Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 16 (2008) 5999–6008
Bioorganic & Medicinal Chemistry

QSAR modeling of acute toxicity by balance of correlations

Andrey A. Toropov,* Bakhtiyor F. Rasulev and Jerzy Leszczynski

Computational Center for Molecular Structure and Interactions, Department of Chemistry, Jackson State University, 1400 J. R. Lynch Street, P.O. Box 17910, Jackson, MS 39217, USA

Received 29 December 2007; accepted 23 April 2008
Available online 26 April 2008

Abstract—Optimal descriptors based on the simplified molecular input line entry system (SMILES) have been utilized in modeling of acute toxicity towards rats. Toxicity of 61 benzene derivatives has been modeled by means of balance of correlations for sets of the training ($n = 27$) and calibration ($n = 24$). The obtained models were evaluated with the external test set ($n = 10$). Comparison of models based on the balance of correlations and models which were obtained on base of the total training (i.e., in case of utilization both training and calibration sets as the united training set) has shown that the balance of correlations gives improvement of statistical quality for the external test set. Predictions based on the one-variable model (based on the correlation balance) are better that the results obtained by the multiple linear regression analysis based on topological and quantum chemical descriptors. A QSAR analysis showed that the electronegativity of the molecule plays an important role in acute toxicity of benzene derivatives studied; presence of electronegative groups increasing toxicity. The presence of nitrogen-containing groups (mostly NH groups) increasing the toxicity that confirmed by both approaches.

1. Introduction

The various reactive benzene derivatives accumulating in the environment may create serious public health and environmental problems. Most of these compounds have carcinogenic activity and may bioaccumulate in the food chain[1] and also been shown to be toxic or mutagenic to microorganisms.[2] Benzene derivatives have been the object of considerable interest for chemists working in the field of environmental science during the past decades. An environmentally oriented approach as Quantitative Structure–Activity (Property) Relationship (QSAR/QSPR) has been efficiently used for the study of toxicity mechanisms of various reactive chemicals. This is a powerful technique, which quantitatively relates variations in biological activity to changes in molecular properties, i.e., the method attempts to link activity data with descriptors chosen via identification of the 'rules' that can be used further to monitor chemical transformations in the environment or in living organisms. Current advances in the QSPR/QSAR technique have been discussed in a paper,[3] which contains a review concerning QSAR analysis of experimental data on biotransformation and toxicity. In this review paper has shown several cases of successful results in defining parameters for the QSAR studies, as well as of the QSAR results based on quantum chemical calculations.

In literature almost no data concerning QSAR studies of benzene derivatives using acute toxicity data for animals. Nevertheless, recently only our group applied the toxicity data for rats in QSAR analysis of benzene derivatives.[4,5] In these studies we have discussed previous toxicity studies of benzene derivatives (all previous studies were based only on algae toxicity or fish toxicity) and then performed a QSAR analysis concerning acute toxicity (LD_{50}) of nitrogroup-containing benzene derivatives and also we made a quantum-chemical analysis, and then used a number of quantum-chemical parameters as descriptors. The satisfactory correlations between oral toxicity for rats and lowest unoccupied molecular orbital (LUMO) energy as well as some topological descriptors have been found in these studies.

High accuracy in prediction of toxicity is impossible[6] because this endpoint is defined not only by molecular structure of toxin, but also depends on the characteristics of the living organism. Nevertheless, predictive models for the toxicity are very useful.[6–10] Majority of the applied models are based on molecular graphs.[7–9]

Keywords: QSAR; Toxicity towards rats; MLRA; Optimal descriptors; SMILES; Balance of correlations.
* Corresponding author. Tel.: +1 601 979 3979; fax +1 601 979 7823; e-mail: aatoropov@yahoo.com

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.04.055

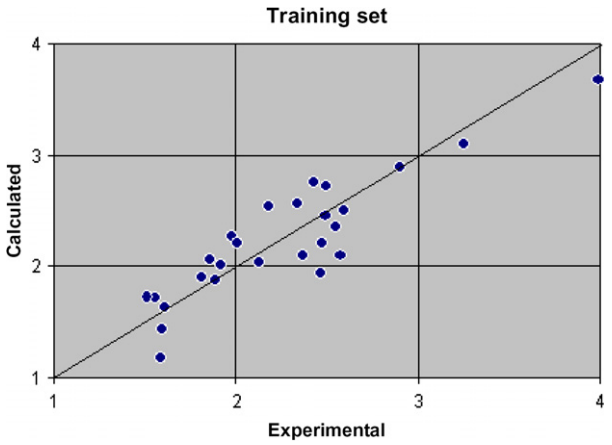

Figure 1. Experimental versus calculated with Eq. 3 values of toxicity towards rats for the training set.

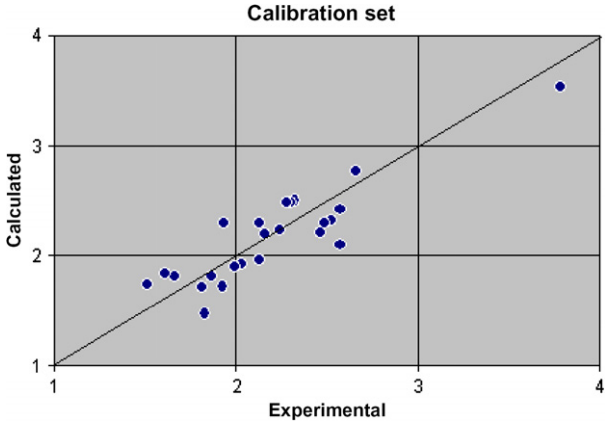

Figure 2. Experimental versus calculated with Eq. 3 values of toxicity towards rats for the calibration set.

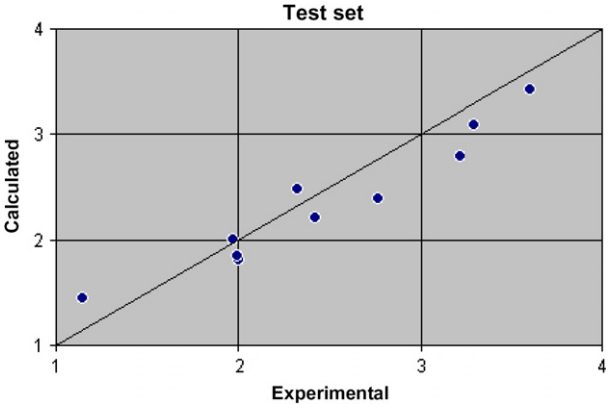

Figure 3. Experimental versus calculated with Eq. 3 values of toxicity towards rats for the test set.

In this study, we decided to use optimal descriptors approach based on the simplified molecular input line entry system (SMILES)[11–13]and additionally to it the multiple linear regression analysis (MLRA) technique to compare performances. A basic idea of the optimal descriptors used in this study has been described in Refs. 14 and 15. It is to be noted that QSPR analysis based on the SMILES notations has been described in Ref. 16.

Previously, the QSPR analysis of alkane properties has shown that even in the case of these simplest hydrocarbons, an overtraining is possible.[15] An overtraining is a situation when statistical characteristics of the QSPR/QSAR model for the training set are better than statistical characteristics for the test set. In order to avoid (or at least to reduce a chance) of the overtraining a 'balance of correlations' principle for training and calibration sets have been utilized.

In this study we significantly extended the dataset to 61 compounds (comparing to our previous papers[4,5]) and the type of compounds were not limited to nitrogroup-containing aromatic compounds as before.[4,5] Recent study included different aromatic compounds with halogen-containing groups, nitrogen-containing groups accompanied by hydroxyl and amino-groups. Similarly to our previous studies the toxicity of the studied species towards rats (LD_{50}, oral) is endpoint under consideration. Numerical data on the endpoint has been taken from web based database of US National Library of Medicine.[17]

The present study is aimed to compare predictive potential of QSAR based on the balance of correlations (using training, calibration and external test sets) and QSAR based on total training and test sets, without calibration set. Additionally, comparison of the models based on optimal descriptors and models obtained by MLRA approach has been performed. A number of descriptors for MLRA have been generated by *DRAGON* software[18,19] and by quantum-chemical calculations using Density Functional Theory.

2. Toxicity data

The dataset for the present study has been collected from a series of 61 compounds that include different aromatic compounds with halogen-containing groups, nitrogen-containing groups accompanied by hydroxyl and amino-groups. The dataset is limited only to benzene derivatives, and number of studied compounds is limited by available acute toxicity data for rats. The activities of the studied compounds are expressed in terms of oral LD_{50} dose for rats and taken from.[17] All original LD_{50} toxicity data (mg/kg) has been converted to molar $\mathrm{Log}(LD_{50})^{-1}$ response variables.

3. Methods

3.1. Optimal descriptors approach

Optimal descriptors have been defined as follows:

$$\mathrm{DCW} = \mathrm{DCW}(\mathrm{SMILES}) = \prod \mathrm{CW}(\mathrm{SA_k}) \qquad (1)$$

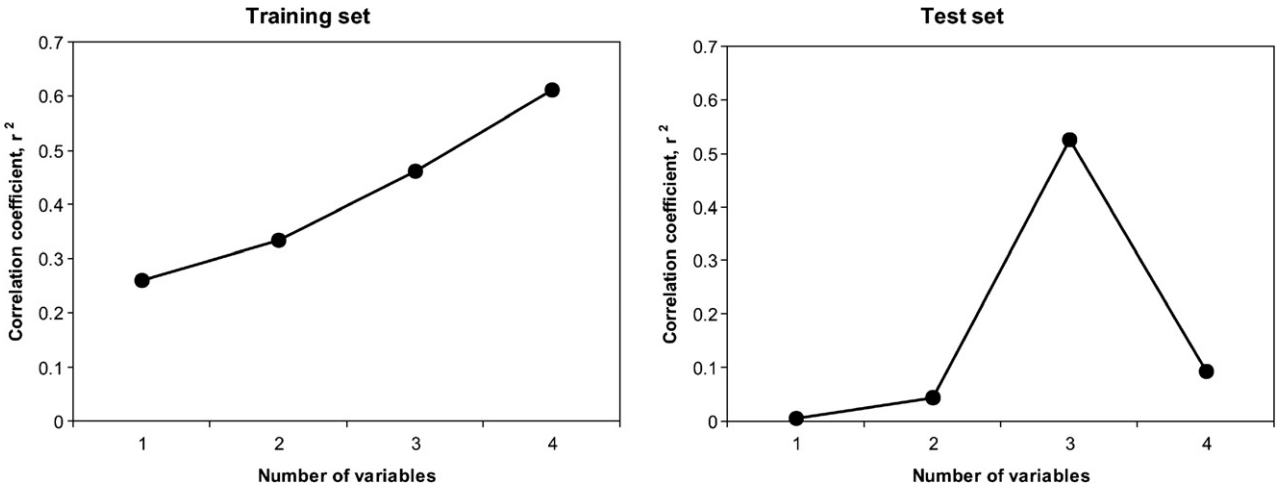

Figure 4. Comparing of correlation coefficients (r^2) of the MLRA models for training and test sets.

Table 1. Statistical characteristics of the models of toxicity based on optimal descriptors

	N	R^2	S	F	N	R^2	S	F	N	R^2	S	F
	Training set				Calibration set				Test set			
1	27	0.791	0.278	95	24	0.791	0.217	83	10	0.930	0.265	106
2	27	0.794	0.289	96	24	0.794	0.216	85	10	0.930	0.258	106
3	27	0.794	0.276	96	24	0.794	0.215	85	10	0.930	0.270	107
					No calibration set							
1	51	0.797	0.244	192					10	0.793	0.353	30
2	51	0.803	0.240	200					10	0.818	0.328	36
3	51	0.817	0.232	218					10	0.836	0.307	41

where SA_k is local or global attribute of SMILES string that encodes molecular structure of given compounds, $CW(SA_k)$ is correlation weights for the SA_k. The values of these parameters have been calculated by hierarchic scheme: first, definition of the four characters attributes ([N+] and [O−]); second, definition of three characters attribute (C = 0), third, definition of two characters attributes (Cl, Br, O=), finally, definition of one character attributes such as C, c, 1, 2, F, O, C. In calculation of the DCW, the parenthesis symbol ')' is replaced by '(' because they are indicators of the same phenomena (branching). In addition to local SMILES attributes the global SMILES attributes have been also utilized for DCW calculation. The number of parentheses in given SMILES fragment denoted as '(xxx'; Number of fluorine atoms denoted as 'Fxxx'; Number of chlorine atoms denoted as 'Clxx'; Number of nitrogen atoms denoted as 'Nxxx'. For instance, the '(001' means presence of one '(' fragment in the SMILES; The 'F002' and 'Cl03' means presence of two fluorine and three chlorine atoms, respectively. If a SMILES attribute SA_k^* is absent in the training set the number of $CW(SA_k^*)$ is assumed equal to 1, in other words influence of the $CW(SA_k^*)$ to the model is blocked.

The described optimizations of correlation weights of invariants in molecular graph[20–32] as well as SMILES fragments[33,34] are search for such values of the correlation weights which produce maximal value of correlation coefficient between modeled endpoint and optimal descriptor for a training set. A predictive potential of obtained models can be estimated with an external test set.

In the present study a novel target function has been used. The target function is defined as

$$\text{Criterion} = R_T + R_C - \text{ABS}(R_T - R_C)\alpha \quad (2)$$

where R_T and R_C are correlation coefficients between the DCW and toxicity for the training and calibration sets, respectively. In other words an attempt to obtain correlation weights which satisfy to conditions: first, R_T and R_C are as large as possible, and second, $|R_T - R_C| \to 0$ has been accomplished.

The comparison of predictive potentials of this novel approaches and the others described in Refs. 20–34 is discussed below. A described application of the training with the calibration sets can be defined as a 'balance of correlations' method. A numerical value of the α has been defined as equal to 0.1. It is a preliminary empirical approximation. The case of the $\alpha = 0$, relates to the case described in Refs. 20–34. In the case of $\alpha \gg 0.1$ there is probability to get similar values of the R_T and R_C, which asymptotically convergent to zero.

3.2. MLRA approach

A set of constitutional, topological, and molecular descriptors were calculated by the DRAGON soft-

Table 2. Correlation weights of the SMILES fragments (SA_k) in case of correlation balance and distribution of the different types of the SA_k in training (N_{TRN}), calibration (N_{CLB}), and test sets (N_{TST})

ID	SA_k	$CW(SA_k)$ in run1	$CW(SA_k)$ in run2	$CW(SA_k)$ in run3	N_{TRN}	N_{CLB}	N_{TST}
1	(	0.9880703	0.9764198	0.9791535	146	150	52
2	1	1.0530526	1.0479764	1.0212917	54	48	20
3	2	0.9888572	0.9765760	1.0035827	6	2	0
4	C	1.0142845	1.0150972	1.0127564	23	35	10
5	Br	1.0333953	1.0381446	1.0229541	1	1	0
6	Cl	1.0384555	1.0376018	1.0405048	28	29	13
7	F	1.0030934	1.0015402	1.0065297	37	34	15
8	O=	0.9598585	0.9483896	0.9638096	13	14	4
9	N	1.0282815	1.0498911	1.0494876	4	11	3
10	O	1.0662223	1.0731354	1.0596372	11	6	1
11	c	0.9868752	0.9892329	0.9837860	178	150	60
12	(000	1.0568560	0.9604498	0.9702864	2	1	0
13	(001	1.0518228	0.9765375	0.9825842	3	3	3
14	(002	1.0517648	0.9975356	1.0055017	11	5	1
15	(003	1.0583493	1.0309084	1.0304712	2	8	3
16	(004	1.0723191	1.0718576	1.0601875	5	3	3
17	(005	0.9946527	1.0069058	1.0113899	2	1	0
18	(006	1.0288649	1.0704023	1.0680352	2	1	0
19	(007	1.0	1.0	1.0	0	1	0
20	(008	1.0	1.0	1.0	0	1	0
21	F000	0.9925902	0.9921891	0.9841599	16	13	4
22	F001	1.0	1.0	1.0	0	1	1
23	F002	1.0541473	1.0648539	1.0291143	1	0	1
24	F003	1.0297308	1.0325473	1.0089149	7	8	4
25	F004	0.9302342	0.9234525	0.9198780	1	1	0
26	F005	1.0657076	1.0750720	1.0341757	2	1	0
27	Cl00	1.0448632	1.0313933	1.0669887	12	10	3
28	Cl01	1.0092429	0.9952787	1.0292403	9	6	4
29	Cl02	1.0021961	0.9932711	1.0160108	3	5	1
30	Cl03	0.9852150	0.9806642	0.9965595	1	1	1
31	Cl04	1.0	1.0	1.0	0	1	1
32	Cl05	1.0399303	1.0447768	1.0343050	2	0	0
33	Cl06	1.0	1.0	1.0	0	1	0
34	N000	0.9914972	0.9841770	1.0005883	12	10	4
35	N001	1.0385932	1.0182994	1.0197261	13	8	5
36	N002	1.0575619	1.0175964	1.0114330	2	4	1
37	N003	1.0	1.0	1.0	0	1	0
38	N004	1.0	1.0	1.0	0	1	0
39	O=C	1.0278297	1.0311915	1.0237725	2	0	0
40	S000	1.0832404	1.0454443	1.0337898	27	24	10
41	[N+]	1.0464742	1.0823108	1.0504987	13	12	4
42	[O−]	1.0458822	1.0466211	1.0558999	13	12	4

ware.[18,19] A set of 1060 molecular descriptors of different kinds was used to describe the chemical diversity of the compounds. The descriptor typology is: (a) constitutional (atom and group fragments); (b) functional groups; (c) atom-centered fragments; (d) empirical; (e) topological; (f) walk counts; (g) various autocorrelations from the molecular graph; (h) Randic molecular profiles from the geometry matrix; (i) geometrical; (j) WHIMs; and (k) GETAWAYs descriptors, and various indicator descriptors. The meaning of these molecular descriptors and the calculation procedures are summarized elsewhere.[19]

Considering the importance of electronical molecular properties for QSAR/QSPR the additional descriptors are calculated—the quantum chemical descriptors.[35–37] Quantum-chemical descriptors have been calculated using Gaussian03 software[36] by Density Functional Theory methodology at the B3LYP/6-31G(d,p) level.

The correlation coefficients for all pair of descriptor variables used in the models were evaluated in order to identify highly correlated descriptors and to detect redundancy in the data set. Any type of redundancy might lead to an overexploitation of a chemical property in the explanation of the dependent variable. Hence, some highly correlated and constant descriptors (cross-correlation $r^2 > 0.9$) were removed from the further consideration. Furthermore, at the process of each model building (i.e., inside of each model) the descriptors with cross-correlation coefficient more than 0.6 are avoided.

The correlation between biological activity and structural properties was obtained by using a variable selection Genetic Algorithm (GA) and MLRA methods. GAs have been applied in recent years as a powerful tool to address many problems in drug design.[38,39] We applied GA to select, from all of the calculated descriptors only the best combinations of those most relevant for

Table 3. Correlation weights of the SMILES fragments (SA_k) in case of maximization of correlation coefficient between toxicity and the DCW over total training set that contains both the training ($n = 27$) and calibration ($n = 25$) sets and distribution of the different types of the SA_k in the total training (N_{TRN}) and test sets (N_{TST})

ID	SA_k	$CW(SA_k)$ in run1	$CW(SA_k)$ in run2	$CW(SA_k)$ in run3	N_{TRN}	N_{TST}
1	(	1.0276256	0.9972809	1.0436414	296	52
2	1	2.4463488	1.6359240	1.8941774	102	20
3	2	0.9433184	1.1685853	0.8594918	8	0
4	C	1.0002822	1.0180630	1.0327060	58	10
5	Br	1.2724119	1.2660093	1.3618481	2	0
6	Cl	0.9532818	1.0945404	0.9347042	57	13
7	F	0.9895725	0.9487478	0.8811733	71	15
8	O=	0.7474085	0.8028133	0.7192619	27	4
9	N	1.0018736	1.1271245	0.9874202	15	3
10	O	1.2799951	1.3128133	1.3992817	17	1
11	c	0.9464335	0.8704898	0.9514151	328	60
12	(000	1.2795321	1.3853821	1.7260898	3	0
13	(001	1.0333221	1.1691273	1.2938916	6	3
14	(002	0.9125820	1.0929377	1.0644908	16	1
15	(003	0.9545480	1.1557480	1.0476771	10	3
16	(004	1.0132886	1.2677634	1.0632671	8	3
17	(005	0.6044526	0.8010361	0.5455415	3	0
18	(006	0.6871017	0.9502461	0.6086850	3	0
19	(007	0.8750000	0.8823810	0.7965168	1	0
20	(008	0.8750000	0.7952312	0.7625324	1	0
21	F000	1.0162946	0.8416272	0.9208133	29	4
22	F001	0.9000000	0.8406559	0.9353365	1	1
23	F002	1.2987071	1.2276563	1.6465696	1	1
24	F003	1.0472735	1.0550265	1.4109339	15	4
25	F004	0.7147366	0.7191233	0.9243081	2	0
26	F005	1.0498907	1.2028628	1.9392530	3	0
27	Cl00	0.9132673	1.0979686	0.7808835	22	3
28	Cl01	0.9227452	1.0061981	0.8187553	15	4
29	Cl02	1.0809374	1.0391437	1.0326652	8	1
30	Cl03	1.2400465	1.0168438	1.2047289	2	1
31	Cl04	1.5375000	1.1526008	1.6889222	1	1
32	Cl05	1.6923412	1.1546570	2.0357822	2	0
33	Cl06	2.0250000	1.2037291	2.5196408	1	0
34	N000	0.9010177	0.8802293	0.7746912	22	4
35	N001	1.2611401	1.1168402	1.2310495	21	5
36	N002	1.3063073	1.0750090	1.3924698	6	1
37	N003	0.9875000	0.9392532	0.8431996	1	0
38	N004	0.8625000	0.8916112	0.7799055	1	0
39	O=C	1.0598952	1.0672631	1.1033048	2	0
40	S000	2.4750575	1.6307267	3.0300762	51	10
41	[N+]	1.0804514	1.0410330	1.3149298	25	4
42	[O−]	1.2628293	1.3923449	1.0923021	25	4

obtaining models with the highest predictive power of toxicity. Finally, the combination GA-MLRA technique was utilized to select the appropriate descriptors and to generate different QSAR models. The GA technique started with a population of 30 random models and 500 iterations for evolution. For GA analysis and the derivation of the QSAR models, the BuildQSAR program[40] has been used.

4. Results

4.1. Optimal descriptor models

Statistical characteristics of the models based on optimal descriptors are presented in Table 1. One can see from the table that the statistical characteristics of the models obtained by balance of correlations are better (training, calibration and test sets) in comparison with models obtained for two sets (training and test, without the calibration set).

Tables 2 and 3 contain numerical data on correlation weights obtained by two ways: by correlation balance and by training without the calibration.

It is clear that if $CW(SA_k) > 1$ the SA_k can be considered as a promoter of an increase in toxicity values. Taking into account probabilistic logic one can state that if $CW(SA_k) > 1$ in all three runs of the Monte Carlo optimization the hypothesis becomes more plausible. Also similar logic may be utilized for case $CW(SA_k) < 1$, in this case one can consider SA_k as a promoter of toxicity lowering. Finally, if CW(SAk) has both larger and lower

Table 4. Model based on optimal descriptors obtained in first run of the Monte Carlo optimization

CAS No.	SMILES	DCW	Expr	Calc	Expr − Calc
Training set					
98-08-8	FC(F)(F)c1ccccc1	1.2148620	0.98900	1.62204	−0.63304
344-07-0	Fc1c(F)c(F)c(F)c(Cl)c1F	1.2157526	1.60700	1.63078	−0.02378
350-57-2	FC(F)(Oc1ccccc1)C(F)F	1.1695817	1.58900	1.17778	0.41122
98-46-4	FC(F)(F)c1cccc(c1)[N+]([O−])=O	1.2991408	2.49600	2.44893	0.04707
392-83-6	FC(F)(F)c1ccccc1Br	1.2554327	1.91800	2.02009	−0.10209
440-60-8	Fc1c(CO)c(F)c(F)c(F)c1F	1.3107764	2.34300	2.56309	−0.22009
446-35-5	O=[N+]([O−])c1ccc(F)cc1F	1.3451660	2.90100	2.90050	0.00050
458-24-2	FC(F)(F)c1cccc(CC(C)NCC)c1	1.3650504	3.25000	3.09559	0.15441
15457-05-3	[O−][N+](=O)c1ccc(cc1)Oc2ccc(cc2[N+]([O−])=O)C(F)(F)F	1.2240780	1.56200	1.71246	−0.15046
42818-60-0	FC(F)(F)c1cc(cc(N)c1)C(C)O	1.2631071	2.36800	2.09539	0.27261
42874-03-3	Clc2cc(ccc2Oc1ccc([N+]([O−])=O)c(OCC)c1)C(F)(F)F	1.2589885	1.85900	2.05498	-0.19598
98-95-3	[O−][N+](=O)c1ccccc1	1.2894665	2.54700	2.35401	0.19299
97-00-7	O=[N+]([O−])c1cc(ccc1Cl)[N+]([O−])=O	1.3269388	2.50000	2.72166	−0.22166
121-73-3	O=[N+]([O−])c1cc(Cl)ccc1	1.2626587	2.57400	2.09099	0.48301
100-14-1	ClCc1ccc(cc1)[N+]([O−])=O	1.2806951	1.97700	2.26795	−0.29095
581-89-5	[O−][N+](=O)c1ccc2ccccc2c1	1.1959862	1.59500	1.43684	0.15816
59-50-7	Oc1ccc(Cl)c(C)c1	1.2408550	1.89200	1.87706	0.01494
82-68-8	Clc1c(c(Cl)c(Cl)c(Cl)c1Cl)[N+]([O−])=O	1.3314534	2.42900	2.76596	−0.33696
88-72-2	O=[N+]([O−])c1ccccc1C	1.3078859	2.18700	2.53473	−0.34773
95-50-1	Clc1ccccc1Cl	1.2473955	2.46800	1.94124	0.52676
95-76-1	Nc1cc(Cl)c(Cl)cc1	1.2744546	2.47300	2.20672	0.26628
95-82-9	Clc1ccc(Cl)c(N)c1	1.2744546	2.00500	2.20672	−0.20172
95-88-5	Oc1ccc(Cl)c(O)c1	1.3043946	2.59300	2.50047	0.09253
98-07-7	ClC(Cl)(Cl)c1ccccc1	1.2251362	1.51300	1.72284	−0.20984
74-11-3	O=C(O)c1ccc(Cl)cc1	1.2574259	2.12700	2.03965	0.08735
87-86-5	Clc1c(O)c(Cl)c(Cl)c(Cl)c1Cl	1.4245514	3.99400	3.67938	0.31462
89-98-5	O=Cc1ccccc1Cl	1.2433129	1.81300	1.90118	−0.08818
Calibration set					
98-16-8	FC(F)(F)c1cc(N)ccc1	1.2855206	2.52600	2.31529	0.21071
320-50-3	FC(F)(F)c1cc(Cl)ccc1Cl	1.2344744	1.66200	1.81446	−0.15246
328-84-7	Clc1ccc(cc1Cl)C(F)(F)F	1.2344744	1.87100	1.81446	0.05654
363-72-4	Fc1cc(F)c(F)c(F)c1F	1.2253223	1.92500	1.72467	0.20033
368-53-6	Nc1cc(cc(N)c1)C(F)(F)F	1.3314415	2.66600	2.76584	−0.09984
393-75-9	O=[N+]([O−])c1cc(cc([N+]([O−])=O)c1Cl)C(F)(F)F	1.2748084	2.46400	2.21019	0.25381
460-00-4	Fc1ccc(Br)cc1[a]	1.2237014	1.81200	1.70876	0.10324
1582-09-8	[O−][N+](=O)c1cc(cc([N+]([O−])=O)c1N(CCC)CCC)C(F)(F)F[a]	1.2767417	2.24000	2.22916	0.01084
29091-05-2	FC(F)(F)c1cc(c(c([N+]([O−])=O)c1N)N(CC)CC)[N+]([O−])=O[a]	1.2458656	2.03100	1.92622	0.10478
54910-89-3	FC(F)(F)c2ccc(OC(CCNC)c1ccccc1)cc2	1.2962155	2.57400	2.42023	0.15377
61988-37-2	FC(F)(Oc1ccc(N)cc1N)C(F)F	1.3047869	2.32100	2.50432	−0.18332
95-69-2	Cc1cc(Cl)ccc1N	1.2840665	2.12700	2.30103	−0.17403
100-00-5	O=[N+]([O−])c1ccc(Cl)cc1	1.2626587	2.57400	2.09099	0.48301
99-54-7	Clc1ccc(cc1Cl)[N+]([O−])=O	1.3020596	2.30400	2.47756	−0.17356
89-61-2	O=[N+]([O−])c1cc(Cl)ccc1Cl	1.3020596	2.28300	2.47756	−0.19456
62-23-7	O=[N+]([O−])c1ccc(cc1)C(O)=O	1.2836974	1.93100	2.29741	−0.36641
51-28-5	O=[N+]([O−])c1cc(ccc1O)[N+]([O−])=O	1.4105045	3.78800	3.54156	0.24644
68-36-0	ClC(Cl)(Cl)c1ccc(cc1)C(Cl)(Cl)Cl[a]	1.2429587	1.99000	1.89770	0.09230
95-49-8	Cc1ccccc1Cl	1.2269280	1.51100	1.74042	−0.22942
95-73-8	Clc1cc(Cl)c(C)cc1	1.2001020	1.82700	1.47722	0.34978
95-79-4	Nc1cc(Cl)ccc1C	1.2840665	2.48500	2.30103	0.18397
95-94-3	Clc1cc(Cl)c(Cl)cc1Cl[a]	1.2731555	2.15800	2.19398	−0.03598
85-34-7	Clc1c(CC(O)=O)c(Cl)ccc1Cl	1.2493531	2.12900	1.96044	0.16856
88-04-0	Oc1cc(C)c(Cl)c(C)c1	1.2364225	1.61200	1.83358	−0.22158
External test set					
327-92-4	O=[N+]([O−])c1cc(c(F)cc1F)[N+]([O−])=O	1.3983359	3.61100	3.42217	0.18883
98-56-6	FC(F)(F)c1ccc(Cl)cc1	1.1971186	1.14300	1.44795	−0.30495
121-17-5	O=[N+]([O−])c1cc(ccc1Cl)C(F)(F)F	1.3031079	2.32200	2.48785	−0.16585
320-60-5	FC(F)(F)c1ccc(Cl)cc1Cl	1.2344744	2.00000	1.81446	0.18554
371-40-4	Fc1ccc(N)cc1[a]	1.2754839	2.42600	2.21682	0.20918
3239-44-9	FC(F)(F)c1cccc(CC(C)NCC)c1	1.3650504	3.30400	3.09559	0.20841
88-73-3	O=[N+]([O−])c1ccccc1Cl	1.2934041	2.76900	2.39264	0.37636
58-90-2	Clc1cc(Cl)c(Cl)c(Cl)c1O[a]	1.3335684	3.21900	2.78671	0.43229
95-74-9	Nc1cc(Cl)c(C)cc1	1.2535431	1.97500	2.00155	−0.02655
87-61-6	Clc1cccc(Cl)c1Cl	1.2372938	1.99600	1.84212	0.15388

[a] SMILES notations which contain SA_k absent in the training set.

Table 5. Calculation of the DCW with CWs of first run the correlation balance approach (Table 2) for SMILES of 'O=[N+]([O−])c1cc(c(F)cc1F)[N+]([O−])=O '(CAS = 327-92-4); DCW = 1.3983359

SA_k	$CW(SA_k)$ in run 1	ID
O=	0.9598585	8
[N+]	1.0464742	41
(	0.9880703	1
[O−]	1.0458822	42
(	0.9880703	1
c	0.9868752	11
1	1.0530526	2
c	0.9868752	11
c	0.9868752	11
(	0.9880703	1
c	0.9868752	11
(	0.9880703	1
F	1.0030934	7
(	0.9880703	1
c	0.9868752	11
c	0.9868752	11
1	1.0530526	2
F	1.0030934	7
(	0.9880703	1
[N+]	1.0464742	41
(	0.9880703	1
[O−]	1.0458822	42
(	0.9880703	1
O=	0.9598585	8
(004	1.0723191	16
F002	1.0541473	23
Cl00	1.0448632	27
S000	1.0832404	40
N002	1.0575619	36

than 1 values over the three runs one can conclude that the SA_k has undefined role. This analysis is of probabilistic nature, however we believe that it can help in search for mechanistic interpretation of toxicity.[25,30]

In this work, separation into training, calibration, and test sets has been done randomly, the only condition was a similarity of diapasons in variation of the toxicity over each set. It is to be noted that in the case of the correlation balance some SA_k are present in the calibration set and they are absent in the training set. It was noted above that the correlation weights of these attributes have been assumed equal to 1. In the case of the model built without the calibration all SA_k are present in the total training set, however, their correlation weights do not improve the statistical characteristics of the model for external test set. The model that has been obtained in the first run of the Monte Carlo optimization by means of the balance of correlations is described as follows

$$\mathrm{Log}(\mathrm{LD}_{50}) = -10.2974(\pm 0.3341) + 9.8114(\pm 0.2609) * \mathrm{DCW} \quad (3)$$

$$n = 27,\ R^2 = 0.7910,\ s = 0.278,\ F = 95 \text{ (training set)}$$
$$n = 24,\ R^2 = 0.7910,\ s = 0.217,\ F = 83 \text{ (calibration set)}$$
$$n = 10,\ R^2 = 0.9296,\ s = 0.265,\ F = 106 \text{ (test set)}$$

Actual and predicted by the Eq. (3) values of the toxicity towards rats are present in the Table 4. An example of the calculation of the DCW with Eq. (1) is demonstrated in Table 5. Graphically correlations between experimental and calculated values of the toxicity are demonstrated in Figures 1–3 for the training, calibration, and test sets, respectively.

A rational definition of the applicability domain for a model is an important task of the QSPR/QSAR analysis.[33,34] In the case of the optimal descriptors calculated with Eq. (1), by the balance of correlations, reasonable definition of the domain of applicability can be formulated as follows: any substances which have SMILES containing majority of the SA_k which contribute in both sets of the training and calibration at least three times (probabilistic logic).

4.2. MLRA models

The MLRA approach has been performed with two purposes: (1) to find structural and electronic features responsible for toxicity exhibition by benzene derivatives and (2) to compare the predictive potentials of two approaches (optimal descriptor and MLRA) using technique with splitted datasets to training and test sets. Statistical characteristics of the one-, two-, three-, and four-variable models obtained by means of the MLRA combined with the variable selection GA are collected in Table 6.

Table 6. Statistical characteristics of one-, two-, three-, and four-variable models of the toxicity towards rats

Model	Variables	Training set (n = 51)			Test set (n = 10)		
		R^2	F	s	R^2	F	s
1 Variable	[a]Mor20e	0.259	17.12	0.471	0.006	0.05	0.798
2 Variables	Mor20e, [b]Dm	0.334	12.05	0.451	0.044	0.37	0.782
3 Variables	[c]C-024, [d]N-067, Dm	0.462	13.44	0.410	0.525	8.85	0.551
4 Variables	[e]nC, [f]RDF045p, N-067, Dm	0.612	18.16	0.352	0.092	0.81	0.763

[a] Mor20e – 3D-MoRSE-signal 20/weighted by atomic Sanderson electronegativities.
[b] Dm – dipole moment, a quantum-chemical descriptor.
[c] C-024 – number of R—CH—R fragments.
[d] N-067 – number of NH fragments.
[e] nC – total number of carbon atoms in molecule.
[f] RDF045p – Radial Distribution Function—4.5/weighted by atomic polarizabilities.

The best three-variable MLRA model is as follows:

$$\begin{aligned}\mathrm{Log}(\mathrm{LD}_{50})^{-1} = &-0.2066(\pm 0.0574)\mathrm{C\text{-}024}\\ &+ 1.2080(\pm 0.1770)\mathrm{N\text{-}067}\\ &+ 0.2228(\pm 0.0614)\mathrm{Dm}\\ &+ 2.1932(\pm 0.2199) \qquad (4)\end{aligned}$$

$$\begin{aligned}&n = 51,\ R^2 = 0.462,\ q^2 = 0.353,\ s = 0.410,\\ &F = 13.44 \text{ (training set)}\\ &n = 10,\ R^2 = 0.525,\ s = 0.551, F = 8.85 \text{ (test set)}\end{aligned}$$

One can see from Table 6 that the best statistical characteristics for the pair training-test sets are obtained in the case of the three-variable model. The model (4) is based on the following descriptors: C-024, N-067, and Dm, where is C-024 is number of R—CH—R fragments (which is involved in aromatic cycles), N-067 is the number of NH fragments and Dm is a dipole moment, calculated at the DFT (B3LYP/6-31G(d,p)) level.

5. Discussion

Three criterions are important in QSAR analysis where considering the influence of the correlation weights (CW) of SMILES attributes. First, stability of the numerical values (i.e., CW more than unit or CW less than unit over all three probes); second, a total number of the given CW at the training: if this number is small, then influence of the CW for the model is problematic; finally, a total number of the given CW in the test set. Again if a total number of CW in test set is small, then influence of the CW on statistical quality of the model should be questioned.

According to these criterions, the SMILE fragment '(' (ID 1) has CW at three runs 0.9880703, 0.9764198, and 0.9791535, respectively; and numbers of the SMILES attribute are 146, 150, and 52 for the training, calibration, and test sets, respectively; and therefore this fragment should be defined as promoter of toxicity lowering (CW is less then 1). The similar logic for SMILE fragment '1' (indicator of the cycle, ID 2), the results leads to conclusion that presence of the '1' is increasing of the toxicity (CW is higher then 1). However, the N000 (ID 34) has undefined influence to toxicity, because for the two runs CW(N000) <1, but in the third run CW(N000) > 1. It should be noted that '2' (ID 3) and 'Br' (ID 5) are not informative for the SMILES-based modeling, because of these attributes are rare. Taking into account these abovementioned rules, a chemist, who familiar with SMILES notations can determine the factors influencing on toxicity from obtained results.

We have compared the performances of one-, two-, three-, and four-variable models constructed by GA-MLRA approach (Fig. 4). As it can be seen, according to these results, only three-variable model showing satisfactory predictive ability ($r^2_{\mathrm{test}} = 0.525$), while the other models fails to give significant prediction values (for one- and two-variable models it can be explained by insufficient information providing by number of descriptors in the model, and for four-variable model it is explained by overfitting that lead to bad prediction.

In the one-variable MLRA model the descriptor Mor20e is a 3-D MoRSE descriptor (3-D Molecular Representation of Structures based on Electron diffraction). 3-D MoRSE descriptors are based on the idea of obtaining information from the 3-D atomic coordinates by the transformation used in electron diffraction studies for preparing theoretical scattering curves.[41] Mor20e descriptor is calculated by summing up the atomic weights viewed by different angular scattering functions and weighted by atomic electronegativities.[19,42]In our case the electronegativity contribution to this descriptor plays main role. And according to the model the higher total electronegativity of molecule the more toxic molecule. The second important descriptor – Dm, dipole moment, involved in three-variable best model (4), and its presence confirming importance of electronegative groups of the studied compounds regarding to toxicity. Electronegative groups increasing dipole moment and consequently intensify toxicity of the studied compounds. C-024 and N-067 descriptors are number of R—CH—R groups and number of NH fragments, respectively. Increasing in number of R—CH—R fragments lowering toxicity value, whereas increasing number of NH groups making compound much more toxic. Interestingly, total number of carbons in molecule nC, also plays certain role, the more carbon atoms the less toxic compound. The member of four-variable model, RDF045p descriptor is a Radial Distribution Function 4.5 descriptor, weighted by atomic polarizabilities. The contribution of this descriptor to the toxicity is less than other descriptors (according to regression coefficients), although effect of this descriptor to toxicity also determines by charges distribution, polarizability, like in case of Dm and Mor20e.

Interestingly, the importance of the number of aromatic fragments in toxicity exhibition has been showed by both methods (descriptor 'c' in optimal descriptors and descriptor C-024 in MLRA). The effect on toxicity increasing of the nitrogen-containing groups presence confirmed by both approaches – SMILE fragment 'N' in optimal descriptors approach and N-067 descriptor in MLRA.

As it was mentioned in the introduction part, in this study we tested an idea of predictive QSAR performances for two cases: (1) splitted dataset to training-calibration-test sets and (2) splitted dataset to training-test sets. Both tests have been performed only for SMILES based optimal descriptors analyses. Looking at the obtained prediction results for 10 compounds we can see that in the case when calibration set used the results is much better. In this case the r^2 for the external test set is much better and F-ratio also showing the significant improving of the model robustness in comparing to dataset splitted only to training-test sets.

An additional comparison of the performances of SMILES based optimal descriptor approach and GA-

MLRA analysis have been performed for the case of splitted dataset to training and test sets. The MLRA analysis showed much lower r^2 values then optimal descriptor approach. Only three-variable model showed satisfactory predictive ability with $r^2 = 0.525$, when the optimal descriptors showed the $r^2 = 0.793$.

6. Conclusions

The balance of correlations for the 'sub training' ($n = 27$) and calibration ($n = 24$) sets provides the statistical quality improvement of the prediction of oral toxicity towards rats for the external test set ($n = 10$) in comparison to the training procedure with combined set that contains both the 'sub training' and calibration ($n = 51$) sets.

The models based on optimal descriptors calculated with SMILES have better statistical characteristics than one-, two-, three-, and four-variable MLR models based on the topological and quantum chemical descriptors. But we can not declare that the optimal descriptors are better tools for the QSAR modeling. We think, however, that the comparison of two mentioned approaches can be useful in QSAR analysis.

Overall, the QSAR analysis showed that the electronegativity of the molecule plays an important role in acute toxicity of benzene derivatives studied; presence of electronegative groups increasing toxicity and involved in the best model the Dipole moment descriptor confirms it. The presence of nitrogen-containing groups (mostly NH groups) increasing the toxicity that confirmed by both approaches.

Acknowledgements

The authors thank for support the High Performance Computational Design of Novel Materials (HPCDNM) Project funded by the U.S. Department of Defense through the U. S. Army Engineer Research and Development Center (Vicksburg, MS) contract #W912HZ-06-C-0057.

References and notes

1. Kriek, E. In *Environmental Carcinogenesis*; Emmelot, P., Kriek, E., Eds.; Elsevier: Amsterdam, 1979; p 143.
2. Won, W. D.; di Salvo, L. H.; Ng, J. *Appl. Environ. Microbiol.* **1976**, *31*, 576.
3. Soffers, A. E. M. F. M.; Boersma, G.; Vaes, W. H. J.; Vervoort, J.; Tyrakowska, B.; Hermens, J. L. M.; Rietjens, I. M. C. M. *Toxicol. In Vitro* **2001**, *15*, 539.
4. Isayev, O.; Rasulev, B. F.; Gorb, L.; Leszczynski, J. *Mol. Divers.* **2006**, *10*, 233.
5. Toropov, A. A.; Rasulev, B. F.; Leszczynski, J. *QSAR Comb. Sci.* **2007**, *5*, 686.
6. Julie, R.; Seward, J. R.; Sinks, G. D.; Schultz, T. W. *Aquat. Toxicol.* **2001**, *53*, 33.
7. Randic, M.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 614.
8. Basak, S. C.; Gute, B. D.; Mills, D.; Hawkins, D. M. *J. Mol. Struct. (Theochem.)* **2003**, *622*, 127.
9. Mekenyana, O.; Nikolova, N.; Schmieder, P. *J. Mol. Struct. (Theochem.)* **2003**, *622*, 147.
10. Cronin, M. T. D.; Netzeva, T. I.; Dearden, J. C.; Edwards, R.; Worgan, A. D. P. *Chem. Res. Toxicol.* **2004**, *17*, 545.
11. Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
12. Weininger, D.; Weininger, A.; Weininger, J. L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97.
13. Weininger, D. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 237.
14. Toropov, A. A.; Toropova, A. P. *Russ. J. Coord. Chem.* **1998**, *24*, 81.
15. Toropov, A. A.; Toropova, A. P. *J. Mol. Struct. (Theochem.)* **2003**, *637*, 1.
16. Vidal, D.; Thormann, M.; Pons, M. *J. Chem. Inf. Model.* **2005**, *45*, 386–393.
17. http://toxnet.nlm.nih.gov/.
18. Todeschini, R.; Consonni, V. *DRAGON software for the Calculation of Molecular Descriptors*, web version 3.0 for Windows, **2003**.
19. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley VCH: Weinheim and New York, 2000.
20. Toropov, A. A.; Schultz, T. W. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 560.
21. Toropov, A. A.; Toropova, A. P. *J. Mol. Struct. (Theochem.)* **2002**, *578*, 129.
22. Toropov, A. A.; Toropova A. P.; *J. Mol. Struct. (Theochem)* **2001**, 538, p. 197.
23. Toropov, A. A.; Roy, K. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 179.
24. Toropov, A. A.; Benfenati, E. *J. Mol. Struct. (Theochem.)* **2004**, *676*, 165.
25. Toropov, A. A.; Benfenati, E. *J. Mol. Struct. (Theochem.)* **2004**, *679*, 225.
26. Toropov, A. A.; Benfenati, E. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1941.
27. Toropov, A. A.; Benfenati, E. *Bioorg. Med. Chem.* **2006**, *14*, 2779.
28. Toropov, A. A.; Benfenati, E. *Bioorg. Med. Chem.* **2006**, *14*, 3923.
29. Toropova, A. P.; Toropov, A. A.; Maksudov, S. Kh. *Chem. Phys. Lett.* **2006**, *428*, 183.
30. Raska, I., Jr.; Toropov, A. *Bioorg. Med. Chem.* **2005**, *13*, 6830.
31. Raska, I., Jr.; Toropov, A. *Eur. J. Med. Chem.* **2006**, *4*, 1271.
32. Toropov, A. A.; Leszczynski, J. *Chem. Phys. Lett.* **2006**, *433*, 125.
33. Toropov, A.; Nesmerak, K.; Raska, I., Jr.; Waisser, K.; Palat, K. *Comput. Biol. Chem.* **2006**, *30*, 434.
34. Toropov, A. A.; Toropova, A. P.; Mukhamedzhanova, D. V.; Gutman, I. *Indian J. Chem.* **2005**, *44A*, 1545.
35. Hehre, W. J.; Radom, L.; Schleyer, P.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.
36. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, Jr., J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.;

Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. Gaussian 98, Revision A.11, Gaussian, Pittsburgh, PA, **1998**.
37. Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
38. Davis, L. *Handbook of Genetic Algorithms*; Van Nostrand Reinhold: New York (USA), 1991.
39. Devillers, J. *Genetic Algorithms in Molecular Modeling*; Academic Press: London, 1996.
40. de Oliveira, D. B.; Gaudio, A. C. *Quant. Struct.-Act. Relat.* **2000**, *19*, 599.
41. Soltzberg, L. J . L. J.; Wilkins, C. L. *J. Am. Chem. Soc.* **1977**, *99*, 439.
42. Schuur, J. J.; Selzer, P.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334.