0.0
*	−1.3243332
1	−1.7135685
−2	−2.3023527
;	−0.5527381

InChI-based optimal descriptors

InChI-based optimal descriptors have been studied with the same three splits into subtraining set, calibration set, and test set. The InChI notations for organometallic compounds were obtained using the ChemSketch software [46].

The InChI-based optimal descriptors are calculated as per the following equation

$$\mathrm{DCW(LimN)} = \sum \mathrm{CW(IA_k)} \quad (3)$$

where IA_k is an attribute of the InChI.

The InChI attributes used for calculation of the DCW with Eq. 2 are the following: “(10”, “(11”, “(12”, “(13”, “(14”, “(15”, “(16”, “(17”, “(18”, “(19”, “(20”, “(21”, “(22”, “(23”, “(24”, “(25”, “(27”, “ (28”, “(2”, “(3”, “(4”, “(5”, “(6”, “(7”, “(8”, “(9”, “(”, “*”, “+2”, “+3”, “+”, “,10”, “,11”, “,12”, “,13”, “,14”, “,15”, “,16”, “,17”, “,18”, “,19”, “,20”, “,22”, “,23”, “,24”, “,25”, “,26”, “,28”, “,29”, “,1”, “,2”, “,3”, “,4”, “,5”, “,6”, “,7”, “,8”, “,9”, “,”, “−10”, “−11”, “−12”, “−13”, “−14”, “−15”, “−16”, “−17”, “−18”, “−19”, “−20”, “−21”, “−22”, “−25”, “−27”, “−28”, “−29”, “−30”, “−1”,”−2”, “−3”, “−4”, “−5”, “−6”, “−7”, “−8”, “−9”, “−”, “.”, “/p-2”, “/”, “0”, “1”, “2”, “3”, “4”, “5”, “6”, “7”, “8”, “9”, “;;;;”, “;;;”, “;;”, “;”, “=”, “Ag”, “As”, “Au”, “C10”, “C11”, “C13”, “C14”, “C15”, “C16”, “C17”, “C18”, “C24”, “C2”, “C3”, “C4”, “C5”, “C6”, “C7”, “C8”, “C9”, “Cl”, “Co”, “Cu”, “C”, “Fe”, “F”, “H10”, “H11”, “H12”, “H14”, “H15”, “H16”, “H17”, “H19”, “H20”, “H34”, “H36”, “H2”, “H3”, “H4”, “H5”, “H7”, “H8”, “H”, “I”, “K”, “Mg”, “Mn”, “N2”, “N3”, “N4”, “N5”, “Na”, “Ni”, “N”, “O2”, “O3”, “O4”, “O5”, “O6”, “O7”, “O8”, “O9”, “O”, “P”, “Sb”, “S”, “b”, “c”, “h10”, “h13”, “h1”, “h2”, “h3”, “h5”, “h7”, “h8”, “h”, “I”, “m11”, “m1”, “m”, “n”, “p”, “q”, “s1”, “t10”, “t13”, “t2”, and “t7”.

Each InChI-attribute reflects some molecular properties [41–44], e.g., ‘−1,’, ‘−2’, …, ‘−11’ are attributes of the connectivity layer; ‘C2’, ‘C3’, …, ‘C8’, ‘H2’, ‘H3’, …, ‘H19’, ‘N2’,…,‘N5’, ‘O2’,…,‘O9’, are components of the formula layer, ‘t5’, ‘s1’, ‘m1’ are components of the stereochemical layer; ‘q’, ‘;’, ‘;;’, ‘;;;’, and ‘;;;;’ are attributes of the electronic charge layer.

In the case of the InChI-based molecular descriptors, the same algorithm of the Monte Carlo optimization (as well as the splits) was used as in the case of the SMILES-based optimal descriptors.

Results

Figure 2 shows the statistical characteristics of the QSAR models for pLD50, which have been calculated via Monte Carlo optimization with LimN values ranging from 10 to 20 for SMILES-based optimal descriptors. It can be determined from Fig. 2 that a satisfactory model takes place for the split A in the case of scheme of the balance of correlations with the LimN = 14. In the cases of split B and C, the SMILES-based optimal descriptors gave unsatisfactory models for both the classic scheme and the balance of correlations.

Figure 3 shows the statistical characteristics of the QSAR models for pLD50, which have been calculated via Monte Carlo optimization with LimN values ranging from 10 to 20 for InChI-based optimal descriptors. It can be determined from Fig. 3 that for all the three splits, the balance of the correlations for the InChI-based optimal descriptors gave reasonable good predictions, but with different LimN values: for split A the best LimN is 13, for split B and split C the best LimN is 16.

Table 4 Distributions of the "strange" SMILES attributes ($^{m}\mathbf{SA_k}$) and InChI attributes ($\mathbf{IA_k}$) in the Split A, Split B, and Split C

	Split A			Split B			Split C		
"Strange" attributes	N(Train)	N(Calib)	N(Test)	N(Train)	N(Calib)	N(Test)	N(Train)	N(Calib)	N(Test)
SMILES, $^{m}\mathbf{SA_k}$									
Hxxxxxxxxxxx	27	21	2	22	28	0	11	23	16
Sxxxxxxxxxxx	21	14	8	16	8	19	10	21	12
Cxxx@xxxxxxx	14	11	2	15	12	0	8	7	12
Fxxx (xxxxxxx	19	0	0	8	11	0	0	19	0
Hxxx@@xxxxxx	11	5	0	5	11	0	1	10	5
Hxxx@xxxxxxx	14	10	2	15	11	0	8	7	11
NxxxCxxxxxxx	16	29	4	9	20	20	15	30	4
Oxxx (xxxxxxx	27	17	3	22	22	3	4	28	15
[O−]. xxxxxxx	10	15	10	16	15	4	16	9	10
[xxx#xxxxxxx	22	0	0	10	12	0	22	0	0
[xxx (xxxxxxx	37	29	5	29	32	10	4	35	32
[xxx-xxxxxxx	23	7	0	11	10	9	15	10	5
[xxxOxxxxxxx	18	3	0	11	10	0	15	5	1
cxxx2xxxxxxx	13	22	0	13	19	3	24	11	0
(xxx [O−] (xxx	11	17	6	10	14	10	17	16	1
CxxxCxxx(xxx	17	23	10	11	18	21	16	28	6
Hxxx@xxxCxxx	14	10	2	15	11	0	8	7	11
[xxx (xxxOxxx	15	12	0	13	14	0	1	15	11
[xxx.xxx [xxx	18	5	2	6	10	9	16	9	0
[xxxCxxx@xxx	14	11	2	15	12	0	8	7	12
[xxxHxxx@xxx	14	10	2	15	11	0	8	7	11
cxxxcxxx (xxx	9	9	2	12	4	4	14	6	0
cxxxcxxx2xxx	7	15	0	9	11	2	15	7	0
InChI, $\mathbf{IA_k}$									
+	20	12	6	18	14	6	11	19	8
0	18	7	4	12	12	5	16	9	4
H3	12	11	10	17	13	3	16	9	8

The N(Train), N(Calib), and N(Test) are the numbers of a given attribute in the subtraining set, calibration set, and test set, respectively

Supplementary material for this study contains numerical data on the statistical characteristics of the SMILES-based and InChI-based models of the toxicity.

The second probe (Split B) of the Monte Carlo optimization for the DCW(16) for InChI-based optimal descriptors gave the following model:

$$\mathrm{pLD50} = 2.0415 - 0.2187^{*}\mathrm{DCW(16)} \tag{4}$$

$$n = 23,\ \ r^2 = 0.710,\ \ s = 0.557,\ \ F = 52\ \text{(subtraining set)};$$

$$n = 23,\ \ r^2 = 0.680,\ \ s = 0.541,\ \ F = 45\ \text{(calibration set)};$$

$$n = 10,\ \ r^2 = 0.803,\ \ s = 0.481,\ \ F = 33\ \text{(test set)}$$

Similar results have been obtained with the other two probes for InChI as well as with ones done on the other splits. Figure 4 shows the model graphically.

Table 2 contains the numerical data on the correlation weights for the DCW(16) calculation (Split B, probe 2). Table 3 contains an example of the DCW(16) calculation for iron pentacarbonyl (CAS 13463-40-6). *Supplementary materials* section contains numerical data on the DCW(16) for all the 56 organometallic compounds together with experimental and calculated pLD50 values using Eq. 4.

Discussion

The SMILES notations are a convenient method of representation from a chemical point of view. InChI representation is not transparent. However, InChI is a more informative format for the QSAR modeling of the toxicity of organometallic compounds by the optimal descriptors (Figs. 1 and 2).

The balance of the correlations was previously used for the QSAR modeling of the toxicity of organic compounds [14]. Thus, while comparing the previous study [14] with this one, this latter approach has demonstrated an advantage over the classic scheme, namely, the elimination of the calibration set.

According to Figs. 1 and 2, the most important attributes of the models (for both cases: SMILES and InChI) are the attributes which take place in the subtraining set approximately 13–16 times. Consequently, attributes which take place in the subtraining 13–16 times for one split, but which take place less than 13 times for an other split, should be classified as "strange". Thus the "strange" attributes are those which are rare at least in one split, but not in all the splits. These "strange" attributes can have negative influence for the statistical quality of the prediction for the external test set.

Table 4 contains the list of the "strange" attributes for the case of the SMILES and the case of the InChI. One can see that the number of "strange" InChI attributes is three, whereas the number of the "strange" SMILES attributes is 23. Since the InChI-based descriptors are robust predictors of the toxicity for all the three splits, one can expect that the number of the "strange" attributes is an informative characteristic of the optimal descriptors: probably, the small number of the "strange" attributes is a promoter of satisfactory quality of the prediction.

Conclusions

1. InChI-based optimal descriptors gave more accurate prediction values of rat toxicity for organometallic compounds than the SMILES-based optimal descriptors; the SMILES-based descriptors gave satisfactory prediction for the split A (only), whereas the InChI-based descriptors gave satisfactory prediction for all three splits;
2. The balance of correlations (i.e., training with two sets: the subtraining set and calibration set) gave more accurate prediction than the classic scheme (i.e., modeling by the scheme of "training-test", without the calibration set);
3. The split into the subtraining set, calibration set, and test set has considerable influence on the predictive ability of the optimal descriptors.

Acknowledgements The authors acknowledge the Marie Curie Fellowship (the contract ID 39036, CHEMPREDICT) and thank Federchimica-AISPEC for financial support.

References

1. Marrero-Ponce Y, Castillo-Garit JA, Castro EA, Torrens F, Rotondo R (2008) 3D-chiral (2.5) atom-based TOMOCOMD-CARDD descriptors: theory and QSAR applications to central chirality codification. J Math Chem 44:755–786. doi:10.1007/s10910-008-9386-3
2. Duchowicz PR, Talevi A, Bruno-Blanch LE, Castro EA (2008) New QSPR study for the prediction of aqueous solubility of drug-like compounds. Bioorg Med Chem 16:7944–7955. doi:10.1016/j.bmc.2008.07.067
3. Ray S, Sengupta C, Roy K (2008) QSAR modeling for lipid peroxidation inhibition potential of flavonoids using topological and structural parameters. Cent Eur J Chem 6:267–276. doi:10.2478/s11532-008-0014-7
4. Roy K, Roy PP (2008) Comparative QSAR studies of CYP1A2 inhibitor flavonoids using 2D and 3D descriptors. Chem Biol Drug Des 72:370–382. doi:10.1111/j.1747-0285.2008.00717.x
5. Toropov AA, Toropova AP, Gutman I (2005) Comparison of QSPR models based on hydrogen-filled graphs and on graphs of atomic orbitals. Croat Chem Acta 78:503–509
6. Toropov AA, Toropova AP, Mukhamedzhanova DV, Gutman I (2005) Simplified molecular input line entry system (SMILES) as an alternative for constructing quantitative structure-property relationships (QSPR). Indian J Chem - Sec A 44:1545–1552
7. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2006) A novel QSAR model for evaluating and predicting the inhibition activity of dipeptidyl aspartyl fluoromethylketones. QSAR Comb Sci 25:928–935. doi:10.1002/qsar.200530208
8. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2006) Prediction of intrinsic viscosity in polymer-solvent combinations using a QSPR model. Polymer (Guildf) 47:3240–3248. doi:10.1016/j.polymer.2006.02.060
9. Puzyn T, Mostrag A, Suzuki N, Falandysz J (2008) QSPR-based estimation of the atmospheric persistence for chloronaphthalene congeners. Atmos Environ 42:6627–6636. doi:10.1016/j.atmosenv.2008.04.048
10. Puzyn T, Suzuki N, Haranczyk M, Rak J (2008) Calculation of quantum-mechanical descriptors for QSPR at the DFT level: is it necessary? J Chem Inf Model 48:1174–1180. doi:10.1021/ci800021p
11. Duchowicz PR, Vitale MG, Castro EA (2008) Partial order ranking for the aqueous toxicity of aromatic mixtures. J Math Chem 44:541–549. doi:10.1007/s10910-007-9327-6
12. Roy K, Ghosh G (2008) QSTR with extended topochemical atom indices. 10. Modeling of toxicity of organic chemicals to humans using different chemometric tools. Chem Biol Drug Des 72:383–394. doi:10.1111/j.1747-0285.2008.00712.x
13. Toropov AA, Benfenati E (2008) Additive SMILES-based optimal descriptors in QSAR modelling bee toxicity: Using rare SMILES attributes to define the applicability domain. Bioorg Med Chem 16:4801–4809. doi:10.1016/j.bmc.2008.03.048
14. Toropov AA, Rasulev BF, Leszczynski J (2008) QSAR modeling of acute toxicity by balance of correlations. Bioorg Med Chem 16:5999–6008. doi:10.1016/j.bmc.2008.04.055
15. Toropov AA, Rasulev BF, Leszczynski J (2007) QSAR modeling of acute toxicity for nitrobenzene derivatives towards rats: comparative analysis by MLRA and optimal descriptors. QSAR Comb Sci 26:686–693. doi:10.1002/qsar.200610135
16. Toropov AA, Benfenati E (2007) Optimisation of correlation weights of SMILES invariants for modelling oral quail toxicity. Eur J Med Chem 42:606–613. doi:10.1016/j.ejmech.2006.11.018
17. Norinder U, Liden P, Bostrom H (2006) Discrimination between modes of toxic action of phenols using rule based methods. Mol Divers 10:207–212. doi:10.1007/s11030-006-9019-3
18. Melagraki G, Afantitis A, Sarimveis H, Igglessi-Markopoulou O, Alexandridis A (2006) A novel RBF neural network training methodology to predict toxicity to *Vibrio fischeri*. Mol Divers 10:213–221. doi:10.1007/s11030-005-9008-y
19. Isayev O, Rasulev B, Gorb L, Leszczynski J (2006) Structure-toxicity relationships of nitroaromatic compounds. Mol Divers 10:233–245. doi:10.1007/s11030-005-9002-4
20. Kuz'min VE, Muratov EN, Artemenko AG, Gorb L, Qasim M, Leszczynski J (2008) The effect of nitroaromatics' composition on

their toxicity in vivo: novel, efficient non-additive 1D QSAR analysis. Chemosphere 72:1373–1380. doi:10.1016/j.chemosphere.2008.04.045
21. Kuz'min VE, Muratov EN, Artemenko AG, Gorb L, Qasim M, Leszczynski J (2008) The effects of characteristics of substituents on toxicity of the nitroaromatics: HiT QSAR study. J Comput Aided Mol Des 22:747–759. doi:10.1007/s10822-008-9211-x
22. Fouchécourt M-O, Beéliveau M, Krishnan K (2001) Quantitative structure-pharmacokinetic relationship modeling. Sci Tot Environ 274:125–135
23. Gombar VK, Kapoor VK (1990) Quantitative structure-activity relationship studies: β-adrenergic blocking activity of 1-(2,4-disubstituted phenoxy)-3-aminopropan-2-ols. Eur J Med Chem 25:689–695. doi:10.1016/0223-5234(90)90134-O
24. Toropov AA, Toropova AP (2000) QSPR modeling of the stability constants of biometal complexes with phosphate derivatives of adenosine. Russ J Coord Chem 26:792–797
25. Toropov AA, Toropova AP (2000) QSPR modeling of the formation constants for complexes using atomic orbital graphs. Russ J Coord Chem 26:398–405
26. Toropov AA, Toropova AP (2001) Prediction of heteroaromatic amine mutagenicity by means of correlation weighting of atomic orbital graphs of local invariants. J Mol Struct (Theochem) 538:287–293. doi:10.1016/S0166-1280(00)00713-2
27. Toropov AA, Toropova AP (2002) QSAR modeling of toxicity on optimization of correlation weights of Morgan extended connectivity. J Mol Struct (Theochem) 578:129–134. doi:10.1016/S0166-1280(01)00695-9
28. Toropov A, Toropova A (2004) Nearest neighboring code and hydrogen bond index in labeled hydrogen-filled graph and graph of atomic orbitals: application to model of normal boiling points of haloalkanes. J Mol Struct (Theochem) 711:173–183. doi:10.1016/j.theochem.2004.10.003
29. Gutman I, Furtula B, Toropov AA, Toropova AP (2005) The graph of atomic orbitals and its basic properties. 2. Zagreb indices. Match 53:225–230
30. Gutman I, Toropov AA, Toropova AP (2005) The graph of atomic orbitals and its basic properties. 1. Wiener index. Match 53:215–224
31. Toropov AA, Toropova AP (2002) Modeling of acyclic carbonyl compounds normal boiling points by correlation weighting of nearest neighboring codes. J Mol Struct (Theochem) 581:11–15. doi:10.1016/S0166-1280(01)00733-3
32. Toropov AA, Benfenati E (2004) QSAR modelling of aldehyde toxicity by means of optimisation of correlation weights of nearest neighbouring codes. J Mol Struct (Theochem) 676:165–169. doi:10.1016/j.theochem.2004.01.023
33. Toropov AA, Benfenati E (2004) QSAR modelling of aldehyde toxicity against a protozoan, Tetrahymena pyriformis by optimization of correlation weights of nearest neighboring codes. J Mol Struct (Theochem) 679:225–228. doi:10.1016/j.theochem.2004.04.020
34. Toropov A, Toropova A (2004) Nearest neighboring code and hydrogen bond index in labeled hydrogen-filled graph and graph of atomic orbitals: application to model of normal boiling points of haloalkanes. J Mol Struct (Theochem) 711:173–183. doi:10.1016/j.theochem.2004.10.003
35. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36. doi:10.1021/ci00057a005
36. Weininger D, Weininger A, Weininger JL (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. J Chem Inf Comput Sci 29:97–101. doi:10.1021/ci00062a008
37. Weininger D (1990) Smiles. 3. Depict. Graphical depiction of chemical structures. J Chem Inf Comput Sci 30:237–243. doi:10.1021/ci00067a005
38. Vidal D, Thormann M, Pons M (2005) LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. J Chem Inf Model 45:386–393. doi:10.1021/ci0496797
39. Vidal D, Thormann M, Pons M (2006) A novel search engine for virtual screening of very large databases. J Chem Inf Model 46:836–843. doi:10.1021/ci050458q
40. Toropov AA, Benfenati E (2007) SMILES in QSPR/QSAR modeling: results and perspectives. Curr Drug Disc Tech 4:77–116
41. Degtyarenko K, Ennis M, Garavelli JS (2007) Good annotation practice for chemical data in biology. In Silico Biol 7:45–56
42. Prasanna MD, Vondrasek J, Wlodawer A, Bhat TN (2005) Application of InChI to curate, index, and query 3-D structures. Proteins 60:1–4. doi:10.1002/prot.20469
43. Coles SJ, Day NE, Murray-Rust P, Rzepa HS, Zhang Y (2005) Enhancement of the chemical semantic web through the use of InChI identifiers. Org Biomol Chem 3:1832–1834. doi:10.1039/b502828k
44. Bertinetto C, Duce C, Micheli A, Solaro R, Starita A, Tiné MR (2007) Prediction of the glass transition temperature of (meth)acrylic polymers containing phenyl groups by recursive neural network. Polymer (Guildf) 48:7121–7129. doi:10.1016/j.polymer.2007.09.043
45. U.S. Library of Medicine (2008). http://toxnet.nlm.nih.gov/
46. ACD/ChemSketch Freeware (2008) version 11.00, Advanced Chemistry Development, Inc., Toronto, ON, Canada. www.acdlabs.com