

# Simplified Molecular Input Line Entry System-Based Optimal Descriptors: Quantitative Structure–Activity Relationship Modeling Mutagenicity of Nitrated Polycyclic Aromatic Hydrocarbons

Andrey A. Toropov<sup>1,2,\*</sup>, Alla P. Toropova<sup>1,2</sup> and Emilio Benfenati<sup>2</sup>

<sup>1</sup>Institute of Geology and Geophysics, Khodzhibaev St. 49, 100041 Tashkent, Uzbekistan

<sup>2</sup>Istituto di Ricerche Farmacologiche, Mario Negri, Via La Masa 19, 20156 Milano, Italy

\*Corresponding author: Andrey A. Toropov, aatoropov@yahoo.com

**We developed a new QSAR model, based on the optimal descriptors, calculated with simplified molecular input line entry system. These descriptors are correlated with mutagenic potential for a training set and correlated with this end-point for a test set. Statistical characteristics of the model are  $n = 28$ ,  $r^2 = 0.902$ ,  $q^2 = 0.892$ ,  $s = 0.554$ ,  $F = 240$  (training set) and  $n = 20$ ,  $r^2 = 0.853$ ,  $q^2 = 0.823$ ,  $s = 0.702$ ,  $F = 105$  (test set).**

**Key words:** mutagenic potency, optimal descriptor, QSAR, SMILES

Received 30 June 2008, revised 12 February 2009 and accepted for publication 22 February 2009

Aromatic amines are a class of ubiquitous environmental pollutants. They are found, e.g. in tobacco smoke, diesel exhaust and tar, and they are used for preparation of industrial products as azo dyes, pesticides, synthetic materials or pharmaceutical products (1). Unfortunately, many aromatic amines are mutagens. Thus, the biological activity in general and mutagenic potentials of aromatic amines in particular are important pieces of information from many points of view, e.g. ecology, human health, risk assessment. There are a number of studies dedicated to mutagenicity models using quantitative structure–activity relationships (QSAR) (2–4).

Polycyclic aromatic hydrocarbons (PAHs), in particular nitrated PAHs (nitro-PAHs) are widespread environmental pollutants found in the exhaust fumes of gasoline and diesel combustion engines, in certain food products as a result of incomplete combustion and in general, in combustion source emissions (5).

Usually, QSARs are based on elucidation of molecular structure by molecular graph. In the last decade, the representation of the molecular structure by simplified molecular input line entry system (SMILES) has become available from internet. The SMILES reflects presence in molecular structure variety of influence attributes, such as functional groups, double/triple bonds, chirality, etc. Thus, SMILES-based QSAR analyses become a tempting alternative to other types of QSAR based on the molecular graph.

The aim of the present report was the estimation of the ability of optimal descriptors calculated with SMILES for the QSAR analysis of the TA100 mutagenicity.

## Method

### Data

Table 1 shows structural details of the 48 nitrated polycyclic hydrocarbons used in this study. Their mutagenic activity (log TA100, i.e. the decimal logarithm of the mutagenic activity) is taken from (5). Three splits into training and test sets were examined. These splits were random but the range of end-points for these sets was almost identical (Table 2). Canonical SMILES for this study have been generated with ACD/ChemSketch<sup>®</sup>.

The SMILES-based optimal descriptors of correlation weights (DCW) were calculated as

$$DCW = CW(b)CW(db)CW(Nn)CW(No)ICW(SS_k) \quad (1)$$

where  $b$  is a number of brackets;  $db$  is a number of double bonds indicated by '=';  $No$  and  $Nn$  are numbers of oxygen and nitrogen atoms, respectively;  $ss_k$  are SMILES attributes ( $SA_k$ ) of two elements. The element SMILES can be a symbol of the SMILES notation (for instance, 'c', 'C', 'n', 'N', =, etc.), or two symbols which are necessary to encode an image (for instance, 'Cl', 'Br', 'N+', etc.); the  $CW(x)$  is the correlation weight for the SMILES attribute  $x$ . Those  $CW$ s are calculated by the Monte Carlo method optimization procedure (6,7) that provides  $CW$ s values which, used in eqn (1) give a maximum correlation coefficient between the descriptor and log TA100. We used the range of the SMILES elements according

**Table 1:** Structures of nitrated polycyclic hydrocarbons

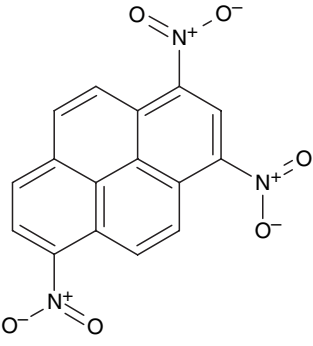
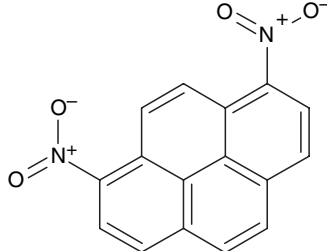
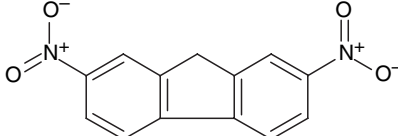
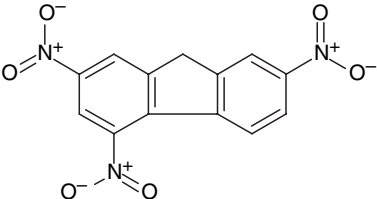
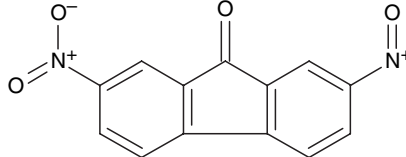
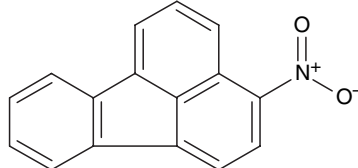
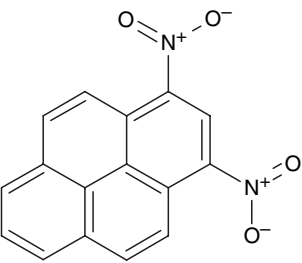
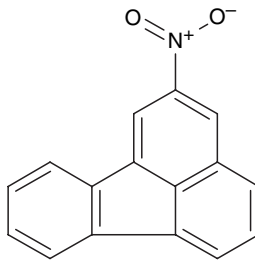
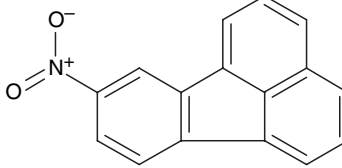
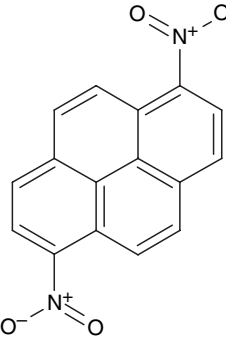
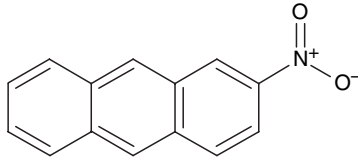
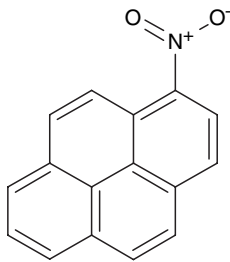
 <p>1</p>	 <p>5</p>	 <p>9</p>
 <p>2</p>	 <p>6</p>	 <p>10</p>
 <p>3</p>	 <p>7</p>	 <p>11</p>
 <p>4</p>	 <p>8</p>	 <p>12</p>

Table 1: (Continued)

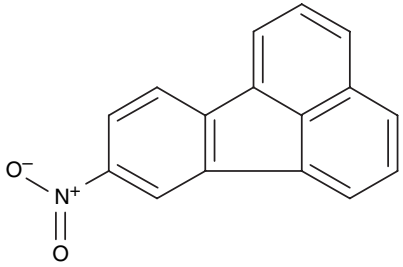
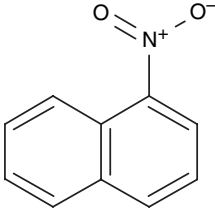
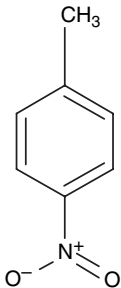
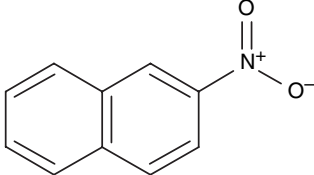
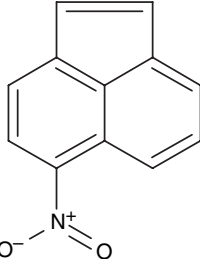
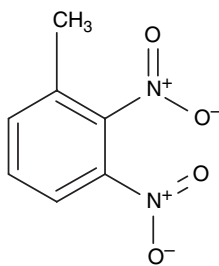
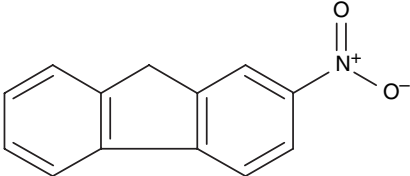
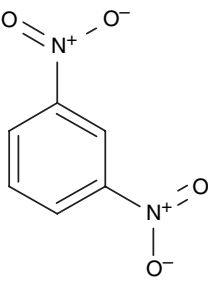
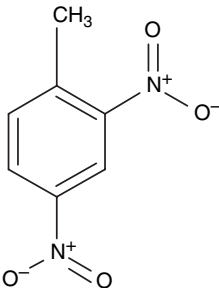
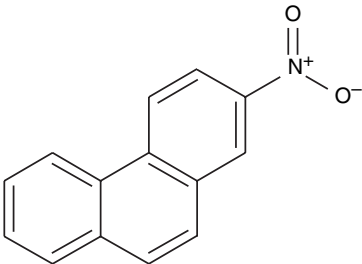
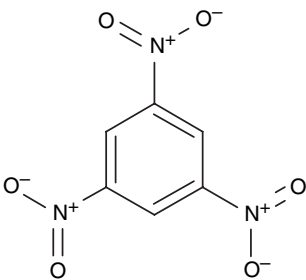
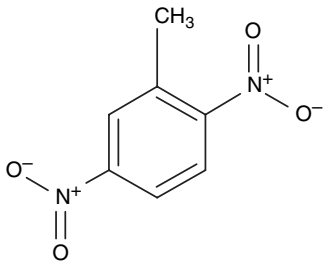
 <p>13</p>	 <p>17</p>	 <p>21</p>
 <p>14</p>	 <p>18</p>	 <p>22</p>
 <p>15</p>	 <p>19</p>	 <p>23</p>
 <p>16</p>	 <p>20</p>	 <p>24</p>

Table 1: (Continued)

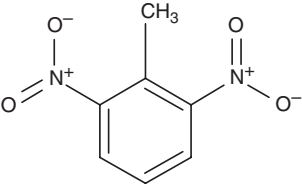
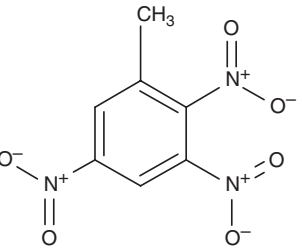
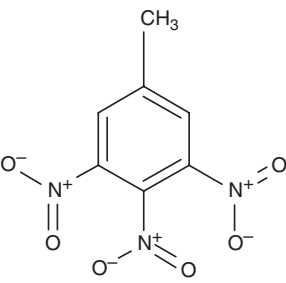
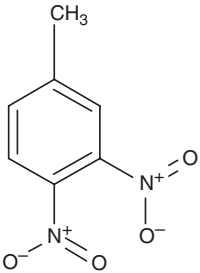
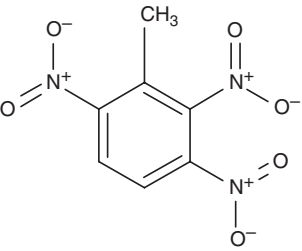
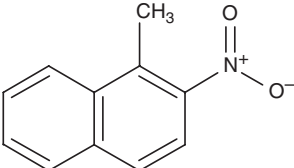
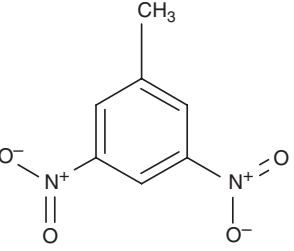
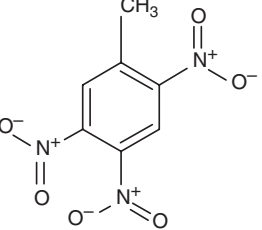
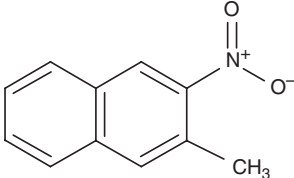
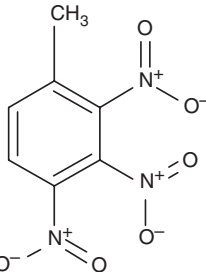
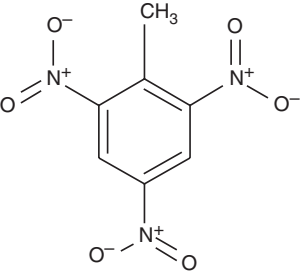
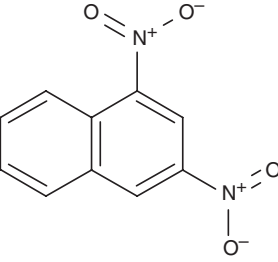
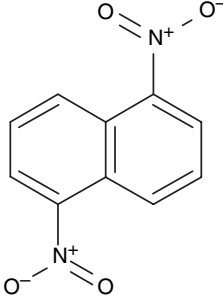
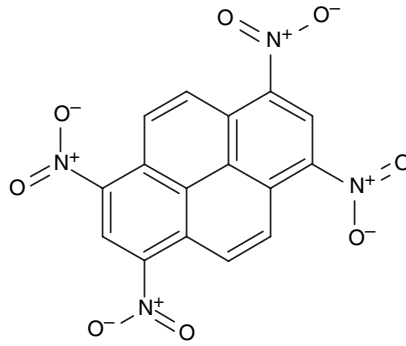
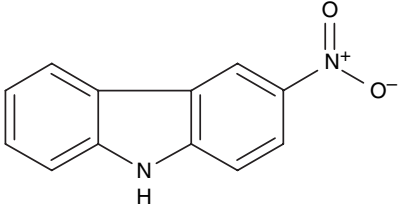
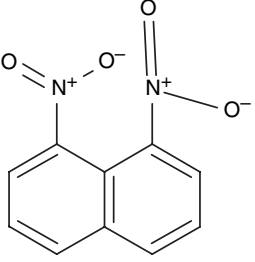
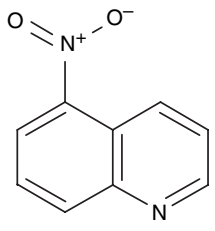
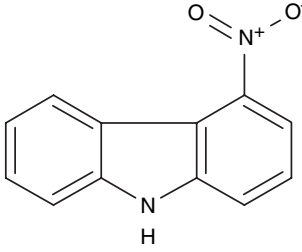
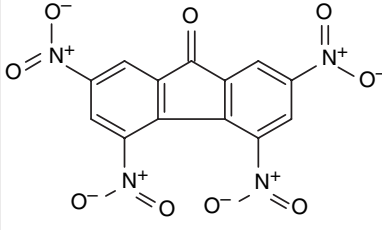
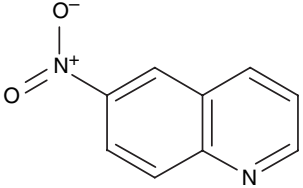
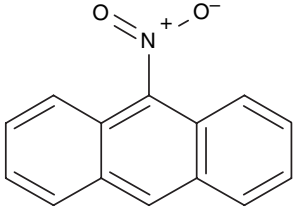
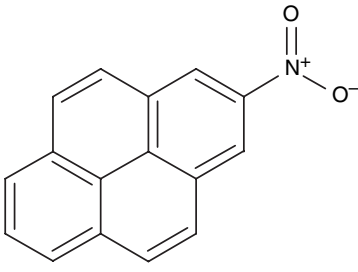
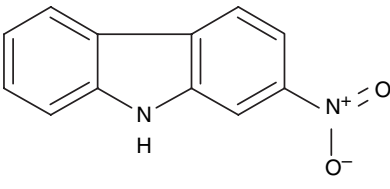
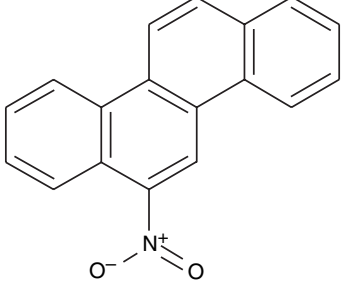
 <p>25</p>	 <p>29</p>	 <p>33</p>
 <p>26</p>	 <p>30</p>	 <p>34</p>
 <p>27</p>	 <p>31</p>	 <p>35</p>
 <p>28</p>	 <p>32</p>	 <p>36</p>

Table 1: (Continued)

 <p>37</p>	 <p>41</p>	 <p>45</p>
 <p>38</p>	 <p>42</p>	 <p>46</p>
 <p>39</p>	 <p>43</p>	 <p>47</p>
 <p>40</p>	 <p>44</p>	 <p>48</p>

**Table 2:** Numbers of compounds selected in the external test sets for splits A, B and C

Split A	Split B	Split C
1	1	3
2	9	4
3	11	7
10	13	9
12	15	11
14	17	25
16	19	27
20	21	29
22	23	33
24	25	43
26	27	
28	29	
30	31	
32	33	
34	35	
38		
40		
42		
44		
46		

to ASCII codes of the symbols. Therefore, every 'AB' composition can have only this version (not both 'AB' together with 'BA').

Simplified molecular input line entry system attributes which are rare in the training set can lead to overtraining. However, these may be blocked by the rule (7): if total number of the  $SA_k$  is less than  $LimN$  then  $CW(SA_k) = 1$ .

By the Monte Carlo optimization of active (i.e. not blocked)  $SA_k$  one can calculate  $CW(SA_k)$  producing as large as possible correlation coefficient between the  $\log TA_{100}$  and DCW calculated with eqn 1 for the training set. Having numerical data on the correlation weights, one can calculate the DCW for test set and estimate predictive potential of the model.

## Results and Discussion

Table 2 contains lists for three versions of the external test sets. Table 3 shows that models for the three splits into training and test sets are similar. For split A, best  $LimN$  is 9 (Figure 1) and for splits B and C best  $LimN$  is 10 (Figures 2 and 3). Again, square of correlation coefficients for test sets are similar. Table 4 shows that this statistical quality is satisfactorily reproduced in series of the Monte Carlo optimization.

As the described scheme of the QSAR analysis gives similar results for different splits, this approach is able to be a robust tool for the modeling.

It is also important that every SMILES attribute has a transparent interpretation, thus the model is convenient for mechanistic elucidation of the mutagenicity  $TA_{100}$  phenomenon.

**Table 3:** Statistical quality of the models for mutagenic potency with  $LimN$  from 0 to 10 (for splits A, B and C)

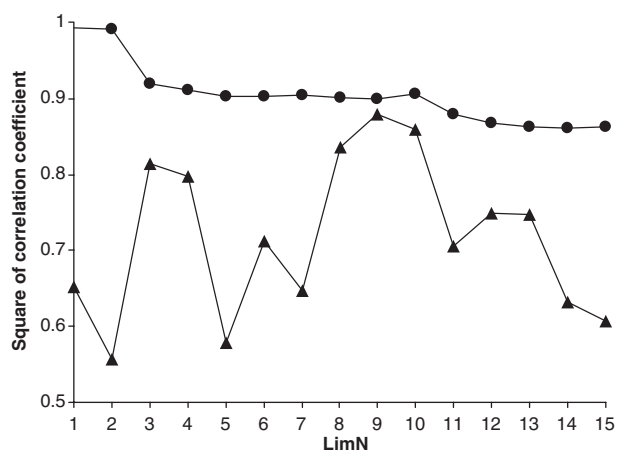
LimN	$N_{act}$	Training set				Test set			
		$n$	$r^2$	$s$	$F$	$n$	$r^2$	$s$	$F$
Split A									
1	58	28	0.9940	0.137	4396	20	0.6530	1.464	36
2	47	28	0.9915	0.163	3119	20	0.5574	1.665	28
3	40	28	0.9189	0.504	295	20	0.8152	0.853	84
4	35	28	0.9107	0.529	265	20	0.7978	0.788	72
5	32	28	0.9036	0.550	244	20	0.5789	1.168	25
6	31	28	0.9024	0.553	241	20	0.7126	0.947	49
7	26	28	0.9047	0.547	247	20	0.6466	1.063	35
8	25	28	0.9007	0.558	236	20	0.8361	0.721	93
<b>9</b>	<b>24</b>	<b>28</b>	<b>0.9003</b>	<b>0.559</b>	<b>235</b>	<b>20</b>	<b>0.8803</b>	<b>0.637</b>	<b>136</b>
10	22	28	0.9069	0.541	253	20	0.8595	0.697	110
11	20	28	0.8804	0.613	192	20	0.7060	0.958	45
12	20	28	0.8675	0.644	172	20	0.7490	0.872	58
13	19	28	0.8622	0.658	163	20	0.7478	0.875	54
14	17	28	0.8619	0.658	162	20	0.6317	1.092	31
15	17	28	0.8627	0.656	164	20	0.6065	1.150	28
Split B									
1	57	33	0.9603	0.348	749	15	0.7825	0.833	47
2	47	33	0.9603	0.347	751	15	0.6525	1.148	25
3	43	33	0.9600	0.349	744	15	0.8000	0.811	53
4	39	33	0.9484	0.396	570	15	0.8189	0.755	59
5	35	33	0.9358	0.442	452	15	0.8091	0.735	55
6	29	33	0.9183	0.498	348	15	0.8531	0.771	76
7	28	33	0.9141	0.511	330	15	0.8531	0.715	76
8	27	33	0.9107	0.521	317	15	0.8333	0.722	65
9	25	33	0.9010	0.549	282	15	0.8497	0.662	74
<b>10</b>	<b>23</b>	<b>33</b>	<b>0.9007</b>	<b>0.549</b>	<b>282</b>	<b>15</b>	<b>0.8714</b>	<b>0.628</b>	<b>90</b>
11	21	33	0.8572	0.659	186	15	0.7353	0.840	37
12	21	33	0.8579	0.657	187	15	0.7175	0.871	34
13	21	33	0.8576	0.658	187	15	0.7097	0.882	32
14	20	33	0.8576	0.658	187	15	0.7306	0.851	35
15	20	33	0.8561	0.662	184	15	0.7426	0.832	38
Split C									
1	58	38	0.9580	0.334	822	10	0.8074	0.950	34
2	46	38	0.9573	0.337	808	10	0.8216	0.937	37
3	43	38	0.9575	0.336	812	10	0.7937	1.015	33
4	40	38	0.9226	0.454	429	10	0.8876	0.754	65
5	37	38	0.9035	0.507	337	10	0.8011	0.992	37
6	35	38	0.9001	0.516	324	10	0.9429	0.656	134
7	35	38	0.9026	0.509	334	10	0.9172	0.740	100
8	28	38	0.8871	0.548	283	10	0.9144	0.763	93
9	26	38	0.8851	0.553	277	10	0.9420	0.722	131
<b>10</b>	<b>24</b>	<b>38</b>	<b>0.8830</b>	<b>0.558</b>	<b>273</b>	<b>10</b>	<b>0.9431</b>	<b>0.678</b>	<b>133</b>
11	22	38	0.8274	0.678	173	10	0.9215	0.706	95
12	21	38	0.8263	0.680	171	10	0.9114	0.723	83
13	21	38	0.8269	0.679	172	10	0.9115	0.726	84
14	21	38	0.8269	0.679	172	10	0.9122	0.730	85
15	21	38	0.8271	0.679	172	10	0.9138	0.727	85

The best  $LimN$  values are boldfaced.

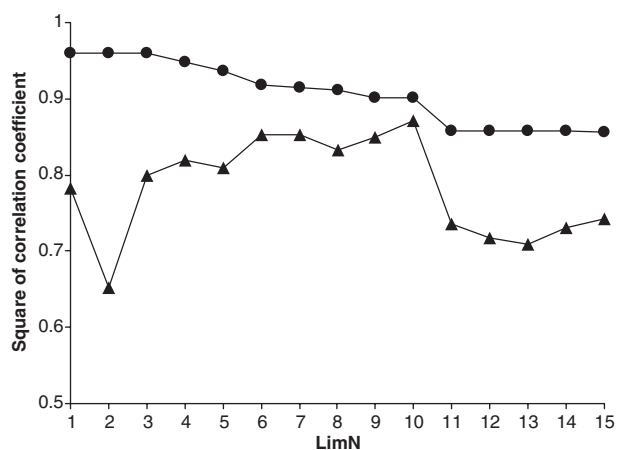
Model of the  $\log TA_{100}$  obtained in the first the Monte Carlo optimization probe with  $limN = 9$  is the following:

$$\log TA_{100} = -269.5393(\pm 1.969) + 269.2754(\pm 1.957) \times DCW \quad (2)$$

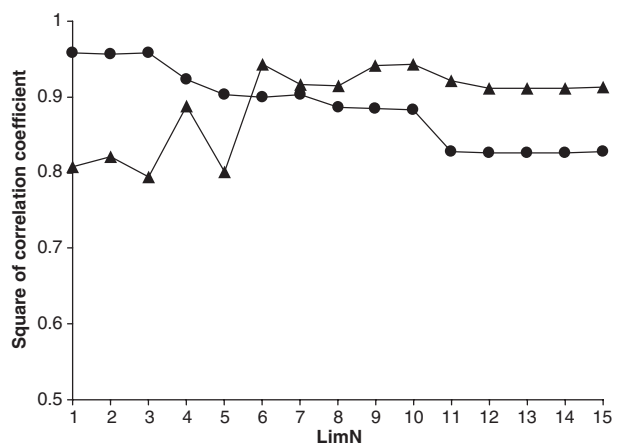
$$n = 28, r^2 = 0.902, q^2 = 0.892, s = 0.554, F = 240 \text{ (training set)}$$



**Figure 1:** Split A: squares of correlation coefficients against the LimN for training (circles) and test (triangles) sets.



**Figure 2:** Split B: squares of correlation coefficients against the LimN for training (circles) and test (triangles) sets.



**Figure 3:** Split C: squares of correlation coefficients against the LimN for training (circles) and test (triangles) sets.

**Table 4:** Statistics for the best models of the split A, B and C

LimN	$N_{act}$	Probe	Training set				Test set			
			$n$	$r^2$	$s$	$F$	$n$	$r^2$	$s$	$F$
Split A										
9	24	1	28	0.9021	0.554	240	20	0.8533	0.702	105
		2		0.8994	0.562	232		0.8884	0.614	143
		3		0.8993	0.562	232		0.8992	0.596	161
		Average		0.9003	0.559	235		0.8803	0.637	136
Split B										
10	23	1	33	0.9013	0.548	283	15	0.8555	0.652	77
		2		0.9039	0.541	291		0.8634	0.645	82
		3		0.8970	0.560	270		0.8953	0.587	111
		Average		0.9007	0.549	282		0.8714	0.628	90
Split C										
10	24	1	38	0.8860	0.551	280	10	0.9436	0.681	134
		2		0.8894	0.543	289		0.9406	0.709	127
		3		0.8737	0.580	249		0.9452	0.643	138
		Average		0.8830	0.558	273		0.9431	0.678	133

$$n = 20, r^2 = 0.853, q^2 = 0.823, s = 0.702, F = 105 \text{ (test set)}$$

Table 5 shows experimental and calculated with eqn 2 logTA100 values. Table 6 contains correlation weights for the DCW calculation. Table 7 shows an example of the DCW calculation. Figures 4 and 5 are demonstrations of the model for training and test sets, respectively (split A). Details for models obtained for splits B and C are represented in the Supporting information.

Various versions of the optimal descriptors (6–14) have also been examined in QSPR/QSAR analyses. Simplified molecular input line entry system-based optimal descriptors have been used for: octanol/water partition coefficient (10), binding affinity (11), anti-HIV-1 activity (12) and water solubility of minerals (13,14). Thus, the described one-variable models of the mutagenic potential cannot be the chance correlations.

Statistical characteristics of the log TA100 model described in Ref. (5) are  $n = 48$ ,  $r^2 = 0.9157$ ,  $s = 0.551$ ,  $F = 41$ . More typical statistical characteristics of the log TA100 models are  $n = 41$ ,  $r^2 = 0.794$  ( $r = 0.891$ ) (2);  $n = 67$ ,  $r^2 = 0.769$  ( $r = 0.877$ ),  $s = 0.708$  (3); and  $n = 42$ ,  $r^2 = 0.901$  (8). Thus, the mutagenicity model calculated with eqn 2 is reasonably good.

Unfortunately, an external test set was not used in Ref. 5. Under such circumstances, an adequate comparison of one-variable models which are calculated with eqn 2 with the 10-variable model from Ref. 5 becomes impossible. However, a positive feature of our model is that it has been validated with an external test set, and this has been repeated thrice with different splits. This proves that our model is predictive. This can be used for the evaluation of compounds not used in the model building. Vice versa in the case of the model from Ref. 5, the predictability is not examined with an external data.

It is well known that the increase of the number of descriptors improves statistical quality of a model that is obtained with the multiple linear regression analysis (MLRA) for the training set (15).

**Table 5:** Mutagenic potency, experimental and calculated with eqn. 2 for the model (split A, LimN = 9)

No.	SMILES	DCW	Experimental	Calculated	Experimental – Calculated
Training set					
4	[O-][N+](=O)c1ccc3ccc4c2c(ccc1c23)ccc4[N+](O-)=O	1.0154622	4.090	3.900	0.190
5	[O-][N+](=O)c4ccc1ccc2ccc([N+](O-)=O)c3ccc4c1c23	1.0187679	4.740	4.790	-0.050
6	[O-][N+](=O)c2ccc3c1ccc(cc1C(=O)c3c2)[N+](O-)=O	1.0113039	2.690	2.780	-0.090
7	[O-][N+](=O)c2cc4cccc3c1cccc1c(c2)c34	1.0111682	3.000	2.743	0.257
8	[O-][N+](=O)c1ccc2cc3cccc3cc2c1	1.0066159	3.050	1.518	1.532
9	[O-][N+](=O)c1ccc2c3ccc(cc3C2c1)[N+](O-)=O	1.0055073	1.270	1.219	0.051
11	[O-][N+](=O)c3ccc4c2cccc1cccc(c12)c4c3	1.0088536	2.600	2.120	0.480
13	[O-][N+](=O)c3ccc4c2cccc1cccc(c12)c4c3	1.0088536	2.090	2.120	-0.030
15	[O-][N+](=O)c1ccc2c3cccc3C2c1	1.0026297	1.080	0.444	0.636
17	[O-][N+](=O)c2cccc1cccc12	0.9993928	0.280	-0.427	0.707
18	[O-][N+](=O)c1ccc2C=Cc3cccc1c23	1.0048653	0.970	1.046	-0.076
19	O=[N+](O-)[c1ccc(c1)[N+](O-)=O	0.9997135	-0.510	-0.341	-0.169
21	Cc1ccc(cc1)[N+](=O)[O-]	0.9926079	-2.100	-2.254	0.154
23	Cc1ccc(cc1[N+](=O)[O-])[N+](O-)=O	0.9961117	-1.290	-1.311	0.021
25	Cc1c(cccc1[N+](=O)[O-])[N+](O-)=O	0.9961117	-1.340	-1.311	-0.029
27	O=[N+](O-)[c1cc(C)cc(c1)[N+](O-)=O	0.9997908	-0.720	-0.320	-0.400
29	O=[N+](O-)[c1cc(cc(C)c1)[N+](O-)=O][N+](O-)=O	1.0036367	0.460	0.715	-0.255
31	O=[N+](O-)[c1cc(c(C)cc1[N+](O-)=O)[N+](O-)=O	1.0036367	1.120	0.715	0.405
33	O=[N+](O-)[c1cc(C)cc1[N+](O-)=O][N+](O-)=O	1.0036367	1.010	0.715	0.295
35	[O-][N+](=O)c1cc2cccc2cc1C	1.0009124	-0.700	-0.018	-0.682
36	[O-][N+](=O)c2cc(cc1cccc12)[N+](O-)=O	1.0022610	0.860	0.345	0.515
37	[O-][N+](=O)c2cccc1c2cccc1[N+](O-)=O	1.0046195	0.910	0.980	-0.070
39	[O-][N+](=O)c3cc2c(c1c(cc(cc1C2 = O)[N+](O-)=O)[N+](O-)=O)c(c3)[N+](O-)=O	1.0098337	2.460	2.384	0.076
41	[O-][N+](=O)c4cc(c1ccc2c(cc([N+](O-)=O)c3ccc4c1c23)[N+](O-)=O)[N+](O-)=O	1.0140685	3.180	3.524	-0.344
43	[O-][N+](=O)c1ccc2ncccc2c1	0.9976078	-1.050	-0.908	-0.142
45	[O-][N+](=O)c1cc2c3cccc3nc2cc1	1.0026297	-1.000	0.444	-1.444
47	[O-][N+](=O)c2c3cccc3cc1cccc12	1.0050876	0.260	1.106	-0.846
48	[O-][N+](=O)c4cc2c(ccc1cccc12)c3cccc34	1.0117534	2.210	2.901	-0.691
Test set					
1	[O-][N+](=O)c1ccc3ccc4c2c(ccc1c23)c(cc4[N+](O-)=O)[N+](O-)=O	1.0185217	3.870	4.724	-0.854
2	[O-][N+](=O)c1cc2Cc3cc(cc(c3c2cc1)[N+](O-)=O)[N+](O-)=O	1.0074911	2.270	1.753	0.517
3	[O-][N+](=O)c1cc([N+](O-)=O)c4ccc3cccc2ccc1c4c23	1.0187679	4.630	4.790	-0.160
10	[O-][N+](=O)c4ccc2c1cccc1c3cccc4c23	1.0129803	3.310	3.231	0.079
12	[O-][N+](=O)c4ccc1ccc2cccc3ccc4c1c23	1.0129803	2.170	3.231	-1.061
14	[O-][N+](=O)c1ccc2cccc2c1	1.0009124	0.370	-0.018	0.388
16	[O-][N+](=O)c1cc2ccc3cccc3c2cc1	1.0066159	1.790	1.518	0.272
20	O=[N+](O-)[c1cc(cc(c1)[N+](O-)=O)[N+](O-)=O	1.0027256	0.720	0.470	0.250
22	O=[N+](O-)[c1ccc(C)c1[N+](O-)=O	1.0006220	-1.260	-0.096	-1.164
24	Cc1cc(ccc1[N+](=O)[O-])[N+](O-)=O	0.9961117	-0.630	-1.311	0.681
26	O=[N+](O-)[c1cc(C)ccc1[N+](O-)=O	1.0006220	-1.300	-0.096	-1.204
28	O=[N+](O-)[c1c(c(C)ccc1[N+](O-)=O)[N+](O-)=O	1.0036367	0.080	0.715	-0.635
30	O=[N+](O-)[c1ccc(c(C)c1[N+](O-)=O)[N+](O-)=O	1.0036367	0.550	0.715	-0.165
32	Cc1c(cc(cc1[N+](=O)[O-])[N+](O-)=O)[N+](O-)=O	0.9991129	0.160	-0.503	0.663
34	[O-][N+](=O)c2ccc1cccc1c2C	1.0022129	0.080	0.332	-0.252
38	[O-][N+](=O)c1cccc2cccc([N+](O-)=O)c12	1.0051028	1.120	1.110	0.010
40	[O-][N+](=O)c1cc2ccc3cccc4ccc(c1)c2c34	1.0098561	2.870	2.390	0.480
42	[O-][N+](=O)c1cccc2ncccc12	0.9960932	-0.700	-1.316	0.616
44	[O-][N+](=O)c1ccc2c3cccc3nc2c1	1.0026297	-0.300	0.444	-0.744
46	[O-][N+](=O)c2cccc3nc1cccc1c23	1.0044901	-0.300	0.945	-1.245

However, it can be accompanied by decrease of the statistical quality of this model for an external test set (16–18). For the toxicity towards rats (16,17) and the inhibition of the protein of cholesterol ester transfer (18), the robust MLRA predictions are the three-variable models or even two-variable models. It should be noted that in above-mentioned studies, one-variable models which are calculated with optimal descriptors (either based on molecular graph or based on SMILES) are better than the MLRA models.

The described computational experiments have shown that SMILES attributes ( $ss_k$ , which contain two elements) can be a robust basis for the predictive model of the mutagenicity, log TA100. There are promoters of increase in the log TA100 value, as well as there are promoters of decrease in the log TA100 value. The promoters of the increase have correlation weights greater than 1 in all three probes of the Monte Carlo optimization. Vice versa the promoters of the decrease in log TA100 value have correlation weights smaller



**Table 6:** Correlation weights for DCW calculation obtained in three probes of the Monte Carlo optimization (split A, LimN = 9)

SMILES attributes, (SA)	CW(SA) in probe 1	CW(CA) in probe 3	CW(CA) in probe 3
<b>b<sup>a</sup></b>			
(001	1.0	1.0	1.0
(002	1.0	1.0	1.0
(003	1.0	1.0	1.0
(004	1.0	1.0	1.0
(005	1.0	1.0	1.0
(007	1.0	1.0	1.0
(008	1.0	1.0	1.0
<b>db</b>			
=001	0.9928542	0.9874177	0.9879436
=002	0.9976385	0.9958808	0.9978417
=003	1.0	1.0	1.0
=004	1.0	1.0	1.0
=005	1.0	1.0	1.0
<b>Nn</b>			
N001	0.9937867	0.9826527	0.9867685
N002	0.9962738	0.9983171	0.9896293
N003	1.0	1.0	1.0
N004	1.0	1.0	1.0
<b>No</b>			
O002	0.9951094	0.9929315	0.9841075
O004	0.9982900	0.9903537	0.9940264
O005	1.0	1.0	1.0
O006	1.0	1.0	1.0
O008	1.0	1.0	1.0
O009	1.0	1.0	1.0
<b>SS<sub>k</sub></b>			
1 (	1.0	1.0	1.0
2 (	1.0	1.0	1.0
2 1	1.0	1.0	1.0
3 (	1.0	1.0	1.0
3 2	1.0	1.0	1.0
4 3	1.0	1.0	1.0
= (	0.9940970	0.9893757	0.9917679
= 2	1.0	1.0	1.0
C (	1.0003676	1.0001318	1.0008531
C 1	1.0	1.0	1.0
C 2	1.0	1.0	1.0
C 3	1.0	1.0	1.0
C =	1.0	1.0	1.0
N +	1.0011843	1.0006022	0.9986831
O (	1.0028353	1.0071538	1.0055170
O -	1.0027073	0.9959818	0.9956816
O =	1.0049233	0.9978888	1.0028051
[ (	0.9998268	1.0000167	1.0012513
[ +	0.9994688	0.9980135	0.9929854
[ -	0.9984738	0.9977314	1.0018611
[ 1	1.0	1.0	1.0
[ 4	1.0	1.0	1.0
[ =	1.0	1.0	1.0
[ N	0.9985995	1.0019456	0.9987970
[ O	0.9943171	1.0026117	0.9966329
[ [	0.9969800	0.9975309	0.9976963
c (	1.0005204	1.0017735	1.0016327
c 1	1.0029115	1.0045423	1.0048511
c 2	1.0016101	1.0024841	1.0030810
c 3	1.0005716	1.0011042	1.0010069
c 4	1.0011128	1.0012849	1.0015388
c C	1.0	1.0	1.0

**Table 6:** (Continued)

SMILES attributes, (SA)	CW(SA) in probe 1	CW(CA) in probe 3	CW(CA) in probe 3
c c	1.0016997	1.0027042	1.0031273
n 2	1.0	1.0	1.0
n 3	1.0	1.0	1.0
n c	1.0	1.0	1.0

<sup>a</sup>As all CW(b) = 1, these SMILES attributes have no influence for this model; however, for other splits CW(b) these attributes have CW(b) ≠ 1.

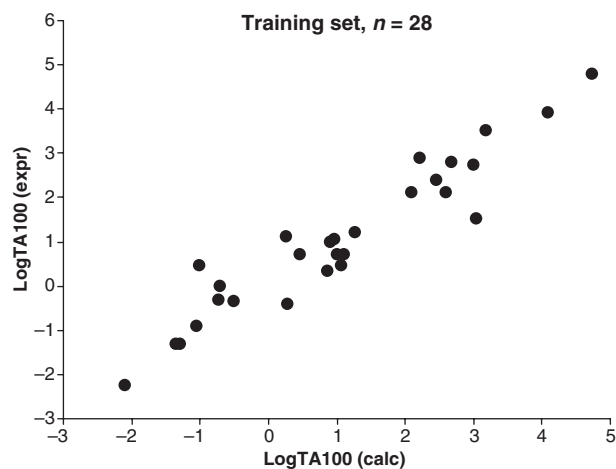
**Table 7:** Example of the DCW calculation (split A, LimN = 9, probe 1) SMILES = '[O-][N+](=O)c1ccc3ccc4c2c(ccc1c23)c(cc4[N+](O-)=O)[N+](O-)=O' No. 1; DCW = 1.0185217

SA	CW(SA) in probe 1	N <sub>TRN</sub>	N <sub>ST</sub>
[ O	0.9943171	50	37
O -	1.0027073	50	37
[ -	0.9984738	50	37
[ [	0.9969800	20	13
[ N	0.9985995	50	37
N +	1.0011843	50	37
[ +	0.9994688	50	37
[ (	0.9998268	127	97
= (	0.9940970	46	32
O =	1.0049233	52	37
O (	1.0028353	33	23
c (	1.0005204	88	56
c 1	1.0029115	89	66
c 1	1.0029115	89	66
c c	1.0016997	154	107
c c	1.0016997	154	107
c 3	1.0005716	44	23
c 3	1.0005716	44	23
c c	1.0016997	154	107
c c	1.0016997	154	107
c 4	1.0011128	23	17
c 4	1.0011128	23	17
c 2	1.0016101	58	40
c 2	1.0016101	58	40
c (	1.0005204	88	56
c (	1.0005204	88	56
c c	1.0016997	154	107
c c	1.0016997	154	107
c 1	1.0029115	89	66
c 1	1.0029115	89	66
c 2	1.0016101	58	40
3 2	1.0	4	5
3 (	1.0	3	1
c (	1.0005204	88	56
c (	1.0005204	88	56
c c	1.0016997	154	107
c 4	1.0011128	23	17
[ 4	1.0	1	1
[ N	0.9985995	50	37
N +	1.0011843	50	37
[ +	0.9994688	50	37
[ (	0.9998268	127	97
[ (	0.9998268	127	97

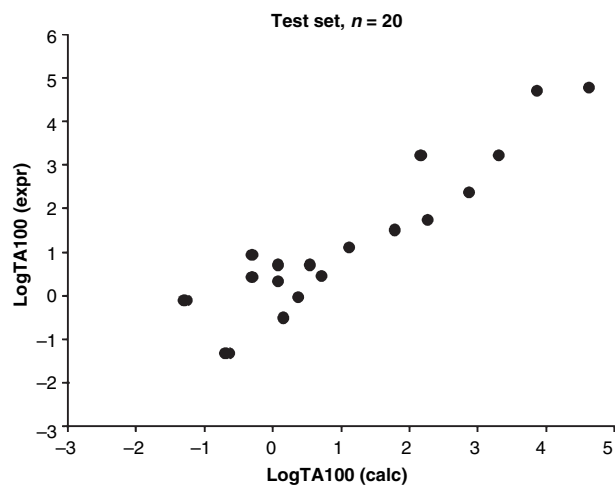
**Table 7:** (Continued)

SA	CW(SA) in probe 1	$N_{TRN}$	$N_{TST}$
[ _O _ _ _ _ _ ]	0.9943171	50	37
O _ _ _ _ _ _ ]	1.0027073	50	37
[ _ _ _ _ _ _ ]	0.9984738	50	37
[ _ _ _ ( _ _ _ _ ]	0.9998268	127	97
= _ _ _ ( _ _ _ _ ]	0.9940970	46	32
O _ _ = _ _ _ _ ]	1.0049233	52	37
O _ _ ( _ _ _ _ ]	1.0028353	33	23
[ _ _ _ ( _ _ _ _ ]	0.9998268	127	97
[ _ _ _ N _ _ _ _ ]	0.9985995	50	37
N _ _ + _ _ _ _ ]	1.0011843	50	37
[ _ _ _ + _ _ _ _ ]	0.9994688	50	37
[ _ _ _ ( _ _ _ _ ]	0.9998268	127	97
[ _ _ _ ( _ _ _ _ ]	0.9998268	127	97
[ _ _ _ O _ _ _ _ ]	0.9943171	50	37
O _ _ _ _ _ _ ]	1.0027073	50	37
[ _ _ _ _ _ _ ]	0.9984738	50	37
[ _ _ _ ( _ _ _ _ ]	0.9998268	127	97
= _ _ _ ( _ _ _ _ ]	0.9940970	46	32
O _ _ = _ _ _ _ ]	1.0049233	52	37
(O05 _ _ _ _ _ ]	1.0	3	6
=O03 _ _ _ _ _ ]	1.0	4	6
N003 _ _ _ _ _ ]	1.0	3	6
O006 _ _ _ _ _ ]	1.0	3	6

$N_{TRN}$  and  $N_{TST}$  are numbers of the SA in the training and the test sets, respectively.

**Figure 4:** Mutagenic potency, experimental and calculated with eqn 2 log TA100 for training set (split A, LimN = 9, probe 1).

than 1 in all three attempts of the Monte Carlo optimization. Finally, there are SMILES attributes with unclear role: they have (in the three runs of the Monte Carlo optimization) values correlation weights greater than 1 and values of correlation weights smaller than 1 (Table 6). Global SMILES attributes, such as, the number of double bonds (=001, =002,...), the number of oxygen atoms (O002, O004,...), the number of nitrogen atoms (N002, N004,...), have improved the predictability of the model, being promoters of the decrease in the log TA100 value (Table 6). All these fragments (or

**Figure 5:** Mutagenic potency, experimental and calculated with eqn 2 log TA100 for test set (split A, LimN = 9, probe 1).

attributes) can be useful in the understanding which molecular features are related to the mutagenic process.

## Conclusions

Optimal SMILES-based descriptors can be used to predict mutagenic potency (TA100) of nitrated polycyclic hydrocarbons. The blocking of rare SMILES attributes, by means of selection of the LimN, is able to improve predictive potential of the model, i.e. improve the statistical characteristics for the external test set. However, the LimN should be selected properly: zero value of the LimN can lead to overtraining, i.e. an excellent model for training set, but poor model for the test set (Table 3); and vice versa if the LimN is too large then the model can become poor for both the training and test sets.

## Acknowledgment

The authors express gratitude for the Marie Curie Fellowship financial support (the contract ID 39036, CHEMPREDICT).

## References

- Glende C., Schmitt H., Erdinger L., Engelhardt G., Boche G. (2001) Transformation of mutagenic aromatic amines into non-mutagenic species by alkyl substituents Part I. Alkylation ortho to the amino function. *Mutat Res*;498:19–37.
- Andrews L.E., Bonin A.M., Fransson L.E., Gillson A.-M.E., Glover S.A. (2006) The role of steric effects in the direct mutagenicity of *N*-acyloxy-*N*-alkoxyamides. *Mutat Res*;605:51–62.
- Okamoto A.K., Gaudio A.C., Marques A.S., Takahata Y. (2005) QSAR study of inhibition by coumarins of IQ induced mutation in *S. typhimurium* TA98. *J Mol Struct (Theochem)*;725:231–238.
- Gramatica P., Papa E., Marrocchi A., Minuti L., Taticchi A. (2007) Quantitative structure–activity relationship modeling of polycyclic

- aromatic hydrocarbon mutagenicity by classification methods based on holistic theoretical molecular descriptors. *Ecotoxicol Environ Saf*;66:353–361.
5. Singh J., Singh S., Shaik B., Deeb O., Sohani N., Agrawal V.K., Khadikar P.V. (2008) Mutagenicity of nitrated polycyclic aromatic hydrocarbons: a QSAR investigation. *Chem Biol Drug Des*;71:230–243.
  6. Toropov A.A., Schultz T.W. (2003) Prediction of aquatic toxicity: use of the optimization of correlation weights of local graph invariants. *J Chem Inf Comput Sci*;43:560–567.
  7. Toropov A.A., Benfenati E. (2007) Optimisation of correlation weights of SMILES invariants for modelling oral quail toxicity. *Eur J Med Chem*;42:606–613.
  8. Gonzalez M.P., Moldes M.C.T., Fall Y., Dias L.C., Helguera A.M. (2005) A topological sub-structural approach to the mutagenic activity in dental monomers. 2. Cycloaliphatic epoxides. *Polymer*;46:2783–2790.
  9. Duchowicz P.R., Castro E.A., Toropov A.A., Benfenati E. (2006) Applications of flexible molecular descriptors in the QSPR–QSPR study of heterocyclic drugs. *Top Heterocycl Chem*;3:1–38.
  10. Raska I., Toropov A., Gutman I., Završnik D., Spirtovic S. (2006) QSPR modeling of octanol/water partition coefficient of vitamins in the book. In: Gutman I., editor. *Mathematical Methods in Chemistry*, Prijepolje: Prijepolje Museum; p. 197–206.
  11. Roy K., Toropov A., Raska I. Jr (2007) QSAR modeling of peripheral versus central benzodiazepine receptor binding affinity of 2-phenylimidazo[1,2-a]pyridineacetamides using optimal descriptors calculated with SMILE. *QSAR Comb Sci*;26:460–468.
  12. Castro E.A., Torrens F., Toropov A.A., Nesterov I.V., Nabiev O.M. (2004) QSAR modeling anti-HIV-1 activities by optimization of correlation weights of local graph invariants. *Mol Simul*;30:691–696.
  13. Toropova A.P., Toropov A.A., Maksudov S.Kh. (2006) QSPR modeling mineral crystal lattice energy by optimal descriptors of the graph of atomic orbitals. *Chem Phys Lett*;428:183–186.
  14. Toropova A.P., Toropov A.A., Gutman I. (2008) QSPR modelling of water solubility of minerals by optimal descriptors calculated with SMILES. *Kragujevac J Sci*;30:65–72.
  15. Topliss J.G., Edwards R.P. (1979) Chance factors in studies of quantitative structure-activity relationships. *J Med Chem*;22:1238–1244.
  16. Toropov A.A., Rasulev B.F., 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.
  17. Toropov A.A., Rasulev B.F., Leszczynski J. (2008) QSAR modeling of acute toxicity by balance of correlations. *Bioorg Med Chem*;16:5999–6008.
  18. Rasulev B.F., Toropov A.A., Hamme A.T., II, Leszczynski J. (2008) Multiple linear regression analysis and optimal descriptors: predicting the cholesteryl ester transfer protein inhibition activity. *QSAR Comb Sci*;27:595–606.

## Note

<sup>a</sup>ACD/ChemSketch Freeware, version 11.00, 2007, Toronto, ON, Canada: Advanced Chemistry Development Inc. Available at: [www.acdlabs.com](http://www.acdlabs.com).

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Split B.** Correlation weights.

**Split B.** Model for first probe with  $\text{LimN} = 10$ .

**Split C.** Correlation weights.

**Split C.** Model for first probe with  $\text{LimN} = 10$ .

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.