

Article

Additive SMILES-Based Carcinogenicity Models: Probabilistic Principles in the Search for Robust Predictions

Andrey A. Toropov^{1,2,*}, Alla P. Toropova^{1,2} and Emilio Benfenati²

¹ Institute of Geology and Geophysics, 100041, Khodzhibaev St. 49, Tashkent, Uzbekistan;
E-Mail: altoropova@mail.ru (A.P.T.)

² Istituto di Ricerche Farmacologiche Mario Negri, 20156, Via La Masa 19, Milano, Italy;
E-Mail: benfenati@marionegri.it (E.B.)

* Author to whom correspondence should be addressed; E-Mail: atoropov@yahoo.com (A.A.T.);
Tel. +390239014595; Fax: +390239014735

Received: 14 May 2009; in revised form: 23 June 2009 / Accepted: 2 July 2009 /

Published: 8 July 2009

Abstract: Optimal descriptors calculated with the simplified molecular input line entry system (SMILES) have been utilized in modeling of carcinogenicity as continuous values ($\log TD_{50}$). These descriptors can be calculated using correlation weights of SMILES attributes calculated by the Monte Carlo method. A considerable subset of these attributes includes rare attributes. The use of these rare attributes can lead to overtraining. One can avoid the influence of the rare attributes if their correlation weights are fixed to zero. A function, limS , has been defined to identify rare attributes. The limS defines the minimum number of occurrences in the set of structures of the training (subtraining) set, to accept attributes as usable. If an attribute is present less than limS , it is considered “rare”, and thus not used. Two systems of building up models were examined: 1. classic training-test system; 2. balance of correlations for the subtraining and calibration sets (together, they are the original training set: the function of the calibration set is imitation of a preliminary test set). Three random splits into subtraining, calibration, and test sets were analysed. Comparison of abovementioned systems has shown that balance of correlations gives more robust prediction of the carcinogenicity for all three splits (split 1: $r_{\text{test}}^2=0.7514$, $s_{\text{test}}=0.684$; split 2: $r_{\text{test}}^2=0.7998$, $s_{\text{test}}=0.600$; split 3: $r_{\text{test}}^2=0.7192$, $s_{\text{test}}=0.728$).

Keywords: QSAR; SMILES; optimal descriptor; carcinogenicity; balance of correlations; applicability domain

1. Introduction

Carcinogenicity is an important endpoint from a toxicological point of view and quantitative structure – activity relationships (QSAR) are a tool for modeling this endpoint [1-3]. Usually, the QSAR analysis is based on molecular descriptors, calculated from molecular graphs [3,4]. However, the simplified molecular input line entry system (SMILES) [5-7] has become a prospective alternative to molecular graphs in QSAR analysis [8-11], owing to an expansion of the databases available via the Internet with molecular structures given in SMILES notation [15,16]. The present study aimed to estimate the ability of the SMILES-based optimal descriptors to be a tool for QSAR analysis of carcinogenicity of non-congeneric chemicals.

2. Materials and Methods

Carcinogenicity data: Experimental values for carcinogenicity were taken from publicly available data sources and further checked for chemical structures [17]. Carcinogenicity is expressed as the potency dose that induces cancer in rats (TD₅₀, in mg/kg body weight). These values have been converted into mmol/kg body weight. The -log(TD₅₀) was examined as endpoint for the modelling. Initially, 401 chemicals have been extracted from [17]. These compounds were selected as substances with numerical data on the carcinogenicity available from [17].

However, this set (401 compounds) contains eight outliers (Table 1): for these compounds the difference between experimental and calculated (by our approach) value of -logTD₅₀ is more than the double the standard error (2s). Probably the high symmetry and the presence of the *N*-nitroso group can lead to the unusual behaviour of these substances. These compounds were removed. Thus, 393 compounds were examined in this study. SMILES notations which were used in this study have been taken from [18].

We randomly split these 393 chemicals three times into training (n=165), calibration (n=167) and test (n=61) sets. The range of -log(TD₅₀) values for these sets is about from -2 to 5 logarithmic units. Below, these splits are denoted the Split1, Split2, and Split3 (The *Supplementary Materials* contain lists of these splits).

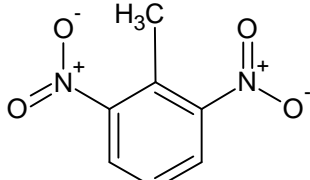
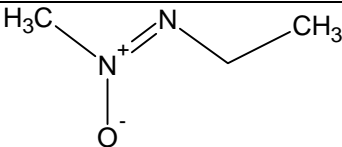
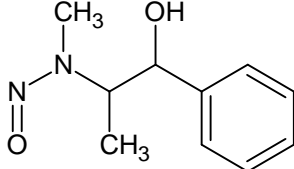
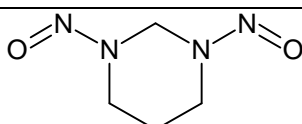
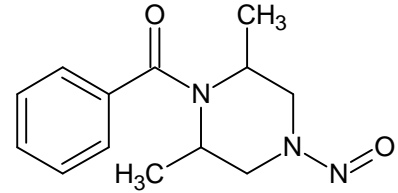
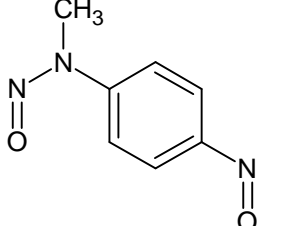
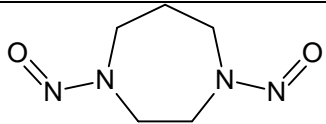
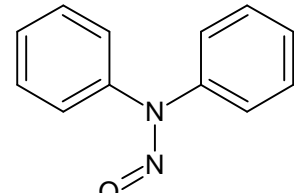
The modification of the descriptor that was used for modeling bee toxicity [10] is the tool for QSAR analysis of the carcinogenicity. This descriptor is calculated as follows:

$$DCW(\text{limS}) = CW(\text{dC}) + \sum CW(^1\text{SA}_k) + \sum CW(^2\text{SA}_k) + \sum CW(^3\text{SA}_k) \quad (1)$$

where ¹SA_k, ²SA_k, ³SA_k are SMILES attributes. ¹SA_k, ²SA_k, and ³SA contain one, two, and three SMILES elements, respectively. The SMILES element can be one (e.g., 'C', 'c', 'N', 'S', etc.), two (e.g., 'Cl', 'Br', etc.), three ('C=O'), and four symbols ('[O-]'). The order of elements in depiction of the ²SA_k or ³SA_k is defined by the ASCII characters. In other words only one version of AB-sequence

or ABC-sequence is possible in the list of the SMILES-attributes (not AB together with BA, or ABC together with CBA).

Table 1. The list of outliers of the QSAR models calculated with SMILES-based optimal descriptors.

Number	Structure	CAS	Chemical name
1		606-20-2	2,6-Dinitrotoluene
2		57497-34-4	Z-Methyl-O,N,N-azoxyethane
3		17608-59-2	N-Nitrosoephedrine
4		15973-99-6	Di(N-nitroso)-perhydropyrimidine
5		61034-40-0	1-Nitroso-4-benzoyl-3,5-dimethylpiperazine
6		99-80-9	N,4-Dinitrosomethylaniline
7		55557-00-1	N,N-Dinitrosohomopiperazine
8		86-30-6	N-Nitrosodiphenylamine

The dC is the difference of the number of 'C' (capital letter) in the given SMILES notation minus the number of 'c' (lowercase letter) in the given SMILES notation. For example, this global SMILES attribute is denoted as '!001', if $dC = N('c') - N('C') = 1$, and as '!-02' if the $dC = -2$. The $CW(dC)$ is the correlation weight of the dC. The symbol "C" (capital letter) is the representation of a carbon atom in the sp^3 configuration. The symbol "c" (lowercase letter) is the representation of a carbon atom in sp^2 configuration. Thus, the dC is a measure of presence of rigid and flexible fragments in molecular architecture. The examined substances contain chlorine that gives an additional 'C'. The chlorine is not rigid fragment in molecular system and we have calculated the dC taking into account the 'C' from chlorine atoms. Table 2 contains an example of the representation of SMILES by the set of SMILES attributes.

Table 2. Example of definition of SMILES attributes (unused positions are indicated by dots).

1S_k	$CW(^1S_k)$	2S_k	$CW(^2S_k)$	dC	$CW(dC)$
C.....	-0.0156855				
O=C.....	-2.8475657	O=C.C.....	0.0	!-02.....	1.2190257

SMILES="CC=O"; CAS= 75-07-0; DCW= -1.6442255.

The $CW(dC)$, $CW(^1SA_k)$, $CW(^2SA_k)$, and $CW(^3SA_k)$ are correlation weights of the above SMILES attributes. By means of the Monte Carlo method one can calculate numerical data for these weights which give maximal value of determination coefficient (square of the correlation coefficient, r^2) for the training set. However, most probably overtraining will result, i.e., an excellent model on the training set will be accompanied by a poor model for the test set. In order to avoid overtraining one can use the correlation balance [11], i.e., split the available chemicals into three sets: subtraining, calibration, and external test set. This approach gave reasonable result for the case of toxicity of 61 compounds [11], however for carcinogenicity of 393 compounds it is not enough. The use of the correlation balance and blocking of rare SMILES attributes [10] can improve the model. The blocking of rare attributes can be done by the scheme: if the number of SMILES from the training (subtraining) set which contain the SMILES attribute SA^* is less than the $limS$, the correlation weight of the SA^* should be fixed equal to zero, $CW(SA^*)=0$.

Without rare attributes the model becomes better for the external test set. However, if $limS$ is too large, the predictive potential of the model decreases, because the low number of active SMILES attribute cannot provide a high quality model. Thus, the central point of the system of modeling is the selection of the most efficient $limS$. The general scheme of the construction of optimal SMILES-based descriptors by the correlation balance method is represented in Figure 1.

This system can be denoted as a **[Subtraining-Calibration-Test] system**. The model can be satisfactory if the N_{111} , i.e., the number of active (not blocked) attributes which are present in subtraining, calibration, and test sets, is as large as possible. The more traditional, "classic" approach is the construction of the model using united training set to predict the endpoint for an external test set. This system can be denoted as **[Training-Test] system**. This model can be satisfactory if the N_{101} , i.e., the number of active attributes which are present in both the training and test set is as large as possible.

The correlation weights were calculated by the Monte Carlo method Optimization. The **[Training-Test] system** is based on correlation weights which provide maximum of the correlation coefficient

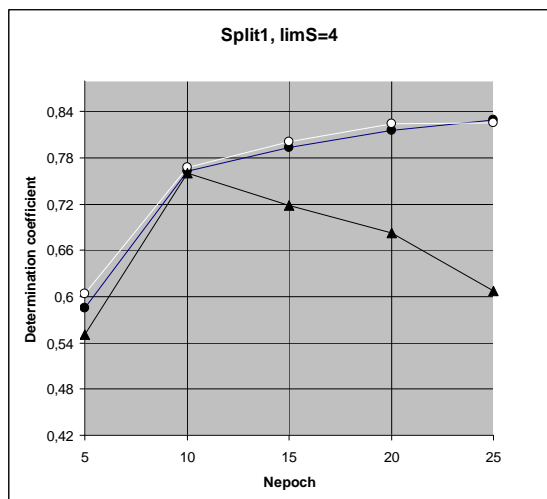
For each attribute SA, $CW(SA)$ is determined initially by setting the start values of all CW s to $1 \pm 0.01 \cdot \text{random}$. The random is the generator of random value of range (0, 1). The regular order of number of attributes (i.e., 1, 2, 3, 4, 5,...) is replaced by a random sequence (e.g., 3, 1, 5, 2, 4,...). A starting value of target function (TF1) is calculated. In a generated random sequence, each attribute correlation weight CW_i was modified with the algorithm:

1. $DCW_i = 0.5 \cdot CW_i$; $Eps = 0.1 \cdot DCW_i$;
2. Calculation of TF1; $CW_i = CW_i + DCW_i$;
3. Calculation of TF2, after modify CW_i ;
4. If $TF2 > TF1$ then $TF1 = TF2$; go to 2
5. $CW_i = CW_i - DCW_i$;
6. $DCW_i = -0.5 \cdot DCW_i$;
7. If absolute value (DCW_i) $> Eps$ then go to 2.

Then, steps of 1–7 are carried out for all CW s (the epoch of the optimization). By computational experiment the optimal number of the epochs has been established (Table 3). This number is 10 (Figure 2).

Figure 2. Results of computational experiments, which were used to establish of the preferable number of epochs of the Monte Carlo optimization (Nepoch). Triangles indicate curves for the test sets. Black circles denote the sub training set. White circles denote the calibration set.

[Subtraining (●) - Calibration (○) - Test (▲)] system



[Training (●) - Test (▲)] system

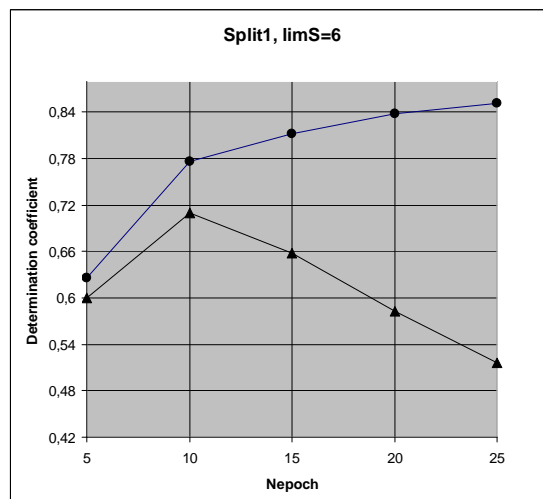


Figure 2. Cont.

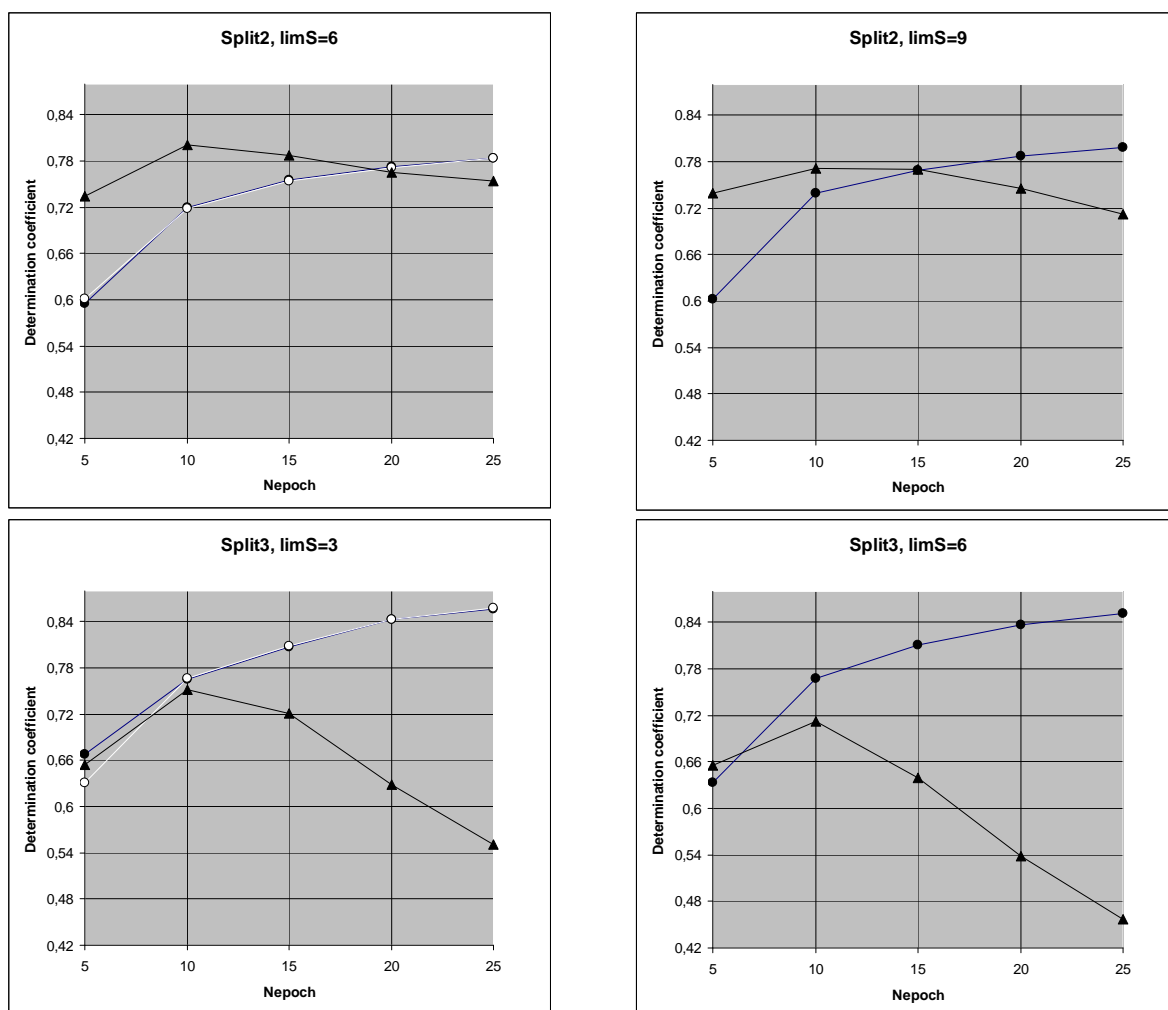


Table 3. Results of computational experiments to establish of number of epochs of the Monte Carlo optimization, Nepoch.

[Subtraining-Calibration-Test] system			
Nepoch	$r^2_{\text{subtraining}}$	$r^2_{\text{calibration}}$	r^2_{test}
Split-1			
5	0.5850	0.6043	0.5513
10	0.7629	0.7675	0.7601
15	0.7939	0.8006	0.7187
20	0.8154	0.8243	0.6827
25	0.8300	0.8262	0.6076
Split-2			
5	0.5947	0.6017	0.7347
10	0.7195	0.7190	0.8011
15	0.7551	0.7538	0.7870
20	0.7732	0.7719	0.7659
25	0.7839	0.7834	0.7538

Table 3. Cont.

Split-3			
5	0.6673	0.6303	0.6548
10	0.7656	0.7669	0.7519
15	0.8077	0.8080	0.7205
20	0.8436	0.8428	0.6288
25	0.8562	0.8581	0.5503
[Training-Test] system			
Split-1			
5	0.6255		0.6003
10	0.7761		0.7098
15	0.8124		0.6579
20	0.8386		0.5826
25	0.8521		0.5158
Split-2			
5	0.6028		0.7397
10	0.7396		0.7719
15	0.7687		0.7705
20	0.7872		0.7452
25	0.7985		0.7123
Split-3			
5	0.6328		0.6559
10	0.7682		0.7127
15	0.8109		0.6397
20	0.8368		0.5378
25	0.8519		0.4573

3. Results

Computational experiments (Figure 3, Table 4) have shown that [Subtraining-Calibration-Test] system gives preferable results in comparison with the [Training-Test] system for all three splits. Thus the correlation balance (i.e., [Subtraining-Calibration-Test] system) improves QSAR model of log(TD₅₀). It is the second successful experiment using the correlation balance for the QSAR analyses [11].

Table 4. Average statistical characteristics of the QSAR model of carcinogenicity (logTD₅₀) for three splits into the subtraining, calibration, and test sets with the limS values of 0-10. For the best models three attempts of the Monte Carlo optimization together with average values are presented, for other models only average values are shown.

SPLIT1

Subtraining set, n=165	Calibration set, n=167	Test set, n=61	SA _k distribution

Table 4. Cont.

[Subtraining-Calibration-Test] system

limS	Nact	R ²	s	F	R ²	s	F	R ²	s	F	W%	N ₁₁₁
0	797	0.8731	0.500	1125	0.8805	0.619	1217	0.5769	0.893	81	42	333
1	622	0.8807	0.485	1203	0.8821	0.621	1235	0.5319	0.942	67	50	314
2	407	0.8275	0.583	783	0.8268	0.703	789	0.6305	0.832	101	70	285
3	321	0.7801	0.658	579	0.7806	0.730	588	0.7102	0.732	145	79	255
4-1	266	0.7622	0.685	522	0.7620	0.734	528	0.7541	0.682	181	82	217
4-2		0.7593	0.689	514	0.7592	0.746	520	0.7483	0.692	175		
4-3		0.7643	0.682	529	0.7647	0.729	536	0.7519	0.678	179		
average		0.7619	0.685	522	0.7619	0.736	528	0.7514	0.684	178		
5	233	0.7247	0.737	429	0.7241	0.770	433	0.7387	0.711	167	85	197
6	203	0.6901	0.781	363	0.6888	0.814	365	0.7129	0.738	148	86	174
7	182	0.6704	0.806	332	0.6710	0.830	337	0.6541	0.812	112	84	153
8	164	0.6528	0.827	307	0.6530	0.844	311	0.7015	0.753	139	87	142
9	152	0.6356	0.847	284	0.6348	0.864	287	0.6378	0.822	105	84	128
10	139	0.6178	0.868	263	0.6218	0.875	271	0.6788	0.777	126	84	117

Training set, n=332	Calibration set, n=0	Test set, n=61	SA_k distribution
--------------------------------	---------------------------------	---------------------------	--

[Training-Test] system

limS	Nact	R ²	s	F	R ²	s	F	R ²	s	F	W%	N ₁₀₁
0	797	0.8868	0.472	2593				0.5429	1.002	71	47	376
1	777	0.8851	0.475	2542				0.5418	0.984	71	46	356
2	542	0.8602	0.524	2032				0.6042	0.910	91	61	330
3	432	0.8313	0.576	1626				0.5575	0.917	74	72	309
4	385	0.8109	0.610	1417				0.5628	0.910	76	75	289
5	344	0.8007	0.626	1327				0.5913	0.871	86	78	267
6-1	312	0.7902	0.642	1243				0.6744	0.769	122	82	255
6-2		0.7875	0.646	1223				0.7138	0.721	147		
6-3		0.7843	0.651	1200				0.6947	0.744	134		
average		0.7873	0.647	1222				0.6943	0.745	135		
7	288	0.7788	0.659	1162				0.6579	0.789	114	83	238
8	268	0.7659	0.678	1080				0.6677	0.777	121	85	227
9	246	0.7363	0.720	922				0.6853	0.757	129	84	207
10	234	0.7224	0.739	859				0.6909	0.750	133	84	196

SPLIT2

Subtraining set, n=165	Calibration set, n=167	Test set, n=61	SA_k distribution
-----------------------------------	-----------------------------------	---------------------------	--

[Subtraining-Calibration-Test] system Split2

limS	Nact	R ²	s	F	R ²	s	F	R ²	s	F	W%	N ₁₁₁
0	797	0.8743	0.507	1134	0.8737	0.540	1142	0.4630	1.055	51	42	337
1	632	0.8740	0.507	1131	0.8736	0.551	1140	0.5003	0.995	59	51	320
2	425	0.8377	0.576	841	0.8367	0.580	846	0.5919	0.820	86	67	286
3	335	0.8048	0.632	673	0.8041	0.633	678	0.5862	0.861	84	78	261
4	284	0.7843	0.664	593	0.7842	0.663	600	0.7042	0.711	141	84	239
5	247	0.7458	0.721	478	0.7448	0.728	482	0.7627	0.671	190	87	214

Table 4. Cont.

6-1	224	0.7315	0.741	444	0.7314	0.748	449	0.7937	0.604	227	84	189
6-2		0.7234	0.752	426	0.7234	0.760	431	0.7922	0.605	225		
6-3		0.7384	0.731	460	0.7384	0.740	466	0.8136	0.593	258		
average		0.7311	0.741	444	0.7310	0.749	449	0.7998	0.600	236		
7	195	0.6978	0.786	376	0.7007	0.781	386	0.7318	0.657	161	84	164
8	178	0.6878	0.799	359	0.6880	0.801	364	0.7223	0.682	153	82	146
9	158	0.6659	0.826	325	0.6692	0.831	334	0.7104	0.709	145	84	133
10	149	0.6472	0.849	299	0.6550	0.847	313	0.6970	0.723	136	84	125

Training set, n=332	Calibration set, N=0	Test set, n=61	SA_k distribution
--------------------------------	---------------------------------	---------------------------	--

[Training-Test] system Split2

limS	Nact	R ²	s	F	R ²	s	F	R ²	s	F	W%	N ₁₀₁
0	797	0.8922	0.468	2734				0.4665	1.013	52	47	372
1	785	0.8950	0.462	2815				0.4711	1.029	53	46	360
2	546	0.8740	0.506	2290				0.5329	0.887	67	61	335
3	442	0.8456	0.561	1807				0.5767	0.845	81	71	315
4	388	0.8194	0.606	1497				0.6130	0.805	94	76	296
5	350	0.8122	0.618	1428				0.5802	0.873	82	79	278
6	321	0.8103	0.621	1412				0.6074	0.840	92	83	267
7	287	0.7848	0.662	1204				0.6689	0.753	120	86	247
8	263	0.7594	0.700	1042				0.7345	0.655	164	87	229
9-1	243	0.7397	0.728	938				0.7472	0.653	174	89	216
9-2		0.7370	0.732	925				0.7862	0.602	217		
9-3		0.7456	0.720	967				0.7604	0.642	187		
average		0.7408	0.726	943				0.7646	0.632	193		
10	228	0.7294	0.742	890				0.7502	0.655	178	86	196

SPLIT3

Subtraining set, n=165	Calibration set, n=167	Test set, n=61	SA_k distribution
-----------------------------------	-----------------------------------	---------------------------	--

[Subtraining-Calibration-Test] system Split3

limS	Nact	R ²	s	F	R ²	s	F	R ²	s	F	W%	N ₁₁₁
0	797	0.8690	0.518	1084	0.8909	0.516	1353	0.5794	0.929	82	42	332
1	614	0.8742	0.508	1134	0.8946	0.513	1402	0.5995	0.896	89	50	309
2	402	0.8266	0.597	778	0.8331	0.614	826	0.6748	0.800	122	69	278
3-1	324	0.7963	0.647	637	0.7982	0.633	652	0.7176	0.729	150	78	254
3-2		0.7919	0.654	620	0.7937	0.639	635	0.6969	0.758	136		
3-3		0.7930	0.652	624	0.7944	0.641	637	0.7431	0.698	171		
average		0.7937	0.651	627	0.7954	0.638	642	0.7192	0.728	152		
4	264	0.7439	0.725	474	0.7462	0.703	485	0.6992	0.765	138	85	224
5	227	0.7127	0.768	404	0.7136	0.738	411	0.6900	0.774	133	86	195
6	198	0.6945	0.792	371	0.7013	0.756	388	0.6899	0.770	133	86	171
7	181	0.6790	0.812	345	0.6843	0.780	358	0.6995	0.758	137	85	154
8	159	0.6432	0.856	294	0.6493	0.815	306	0.7061	0.749	142	84	134
9	147	0.6219	0.881	268	0.6533	0.820	311	0.6934	0.775	134	84	123
10	140	0.5952	0.911	240	0.6269	0.849	277	0.6300	0.842	101	83	116

Table 4. Cont.

Training set, n=332	Calibration set, n=0	Test set, n=61	SA _k distribution
------------------------	-------------------------	-------------------	---------------------------------

[Training-Test] system Split3

limS	N _{act}	R ²	s	F	R ²	s	F	R ²	s	F	W%	N ₁₀₁
0	797	0.8930	0.457	2756				0.5532	1.009	73	47	377
1	776	0.8932	0.457	2763				0.5529	0.996	73	46	356
2	540	0.8699	0.504	2209				0.5998	0.922	89	61	327
3	434	0.8349	0.568	1674				0.5908	0.896	88	72	311
4	388	0.8220	0.590	1528				0.6068	0.865	92	75	291
5	348	0.8030	0.620	1346				0.6650	0.796	117	78	272
6-1	320	0.7773	0.660	1152				0.7017	0.751	139	82	261
6-2		0.7942	0.634	1273				0.6967	0.761	136		
6-3		0.7834	0.651	1193				0.7171	0.735	150		
average		0.7850	0.648	1206				0.7051	0.749	141		
7	288	0.7598	0.685	1045				0.6807	0.778	126	84	241
8	271	0.7637	0.679	1067				0.6520	0.817	112	85	229
9	244	0.7318	0.724	901				0.6833	0.778	127	86	210
10	232	0.7288	0.728	887				0.6826	0.781	127	84	196

A useful characteristic of these models is $W\% = N_{111}/N_{act}$, where N_{111} is the number of non blocked attributes which take place in subtraining, calibration, and test set; N_{act} is the total number of attributes which are not blocked for a given limS. There is a correlation between W% and the determination coefficient for the test set (Figure 4, Table 4). One can see from the results that good prediction occurs if the W% is higher than 80 (excepting [Subtraining-Calibration-Test] for the Split3: in this case $W\%=78$).

The model obtained in the first probe of the Monte Carlo optimization for the split1 with limS=4 is calculated as follows:

$$-\log(TD_{50}) = -0.5981 (0.0074) + 0.1118 (0.0004) * DCW(4) \quad (3)$$

n=165, $r^2=0.7622$, s=0.685, F=522 (subtraining set)

n=167, $r^2=0.7620$, s=0.734, F=528 (calibration set)

n=61, $r^2=0.7541$, s=0.682, F=181 (test set)

Y-scrambling[19,20] for the test set ($N_{shifting}=300[20]$) gave $r^2_{scrambling}=0.0996$

Figure 5 shows the model calculated with Equation 3, graphically. The *Supplementary Materials* contains numerical data on the experimental and calculated values with Equation 3 (split1 with limS=4). Table 5 contains numerical data on the correlation weights of SMILES attributes obtained in three probes of the Monte Carlo optimization.

Figure 3. Comparison of the [subtraining-calibration-test] system and the [training-test] system for three splits.

[Subtraining (●) - Calibration (○) - Test (▲)] system

[Training (●) - Test (▲)] system

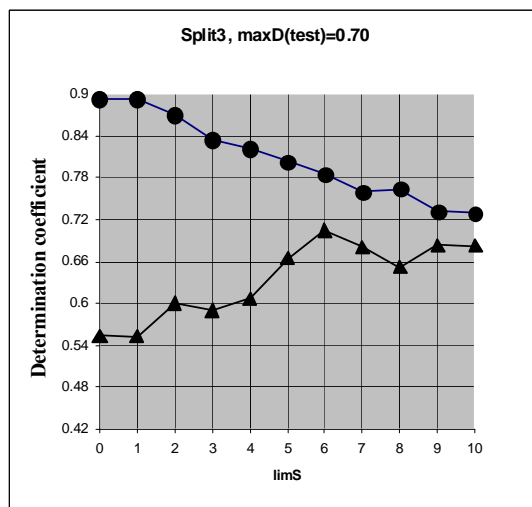
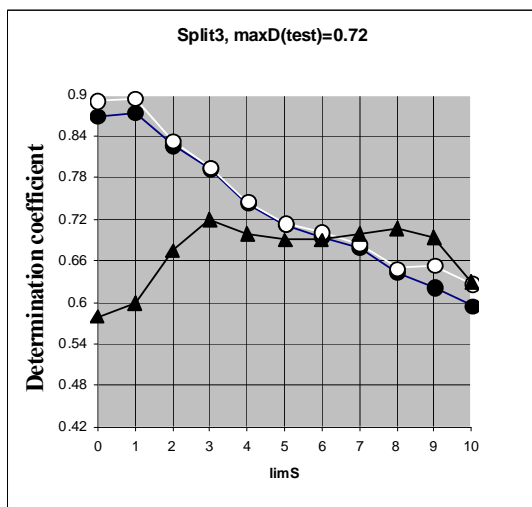
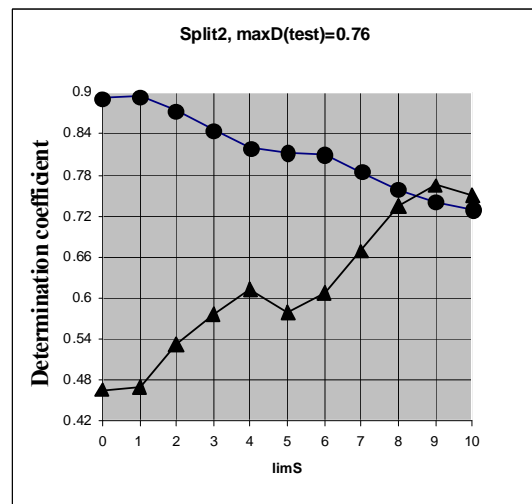
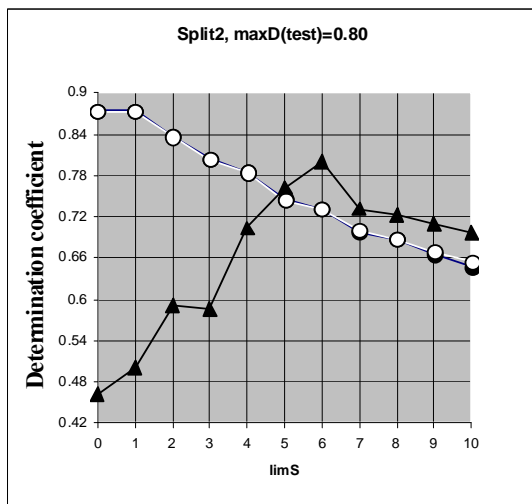
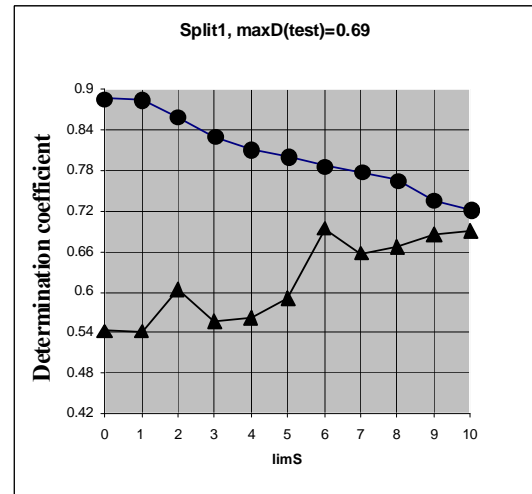
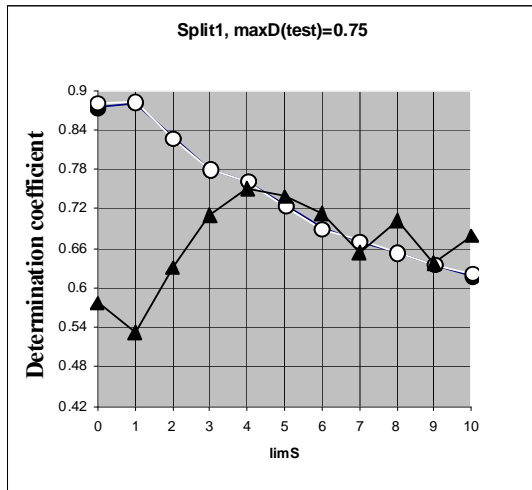


Figure 4. Correlations between the determination coefficient for test set and W% for the three splits (see data from Table 4).

[Subtraining-Calibration-Test] system

[Training-Test] system

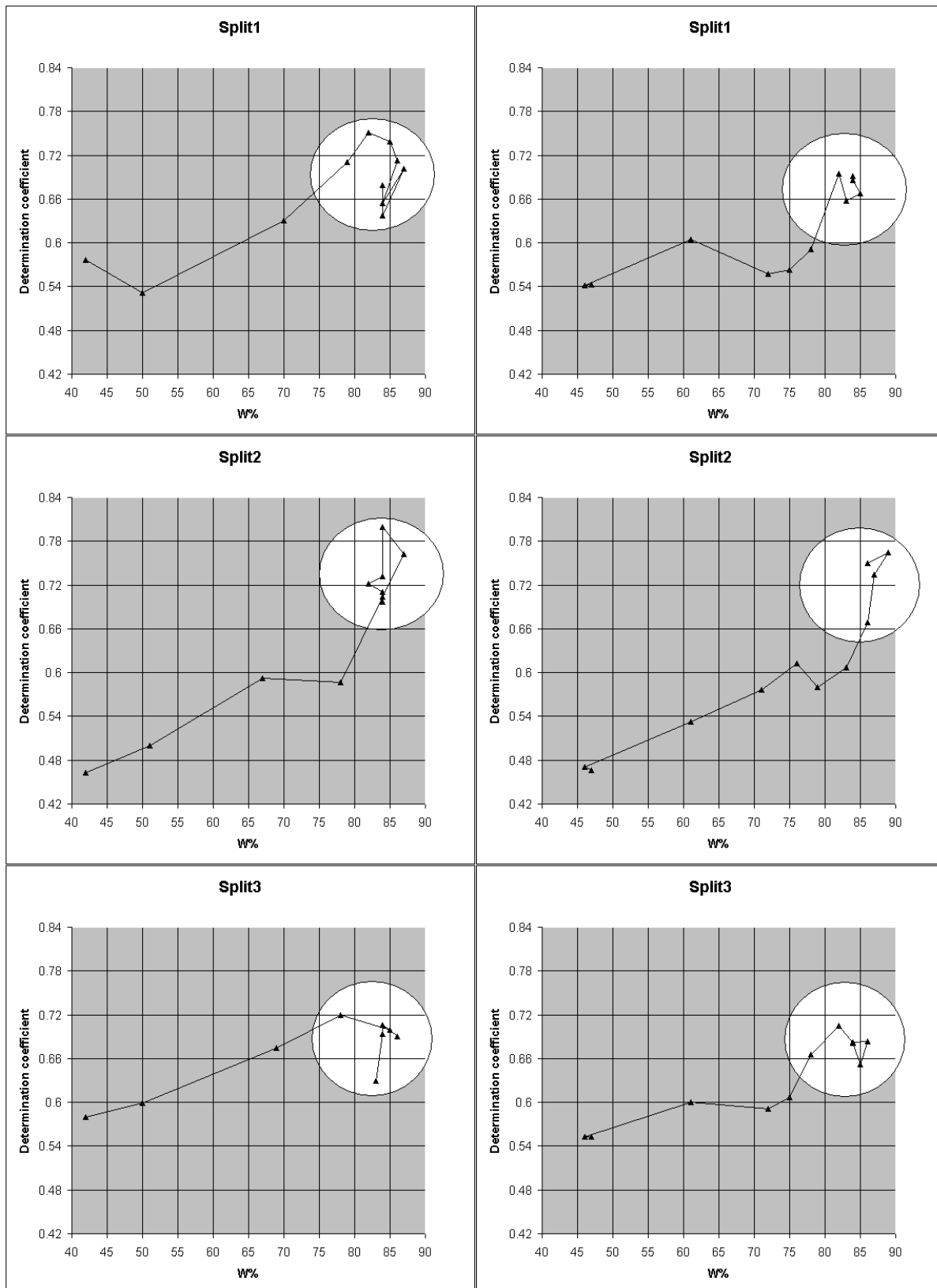


Table 5. Correlation weights for calculation with Equation 1 DCW(4). N(Subtr), N(calib), and N(Test) are numbers of a given SMILES attribute in the subtraining set, calibration set, and test set, respectively. The rare attributes are omitted.

SMILES-Attributes (SA)	CW(SA) probe 1	CW(SA) probe 2	CW(SA) probe 3	N(Subtr)	N(Calib)	N(Test)
dC						
!-01.....	2.7522274	2.8704615	3.5346711	5	4	0
!-02.....	1.2190257	2.1277910	1.8790680	10	9	3
!-03.....	6.6784389	8.0311759	7.1271958	15	10	3
!-04.....	1.4326102	1.6702225	1.9340790	17	22	8
!-05.....	3.9671055	4.0344924	4.1729635	9	11	6
!-06.....	5.8564637	5.8794012	6.4409754	8	11	7
!-07.....	5.4970475	5.1611240	5.2308474	5	3	0
!-08.....	9.1295923	9.5122328	9.0035813	4	3	1
!-21.....	-1.6383248	1.8781962	0.0037831	4	0	0
!000.....	3.6271821	4.6894405	3.7495506	6	7	1
!002.....	1.5603260	1.7450611	1.4951171	4	4	1
!003.....	-1.2514096	-1.3248590	-1.1256941	5	8	2
!004.....	0.7359726	1.0643522	1.2450258	11	8	1
!005.....	0.9702817	0.6240636	1.2144260	13	9	4
!006.....	4.1543029	4.9338830	4.9975361	7	5	2
!007.....	-3.7770327	-3.5039029	-3.0945823	4	3	3
!010.....	0.5049355	-0.2636435	0.3157527	6	8	3
!012.....	3.2511213	3.2471578	4.5049864	6	3	1
¹SA_k						
#.....	3.3706294	3.3739877	2.0643948	5	3	0
(.....	-1.6866726	-1.3666396	-1.5485382	708	780	260
/.....	-0.4913426	0.1880630	-1.0975733	17	24	4
1.....	-1.4970879	-0.8440743	-0.0771659	222	222	88
2.....	-0.1050677	-1.1891334	-1.1138329	130	132	48
3.....	-1.3433340	0.0456678	-0.1828115	60	60	20
4.....	3.4954870	3.1562107	3.5453447	20	18	8
5.....	2.8128037	3.3899959	1.6902086	10	10	4
=.....	-1.8660845	-2.1441609	-1.8865449	77	79	23
C.....	-0.0156855	0.0453525	0.2198595	765	736	290
Br.....	0.5327181	0.2779344	0.8454938	23	8	1
Cl.....	2.9838590	2.1906970	3.1890603	61	85	13
F.....	-0.4680666	-1.0425952	0.2836492	15	19	8
O=C.....	-2.8475657	-2.4376628	-2.9332073	33	21	13
O=.....	0.7369372	0.0037805	1.4086398	140	132	47
N.....	1.1227501	1.1982965	1.4193640	196	201	76
O.....	-1.2649109	-0.4408418	-0.1501499	138	143	45
S.....	2.3712714	2.6251760	2.5313565	13	12	7
[N+].....	1.9345689	1.6543771	3.0457447	26	31	12
[O-].....	5.9250900	5.6230564	6.9653600	26	32	12
[.....	-2.1531745	-2.8080919	-1.9966710	4	6	0
\.....	3.3565892	3.4338813	2.9027414	14	29	7
c.....	-0.0357264	0.0373181	0.0419142	653	679	247
n.....	-0.6564241	-0.1251570	-1.4184164	37	44	23
o.....	-1.0665085	-1.3777640	-0.2485470	16	12	7
s.....	-0.0527175	-0.9991370	-1.0040993	7	6	7
²SA_k						
(...(.....	-0.0735964	-0.1751550	-0.4970432	18	28	4
/...(.....	-0.9972903	-1.5270762	-0.7479799	7	10	2
1...(.....	2.3733452	2.4975765	2.4084744	37	45	15
2...(.....	0.0608136	0.1227718	-0.1848792	14	15	6
2...1.....	5.7529509	6.6287931	7.4713129	5	6	2

Table 5. Cont.

3...(.....	-1.6283079	-0.2528134	-2.0499858	6	3	0
3..2.....	-1.5915226	-1.5268568	-1.9365673	4	6	3
=...(.....	1.8468333	2.3079839	2.4377851	14	17	3
=..1.....	2.5039102	-0.5608527	-0.0325572	7	5	1
=..2.....	-3.2847055	-2.5042239	-2.0658154	7	5	2
=..3.....	3.4389111	0.8278003	3.5625409	6	5	4
C...#.....	-0.1836222	-0.9385750	-0.4107562	6	3	0
C...(.....	-0.7573851	0.0280162	0.7466835	443	456	163
C.../.....	1.1245359	0.0042443	0.8873189	13	13	2
C...1.....	-0.4566262	0.0357389	-1.3911620	74	73	30
C...2.....	0.3115003	1.0911890	0.8401947	46	47	12
C...3.....	3.6534836	3.4678053	2.9866499	40	22	8
C...4.....	-0.7152591	-0.7967591	-1.2502575	17	13	5
C...5.....	3.6909807	4.4229382	4.3015822	10	7	6
C...=.....	-0.5319093	-0.5807285	-0.4569253	98	101	29
C...C.....	-0.4098212	-0.6663667	-0.4722713	244	211	113
Br...(.....	-1.2467411	-0.7804139	-0.9676602	24	7	0
Br..C.....	5.8039394	6.6601591	5.9721683	9	5	1
Cl...(.....	-0.2165917	-0.6443513	-0.7389015	68	104	11
Cl..C.....	6.8768666	7.6839570	7.4343341	17	18	5
F...(.....	0.2020867	-0.0538118	-0.1868874	24	22	12
O=C...(.....	0.7311485	-1.6257029	0.2188934	18	8	5
O=C.1.....	4.3778160	4.7809340	4.0011131	9	6	1
O=...(.....	-0.5612999	-1.5272716	-1.1454337	177	158	60
O=.1.....	-2.5019413	-3.3715192	-4.0028237	4	2	0
N...#.....	-3.8725309	-4.4992215	-3.7843832	4	2	0
N...(.....	0.0666245	0.7453778	-0.1289674	140	165	56
N.../.....	0.8133323	0.0606093	-0.1893841	9	12	2
N...1.....	1.8744868	1.0335496	1.5038557	23	17	10
N...2.....	1.4979132	1.4959901	1.4961647	6	9	3
N...=.....	-1.3157419	0.1537739	-0.3882898	12	16	5
N...C.....	1.4051238	0.9827410	1.0619180	63	70	24
N...O=.....	6.1270291	7.3000058	4.8170313	39	34	13
N...N.....	3.1922498	3.5013150	4.1321688	14	8	8
O...(.....	-0.1195150	-0.1976562	-0.7838811	106	111	31
O...1.....	-0.7620380	-1.4388311	-1.9361803	19	13	5
O...2.....	-2.5618134	-3.2668394	-2.8747322	9	5	3
O...C.....	1.0444154	1.0339726	0.9105754	90	96	29
S...(.....	-0.7479741	0.4990928	-0.0132133	7	8	4
S...=.....	1.5009045	0.6752299	-0.2807681	5	7	2
S...C.....	0.2470117	-1.2535030	-0.5349209	6	1	4
[N+](.....	3.2516821	1.6524330	1.3748828	40	37	17
[O-](.....	-0.4532482	-0.8221590	-1.3547359	39	48	18
[O-][N+]....	0.2616708	0.6284804	-1.2536848	5	6	2
\...(.....	0.2506876	-0.8700648	-1.1268254	5	11	1
\...C.....	2.1710329	2.6262343	1.7619375	11	26	6
\...N.....	-3.1201815	-3.8706242	-3.0325970	4	12	3
c...(.....	0.3275817	-0.1910585	-0.5343311	183	238	94
c...1.....	0.5127781	0.1714980	0.8236519	196	204	75
c...2.....	0.1139969	1.4331593	2.2509927	129	122	38
c...3.....	1.5045372	1.4375414	0.1592669	41	50	15
c...4.....	0.9391582	0.2451376	0.7605772	9	10	10
c...C.....	-1.6459258	0.0580657	-0.4333240	15	19	1
c...Cl.....	-1.9973422	-2.7517912	-3.6905785	5	7	3
c...N.....	1.0896408	0.1897391	0.7548234	26	19	12
c...O.....	2.4331156	1.2515997	0.9178503	22	18	6
c...c.....	-0.2252497	-0.6284915	-0.8624749	316	305	106

Table 5. Cont.

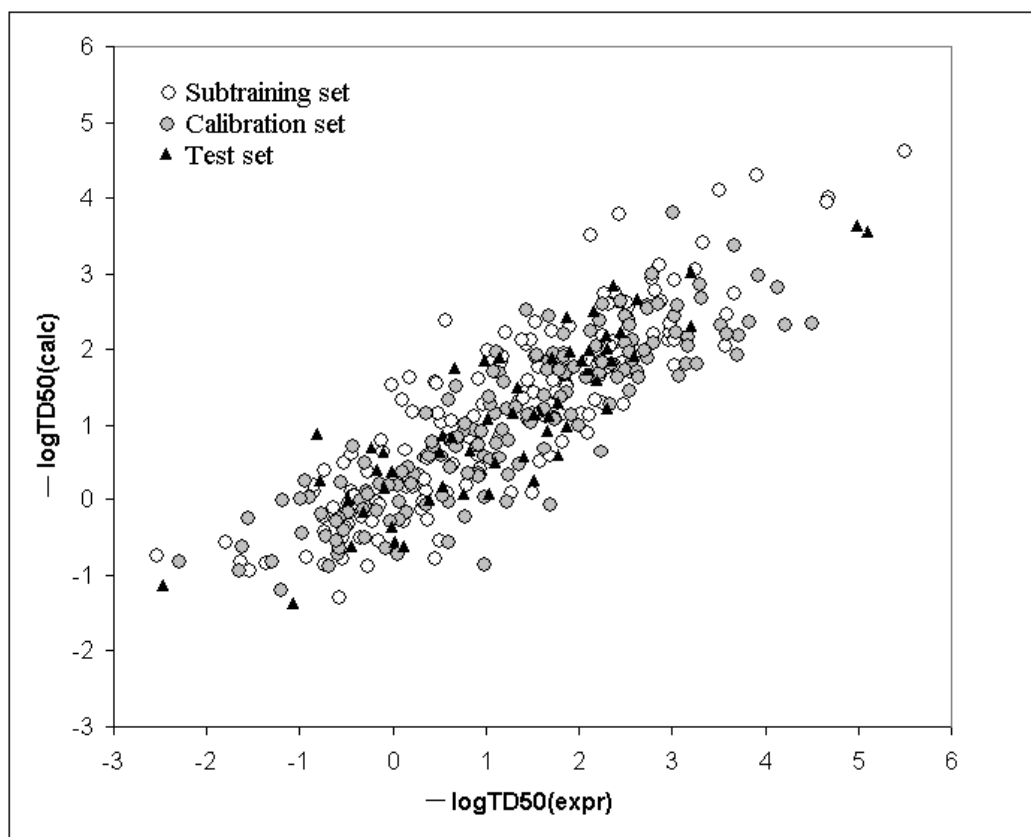
n...(.....	0.8765637	-0.9401023	0.2494319	11	8	6
n...1.....	1.6235455	1.2765370	1.8586068	16	15	10
n...2.....	2.0303385	3.9797145	3.5009942	6	11	8
n...3.....	4.1295873	3.5576218	4.1220666	4	6	3
n...c.....	2.3101715	1.5599987	1.7164852	25	40	17
o...(.....	-3.9990305	-2.9953875	-3.7613936	8	9	6
o...1.....	5.8875962	7.1857177	6.8437068	5	5	2
o...2.....	-0.6199000	-0.1044473	-0.2658548	7	6	5
o...c.....	5.5725605	4.2549084	3.0289124	8	3	1
s...1.....	0.9407152	1.9765325	2.0040980	6	6	7
s...c.....	-0.4960014	0.0005352	0.0003709	4	2	6
³ SA _k						
(...C...(...	3.9980286	3.5630634	2.4208800	95	102	42
(...Br...(...	0.0004754	0.5262619	-1.7321140	9	3	0
(...Cl...(...	0.6233843	-0.3777470	1.3569611	29	44	5
(...F...(...	1.9216525	2.4994253	1.1337512	11	9	5
(...O=(...	1.2586684	0.9193678	0.2775945	71	68	28
(...N...(...	2.1549141	1.5500875	1.3741232	29	40	11
(...O...(...	1.3921586	0.5590769	0.3741692	33	34	8
(...[N+](...	2.2543551	4.4491588	4.0323477	15	8	5
(...[O-](...	-1.5354823	-1.4737180	-3.3779724	19	23	9
(...c...(...	-1.0930229	-0.9342219	0.4332127	12	18	0
/...C...(...	4.4970818	4.0048287	3.0000737	5	4	0
1...C...(...	4.2142525	2.8161885	3.1850279	16	16	5
1...O...(...	2.2528410	0.3797661	1.4952742	4	2	0
1...c...(...	0.9335129	0.5288083	0.8716499	18	35	12
2...C...(...	-2.2544539	-1.0919167	-2.6427067	10	13	4
2...c...(...	2.9972422	3.4417285	3.8400379	29	28	12
2...c...1...	1.9960734	0.5619955	-0.0932258	7	8	2
2...o...(...	1.0110528	1.4973281	1.0603210	4	4	4
3...C...(...	-2.2814415	-2.8156671	-1.9977117	7	6	1
3...C...2...	6.4980735	8.0020020	8.2529271	4	0	0
3...c...(...	-0.2171379	1.0578016	1.2483687	9	9	4
3...c...2...	5.2502023	4.2226554	4.7487134	8	7	2
4...C...(...	1.7464204	-0.6210801	-0.2842473	6	4	0
=...C...3...	7.0029958	7.5641676	6.7549811	8	4	0
=...C...1...	2.9999456	2.8474909	4.2545986	12	8	4
=...C...(...	1.5713076	0.9521538	0.5309943	18	18	2
=...C.../...	5.2495800	5.4994084	6.0049990	6	7	2
=...N.../...	5.8119346	5.5954944	6.3789368	7	8	2
C...(...C...	0.5513380	-0.2378199	-1.0864076	69	64	24
C...(...1...	-1.1216831	-3.3768328	-2.4951909	9	10	3
C...(...=...	6.2535976	5.1587497	4.4360925	9	11	3
C...(...(...	-0.3146880	-1.6918223	-1.2834724	11	22	3
C.../...(...	-3.0637430	-1.8160468	-2.9987839	5	4	1
C...1...C...	5.4954661	3.9189550	4.6140280	8	10	5
C...1...(...	1.3718529	1.3079333	1.4795097	8	13	1
C...1...=...	0.2476856	1.3169467	0.9333222	6	4	1
C...2...(...	-0.6449419	-0.8430901	-0.7370108	5	9	0
C...2...C...	5.9965379	5.9966533	6.0017257	8	6	1
C...2...=...	0.0028591	-0.8747150	-0.3764369	7	5	2
C...3...(...	6.5045056	6.2477966	5.7492216	5	3	0
C...3...=...	5.6231609	6.0002916	5.1293498	5	3	2
C...3...C...	-3.5000076	-2.9954231	-3.0028309	11	2	3
C...4...C...	-3.0021372	-4.5016046	-2.9968826	4	1	1
C...=...1...	0.4331526	2.6582434	1.9050556	7	5	1

Table 5. Cont.

C...=(...	-2.4953051	-0.6283873	-0.7473193	8	11	1
C...=...C...	1.6093540	2.7975919	2.1897927	33	34	11
C...=...3...	0.2971809	1.5013043	-0.9639603	5	3	2
C...=...2...	1.6275637	3.7172258	1.7773288	7	4	0
C...C...3...	5.0746157	4.6119002	4.5908880	16	8	5
C...C...=...	-0.4018420	-1.0583619	-1.0977818	36	26	5
C...C...1...	1.8775937	0.9384077	1.3828674	31	27	17
C...C...2...	0.7079969	-0.1879112	0.6274962	21	19	3
C...C...(...	1.0123078	0.7924398	0.4047303	109	109	42
C...C...4...	-1.4024533	-0.7803070	-0.3715353	7	5	2
C...C...C...	-0.0428402	0.1882515	-0.1515135	77	58	59
C...Br...(...	1.6298040	2.5038504	1.0267176	4	1	0
C...Cl...(...	-1.4962815	-0.4993582	-1.2516847	4	4	1
C...N...1...	-0.2691219	-0.3795019	-1.0000894	8	8	1
C...N...(...	1.4422504	0.9731136	0.4339815	36	35	18
C...O...2...	5.1236472	4.0895698	3.5121038	5	3	1
C...O...(...	3.2515162	2.2843408	2.5780816	28	32	12
C...O...C...	4.3741698	4.3105041	3.0634685	8	10	4
C...O...1...	2.8733789	3.1267319	2.9673439	13	9	3
C...\...C...	-2.8461979	-3.8759584	-2.2789129	4	8	1
C...c...2...	6.0006860	4.5315715	5.2468516	4	5	0
C...c...1...	2.4356912	0.4351720	1.2472872	10	13	1
Br...(C...	2.0615855	1.1437599	1.9040185	4	6	0
Br.C...(...	1.4981969	0.6256406	0.0009660	7	3	0
Cl...(C...	-1.2075807	-0.6362162	-1.9255180	9	6	0
Cl...(C...	-1.1526609	-0.2476848	-1.8147825	27	32	4
Cl...(Cl...	0.5049208	3.2539441	0.6886852	4	7	0
Cl.C...C...	-0.0014586	0.0039516	-1.2464369	9	10	2
Cl.c...1...	-0.2533902	1.6295626	2.2512123	4	6	3
F...(C...	1.6863754	1.5015167	0.5346605	5	8	2
F...(C...	-1.7457403	-2.1139078	-1.1982770	6	8	4
O=C.1...C...	5.1279970	2.8669862	3.6914477	4	3	1
O=...(C...	0.8107109	0.7780984	-0.4033174	92	68	31
O=.1...C...	-3.7510578	-4.1255552	-4.1222938	4	2	0
O=.N...(...	9.5435183	10.0636543	10.0315899	24	28	6
N...#...C...	-4.5000318	-4.5014930	-4.5004055	4	1	0
N...(N...	1.1916803	0.9983579	1.6288201	12	10	2
N...(1...	-0.1264931	-0.7538114	0.0454142	5	7	0
N...(C...	3.5018836	2.2822748	3.0336305	55	62	30
N...(O=...	-2.3138510	-2.3145054	-1.5031092	23	14	5
N...(O=C...	-1.4990141	-0.8704228	-1.2494712	6	5	2
N...1...C...	2.6914072	2.5647275	2.7184638	12	13	4
N...2...C...	-0.4978517	-0.0000051	1.0021092	5	6	1
N...C...(...	-0.8104915	-0.4341890	-0.9684013	25	24	8
N...C...C...	-1.2520104	-0.7226801	-1.0008159	22	26	6
N...O=...(...	2.6715972	1.6272171	3.6550784	11	11	1
N...N...1...	0.0042593	1.4970138	1.1825822	5	3	3
N...N...O=...	4.2459870	3.7466101	3.2221499	10	6	5
N...N...(...	4.7536740	5.4055258	4.6294702	6	4	2
N...c...2...	-3.6269860	-2.8792235	-4.2532265	5	3	1
N...c...1...	-0.1899251	0.2338572	0.2627307	20	15	11
O...(O=...	-0.6219316	1.4395827	1.2532594	19	17	5
O...(C...	0.9395840	-0.4347300	0.4467587	52	40	17
O...(C...	11.5040533	12.0049318	11.9989881	4	4	1
O...(O=C...	4.9368982	4.9994000	4.7472261	7	2	0
O...C...1...	-0.4987416	0.4978181	0.9413176	4	4	3

Table 5. Cont.

O...C...C...	-2.6559832	-3.2412835	-3.2024047	35	37	10
O...C...(...	-0.1201364	-0.3106109	-1.2477221	27	31	10
O...c...1...	-2.7623334	-2.0920952	-2.1242060	14	8	3
O...c...2...	-1.4980614	-3.5286748	-3.6223778	7	6	0
S...C...C...	-0.0034408	1.5042803	0.7537994	4	0	2
[N+](...C...	9.2539066	9.4417355	7.5006239	6	4	1
[N+](...2...	5.4109375	4.6273127	3.2821267	6	3	4
[N+](...O=...	-0.3787790	0.3109436	0.1872916	4	7	2
[O-](...[N+]	-3.8743809	-1.4388021	-1.2453181	18	22	9
[O-](...O=...	-4.0585677	-2.6242009	-2.5577359	15	11	6
[O-][N+](...	-3.5045096	-1.4982980	-2.4892797	5	5	2
\...C...=...	-1.3136029	-1.8755430	-1.2854492	4	11	1
\...C...(...	-3.5018378	-4.4994516	-3.8096741	5	7	2
c...([O-]	3.9992170	4.0612326	4.4978479	4	10	3
c...(...c...	1.7523875	2.5921235	0.9359654	24	19	13
c...(...Br...	1.3392341	0.5340779	1.1889034	17	0	0
c...(...C...	1.0002010	0.2472155	-0.2478234	19	41	13
c...(...Cl...	1.1825597	2.5039425	0.9088978	15	23	7
c...(...O...	1.1553993	0.9107402	1.8401083	10	34	6
c...(...N...	-0.4647652	-0.5109250	0.5049629	13	41	11
c...(...1...	3.2546395	1.6826438	1.7521385	17	17	10
c...(...O=...	2.0008786	2.9044603	2.8172049	15	17	7
c...(...F...	2.5615727	2.2483847	2.9956341	6	0	2
c...1...O...	0.3157218	0.2460029	0.0026575	7	3	3
c...1...C...	-0.3426326	-0.4053179	-0.0587013	10	10	4
c...1...(...	4.1291472	3.7385617	4.5010102	15	17	6
c...1...c...	2.5270201	4.1254591	1.6392878	64	69	24
c...2...c...	3.1834674	3.5765233	2.5649848	46	41	10
c...2...O...	-2.1902681	-0.8169002	-1.5600959	6	5	0
c...2...C...	2.0599980	3.3166075	2.0508240	6	4	2
c...2...(...	-3.4837706	-2.7454921	-2.2497303	5	2	2
c...3...c...	0.5671953	0.3280375	2.5577465	14	15	6
c...C...C...	-0.4968006	-1.0630716	-0.7464899	4	6	0
c...N...(...	4.1212930	5.1280879	3.2342653	9	3	3
c...O...(...	7.8795041	8.7529500	8.5448576	5	2	0
c...O...C...	0.4969760	1.0602889	1.0035994	10	8	0
c...c...2...	-1.0046754	-0.7477295	-1.3148992	59	58	20
c...c...c...	-0.9189229	-1.1362229	-0.9886103	171	148	50
c...c...1...	0.9684404	1.0961687	1.0267859	111	101	36
c...c...3...	-1.4056269	-2.8661586	-1.7490548	18	24	6
c...c...4...	-1.2498300	0.6278968	0.4997288	5	5	4
c...c...(...	-0.5592802	-0.7452834	-0.2831183	87	110	45
c...n...1...	0.4037274	0.7545182	1.8635708	8	9	8
n...1...c...	1.1446162	1.1906216	0.1368885	11	8	8
n...c...c...	-4.4951810	-4.4955509	-4.2500468	5	11	1
n...c...(...	-1.7475062	-0.9098866	-0.0016730	10	13	13
o...(...c...	1.9983265	-0.3077248	1.1610603	5	5	5
o...1...(...	-0.8795536	-0.8151611	-1.4961309	4	3	2
s...1...(...	3.0007359	3.3126224	2.8719278	5	5	6

Figure 5. Graphical representation of the model for $\log\text{TD}_{50}$ calculated with Equation 3.

4. Discussion

One can see that the statistical characteristics of this model are reasonably good. As additional validation we have calculated Y-scrambling criterion, randomly shifting the carcinogenicity values [16,17]. If after the shifting (300 exchanges recommended in Ref.[17]) the correlation coefficient is less than 0.2, the correlation of our model can be classified as not chance correlation. Thus, the Y-scrambling has shown that the Equation 3 gives robust prediction (not chance correlation) for the test set.

In our previous study we examined different equations for the carcinogenicity model, and only one split into the subtraining, calibration and test set [15]. Examination of three splits indicates that good results occur for all three splits (Table 4). Thus, we expect that the present model is more robust, also considering the Y-scrambling test.

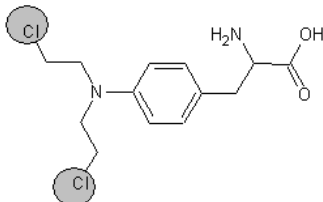
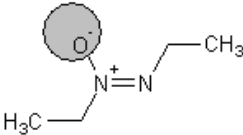
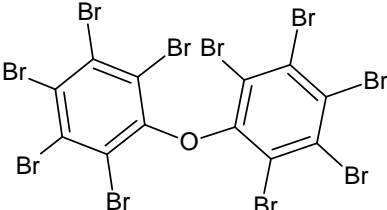
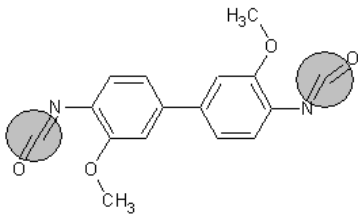
One can see from Table 5 that there are three categories of SMILES attributes: category 1 is the set of SMILES attributes with the correlation weight more than zero in all three probes of the Monte Carlo optimization; category 2 is the set of SMILES attributes with the correlation weight less than zero in all three probes; category 3 is the set of SMILES attributes with non consistent values, which have both correlation weights more than zero and correlation weights less zero in the three probes of the optimization. We can say that the category 1 contains promoters of $\log\text{TD}_{50}$ increase; category 2 contains promoters of $\log\text{TD}_{50}$ decrease; category 3 contains attributes with unclear influence on $\log\text{TD}_{50}$.

The !-02, #, Cl, S, [N+], and [O-] SMILES elements are promoters of logTD₅₀ increase, thus of carcinogenicity. However it is necessary to take into account the value of correlation weight as well as the number of the given attribute in the subtraining set. Taking this into account, one can detect that the strongest promoters of the logTD₅₀ increase are Cl (number Cl in the subtraining set is 61, the range of correlation weights of the Cl in three probes is 2.19 - 3.19) and [O-] (the number of [O-] in the subtraining set is 26, the range of correlation weights in three probes is 5.92 - 6.96).

A similar analysis can be done for the promoters of logTD₅₀ decrease. For instance, the number of bracket s(' in the subtraining set is 708 and the range of correlation weights of bracket is from -1.366 till -1.686; the number of '=' in the subtraining set is 77 and the range of correlation weight is from -1.866 till 2.144. Table 6 contains examples of compounds, which contain the mentioned SMILES attributes. Thus, the analysis of the correlation weights of SMILES attributes can help in searching for agents of the carcinogenicity phenomenon.

An important feature of our model is that SMILES attributes are used for the QSAR predicted values and not only as tool for a binary classification (carcinogenic or not). Our model, which provides continuous values, can be used for risk assessment calculations, where a dose is necessary.

Table 6. Examples of compounds which contain promoters of increase/decrease of the logTD₅₀.

Structure	CAS and SMILES	logTD ₅₀
	148-82-3 <chem>O=C(O)C(N)Cc1ccc(cc1)N(CCCl)CCCl</chem>	3.512
	16301-26-1 <chem>[O-][N+](CC)=N\CC</chem>	3.667
	1163-19-5 <chem>BrC2c(Oc1c(Br)c(Br)c(Br)c(Br)c1Br)c(Br)c(Br)c(Br)c2Br</chem>	-0.542*
	91-93-0 <chem>COc1cc(ccc1/N=C=O)c2ccc(\N=C=O)c(OC)c2</chem>	-0.740*

*) One can see that aromatic bonds are indicated in SMILES by 'c' (lower case), thus '=' is indicator of local double bonds which are not a part of aromatic fragments.

The applicability domain for these models can be defined from a probabilistic point of view: one can estimate the carcinogenic potential of compound if the SMILES of this compound does not contain rare SMILES attributes. A stronger definition of the applicability domain can be formulated taking into account the roles of the attributes (as promoters of logTD₅₀ increase/decrease): thus, one can estimate the carcinogenic potential of a compound if the SMILES of the compound contains solely apparent promoters of logTD₅₀ increase and/or decrease (without of SMILES attributes with unclear role).

5. Conclusions

- Optimal descriptors calculated by the Monte Carlo method can provide reasonable prediction for the carcinogenicity log(TD50).
- Blocking of rare SMILES attributes can improve statistical quality of the predicting. Splits into subtraining, calibration and test sets, as well splits into the training and test sets have influence to statistical characteristics of the models. In our case, in three splits examined in this study these characteristics are similar.
- The correlation balance, i.e., the **[Subtraining-Calibration-Test] system** gave models which are better in comparison with models obtained with the more traditional **[Training-Test] system**.

Acknowledgements

The authors thank the Marie Curie Fellowship (the contract ID 39036, CHEMPREDICT) and the EC project CAESAR (Project no. 022674 (SSPI)) for financial support.

References and Notes

1. Benfenati, E.; Benigni, R.; Demarini, D.M.; Helma, C.; Kirkland, D.; Martin, T.M.; Mazzatorta, P.; Ouedraogo-Arras, G.; Richard, A.M.; Schilter, B.; Schoonen, W.G.; Snyder, R.D.; Yang, C. Predictive Models for Carcinogenicity and Mutagenicity: Frameworks, State-of-the-Art, and Perspectives. *J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev.* **2009**, *27*, 57-90.
2. Benigni, R.; Netzeva, T.; Benfenati, E.; Bossa, C.; Franke, R.; Helma, C.; Hulzebos, E.; Marchant, C.; Richard, A.; Woo, Y.-T.; Yang, C. The expanding role of predictive toxicology: An update on the (Q)SAR models for mutagens and carcinogens. *J. Environ. Sci. Health C* **2007**, *25*, 53-97.
3. Benigni, R. Structure-activity relationship studies of chemical mutagens and carcinogens: Mechanistic investigations and prediction approaches. *Chem. Rev.* **2005**, *105*, 1767-1800.
4. Contrera, J.F.; MacLaughlin, P.; Hall, L.H.; Kier, L.B. QSAR modeling of carcinogenic risk using discriminant analysis and topological molecular descriptors. *Curr. Drug Dis. Technol.* **2005**, *2*, 55-67.
5. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36.

6. Weininger, D.; Weininger, A.; Weininger, J.L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97-101.
7. Weininger, D. SMILES. 3. DEPICT. Graphical depiction of chemical structures. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 237-243.
8. Vidal, D.; Thormann, M.; Pons, M. LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386-393.
9. Toropov, A.A.; Benfenati, E. Optimisation of correlation weights of SMILES invariants for modelling oral quail toxicity. *Eur. J. Med. Chem.* **2007**, *42*, 606-613.
10. Toropov, A.A.; Benfenati, E. Additive SMILES-based optimal descriptors in QSAR modelling bee toxicity: Using rare SMILES attributes to define the applicability domain. *Bioorg. Med. Chem.* **2008**, *16*, 4801-4809.
11. Toropov, A.A.; Rasulev, B.F.; Leszczynski, J. QSAR modeling of acute toxicity by balance of correlations. *Bioorg. Med. Chem.* **2008**, *16*, 5999-6008.
12. Toropov, A.A.; Toropova, A.P. QSAR Modeling of Mutagenicity Based on Graphs of Atomic Orbitals *Internet Electron. J. Mol. Des.* **2002**, *1*, 108-114.
13. Marino, D.J.G.; Peruzzo, P.J.; Castro, E.A.; Toropov, A.A. QSAR Carcinogenic Study of Methylated Polycyclic Aromatic Hydrocarbons Based on Topological Descriptors Derived from Distance Matrices and Correlation Weights of Local Graph Invariants *Internet Electron. J. Mol. Des.* **2002**, *1*, 115-133.
14. Peruzzo, P.J.; Marino, D.J.G.; Castro, E.A.; Toropov, A.A. QSPR Modeling of Lipophilicity by Means of Correlation Weights of Local Graph Invariants *Internet Electron. J. Mol. Des.* **2003**, *2*, 334-347.
15. Available online: <http://chem.sis.nlm.nih.gov/chemidplus/>.
16. Available online: <http://webbook.nist.gov/chemistry/>.
17. Available online: http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html/.
18. Toropov, A.A.; Toropova, A.P.; Benfenati, E.; Manganaro, A. QSAR modelling of carcinogenicity by balance of correlations. *Mol. Divers.* 2009, in press.
19. Mazzatorta, P.; Smiesko, M.; Piparo, E.; Benfenati, E. QSAR model for predicting pesticide aquatic toxicity. *J. Chem. Inf. Model.* **2005**, *45*, 1767-1774.
20. Fatemi, M.H.; Haghadi, M. Quantitative structure-property relationship prediction of permeability coefficients for some organic compounds through polyethylene membrane. *J. Mol. Struct.* **2008**, *886*, 43-50.