Cent. Eur. J. Chem. • 8(5) • 2010 • 1047–1052
DOI: 10.2478/s11532-010-0072-5

VERSITA

Central European Journal of **Chemistry**

QSPR modelling of normal boiling points and octanol/water partition coefficient for acyclic and cyclic hydrocarbons using SMILES-based optimal descriptors

Research Article

A.A. Toropov[*1,2], A.P. Toropova[1,2], E. Benfenati[2]

[1]*Institute of Geology and Geophysics, 100041 Tashkent, Uzbekistan*

[2]*Istituto di Ricerche Farmacologiche Mario Negri, 20156 Milano, Italy*

Received 26 November 2009; Accepted 8 May 2010

Abstract: Predictive quantitative structure – property relationships (QSPR) have been established for normal boiling points and octanol/water partition coefficient for acyclic and cyclic hydrocarbons using optimal descriptors calculated with simplified molecular input line entry system (SMILES). The probabilistic criteria for a rational definition of the domain of applicability of these models are discussed.

Keywords: *QSPR • SMILES • Normal boiling point • Octanol/water partition coefficient • Applicability domain*

1.Introduction

Quantitative structure – property/activity relationships (QSPR/QSAR) [1-3] are aimed to answer: what factors operate during a physical, chemical, and/or biochemical phenomenon? Representation of the molecular structure for these searches is based on molecular descriptors which as a rule, are presented as atoms (vertices) and their bonds (edges) in molecular graphs [4]. A conceptual alternative to molecular graphs is SMILES [5-7]. Fragment methods, in general, and SMILES-based optimal descriptors, in particular, are clear enough to interpret: each molecular fragment shows a defined influence of the phenomenon.

There are several reasons to search for SMILES-based QSPR/QSAR models. The first, comparison of models based on the molecular graph and those based on SMILES can be useful from a heuristic point of view. The second, the number of databases available on the Internet gradually increases. The third, SMILES notation can be built for substances which cannot be represented by molecular graphs, *e.g.* for mixtures or for inorganic substances.

The present study evaluated the ability of the SMILES-based fragment method for QSPR modelling of normal boiling points and octanol/water partition coefficient.

2. Experimental Procedure

2.1. Method

Data on normal boiling points (BP) and octanol/water partition coefficient (logP) for 90 cyclic and acyclic hydrocarbons were taken from [8]. Canonical SMILES [9] were used. The training set (n=45) and test set (n=45) were split randomly.

Descriptors of the correlation weights (DCW) were calculated by

* *E-mail: aatoropov@yahoo.com*

$$DCW = \Pi\, CW(s_k)\ \Pi\, CW(ss_k)\ \Pi\, CW(sss_k) \quad (1)$$

where s_k, ss_k, and sss_k are SMILES attributes (SA_k) of one, two or three elements, respectively. The element of the SMILES can be a symbol of the SMILES notation (for instance, 'c', 'C', 'n', 'N', =, *etc.*), or two symbols of the SMILES encoding an image (for instance, 'Cl', 'Br', 'N+', 'O-', *etc.*); CW(x) is the correlation weight for the SMILES attribute x. The CWs are calculated by the Monte Carlo optimization procedure [6,7] that provides values of the CWs which used in Eq. 1 give a maximum for the correlation coefficient between the descriptor and property of interest. The notation of the SMILES elements was used according to the ASCII codes of the symbols. Each 'AB' composition can be represented in only one way (thus only 'AB' and not 'BA', and similarly only 'ABC' not 'CBA').

There are 66 SMILES attributes for the 90 substances under consideration. Some of the attributes are rare or absent in the training set. The threshold LimN described in [7] can be used for rational selection of the attributes: if a SMILES attribute occurs in the training set less than LimN times its correlation weight (CW) is defined as 1. Eq. 1 shows that CW=1 can not change the DCW value.

The prevalence of SA_k in the training and test sets is important. The relative prevalence (RP) of SA_k in the two sets can be calculated as

$$\Delta P(SA_k) = \frac{NS_{TR}(SA_k)}{NS_{TR}} - \frac{NS_{TS}(SA_k)}{NS_{TS}} \quad (2)$$

where $NS_{TR}(SA_k)$ and $NS_{TS}(SA_k)$ are the numbers of

Table 1. Probabilistic characteristics of the active SMILES attributes (SA_k) with LimN=5 and their correlation weights for modelling the normal boiling points $CW(SA_k)_{BP}$ and the octanol/water partition coefficient $CW(SA_k)_{logP}$

SA_k	ID	No.	$CW(SA_k)_{BP}$	$CW(SA_k)_{logP}$	$N_{TR}(SA_k)$	$N_{TS}(SA_k)$	$NS_{TR}(SA_k)$	$NS_{TS}(SA_k)$	$dP(SA_k)$
(_____	1	1	0.9993545	1.0009505	86	100	28	33	-0.11111
1_____	2	2	1.0104724	0.9983162	60	60	30	30	0.0
2_____	3	3	1.0067214	1.0010740	28	30	14	15	-0.02222
3_____	4	4	0.9957615	0.9965525	6	6	3	3	0.0
C_____	6	5	1.0041399	1.0059730	186	182	41	40	0.02222
c_____	8	6	1.0043571	1.0049424	256	262	30	30	0.0
2___1_____	12	7	1.0040442	0.9988117	5	5	5	5	0.0
C___(_____	14	8	1.0036837	1.0001987	106	134	25	29	-0.08889
C___1_____	15	9	1.0079944	1.0025893	5	10	5	10	-0.11111
C___C_____	18	10	1.0024780	1.0021379	84	65	26	25	0.02222
c___(_____	20	11	1.0026192	0.9970252	54	56	18	18	0.0
c___1_____	21	12	1.0006219	0.9962378	90	91	30	30	0.0
c___2_____	22	13	1.0039974	1.0019023	41	44	13	15	-0.04444
c___3_____	23	14	1.0054033	1.0035743	10	10	3	3	0.0
c___C_____	24	15	1.0058187	1.0063872	25	25	25	25	0.0
c___c_____	25	16	0.9987972	1.0016258	144	146	30	30	0.0
(___C___(___	26	17	0.9996478	1.0006016	31	44	19	25	-0.13333
1___c___(___	29	18	0.9989729	0.9989317	8	7	8	7	0.02222
2___c___1___	31	19	1.0008488	1.0026416	5	4	5	4	0.02222
C___(___(___	35	20	0.9995191	1.0019752	6	4	3	2	0.02222
C___(___C___	36	21	0.9992593	0.9993537	31	46	12	15	-0.06667
C___C___(___	39	22	1.0019924	1.0008759	41	39	14	17	-0.06667
C___C___C___	41	23	1.0040603	1.0038414	40	24	17	10	0.15556
C___c___1___	45	24	1.0045198	1.0021810	21	23	21	23	-0.04444
c___(___C___	48	25	0.9990960	1.0036468	37	38	14	14	0.0
c___(___c___	51	26	0.9978614	1.0079281	6	6	5	5	0.0
c___1___c___	52	27	1.0013400	1.0035067	30	31	30	30	0.0
c___1___2___	54	28	1.0030506	1.0018722	5	5	5	5	0.0
c___1___C___	55	29	1.0007957	1.0052779	5	10	5	10	-0.11111
c___2___c___	57	30	0.9987250	0.9946108	19	19	13	14	-0.02222
c___C___C___	61	31	0.9990522	1.0035252	8	8	8	8	0.0
c___c___2___	62	32	1.0012829	0.9998930	28	31	13	15	-0.04444
c___c___(___	63	33	1.0005657	1.0013465	36	31	17	18	-0.02222
c___c___3___	64	34	1.0060274	0.9990971	8	7	3	3	0.0
c___c___1___	65	35	1.0019073	0.9998137	52	51	30	30	0.0
c___c___c___	66	36	1.0029710	0.9990810	82	86	26	27	-0.02222

ID = number of SA_k in the whole list of 514 attributes
No. = number in list of active attributes
$N_{TR}(SA_k)$ = the total number of SA_k in the training set
$N_{TS}(SA_k)$ = the total number of SA_k in the test set
$NS_{TR}(SA_k)$ = number of SM ILES in the training set containing SA_k
$NS_{TS}(SA_k)$ = number of SMILES in the test set containing SA_k
$dP(SA_k)$ = relative prevalence of SA_k in the training and test sets.

Table 2. Example of the DCW calculation for modelling normal boiling points SMILES = "CCCC" No.1; CAS = 106-97-8 DCW = 1.0325737

SAk	CW(SA_k)
C_____	1.0041399
C_____	1.0041399
C_____	1.0041399
C_____	1.0041399
C___C_____	1.0024780
C___C_____	1.0024780
C___C_____	1.0024780
C___C___C___	1.0040603
C___C___C___	1.0040603

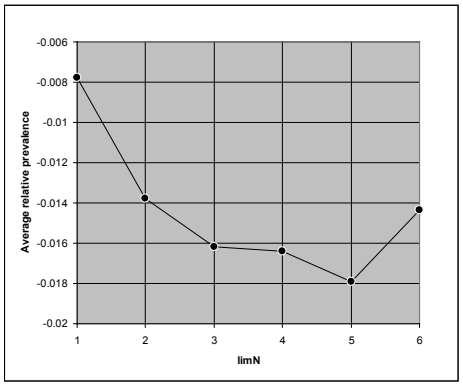


Figure 1. LimN versus the average relative prevalence

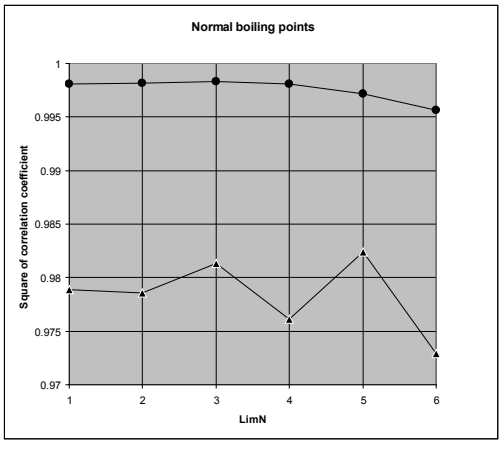


Figure 2. Square of correlation coefficient versus the LimN for the training (circles) and test (triangles) sets for modelling normal boiling points.

SMILES containing the SA_k in the training and test sets, respectively and NS_{TR} and NS_{TS} are the total numbers of SMILES in the two sets. The ideal situation is $\Delta P(SA_k) = 0$, but $\Delta P(SA_k) > 0$ (RP of SA_k in the training set is higher than in the test set) and $\Delta P(SA_k) < 0$ (RP of the SA_k is higher in the test set) are also possible.

The average RP over all active (not blocked) SA_k is calculated as follows

$$\overline{\Delta P} = \frac{\sum \Delta P(SA_k)}{N_{act}} \quad (3)$$

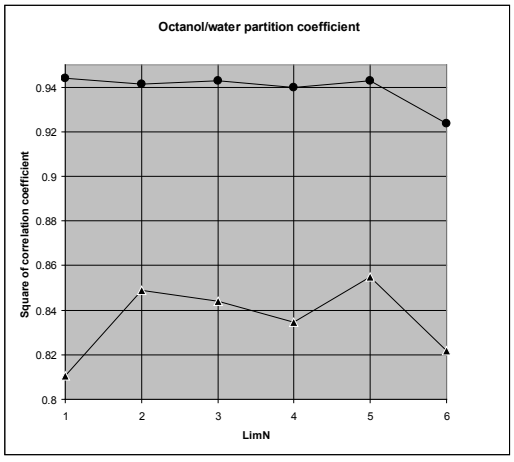


Figure 3. Square of correlation coefficient versus the LimN for the training (circles) and test (triangles) sets for modelling the octanol/water partition coefficient

where N_{act} is the total number of active SA_k. This parameter is a characteristic of the split into the training and test set, and the ideal situation is

$$\overline{\Delta P} = 0 \quad (4)$$

However, this situation hardly is met.

The realistic estimation of the predictability for QSPR/QSAR models becomes a central problem for this field of theoretical chemistry (*i.e.*, for the QSPR/QSAR analyses). Criterions which are calculated with Eqs. 2 and 3 are aimed to involve probabilistic principles for solution of the above-mentioned problem (estimation of the predictability).

3. Results and Discussion

If LimN=5, the curve for the plot of the limN versus $\overline{\Delta P}$ has a minimum (Fig. 1). The statistical quality of the models for normal boiling points and octanol/water partition coefficient is best when LimN=5. Table 1 shows the probabilistic characteristics of the SMILES attributes which are active when LimN=5. Table 2 gives an example of the DCW calculation (for normal boiling points, limN=5).

The following models are obtained using SMILES-based optimal descriptors:

BP =-2422.9264(±4.401) +
+2359.8772(±3.950) * DCW (5)

n=45, r^2=0.997, q^2=0.9969, s=4.39 (°C), F=15511 (training set)

Table 3. Experimental and values calculated with Eqs. 5 and 6 for normal boiling points and octanol/water partition coefficients (logP)

CAS	Name	SMILES	Normal boiling points, °C			Octanol/water partition coefficient		
	Training set		DCW	Exp	Calc	DCW	Exp	Calc
106-97-8	Butane	CCCC	1.0325737	-0.5	13.821	1.0386231	2.890	2.904
109-66-0	Pentane	CCCCC	1.0436381	36.0	39.931	1.0510827	3.390	3.419
110-54-3	n-Hexane	CCCCCC	1.0548211	68.7	66.322	1.0636918	3.900	3.939
96-14-0	3-Methylpentane	CC(CC)CC	1.0534905	63.2	63.182	1.0481662	3.600	3.298
142-82-5	Heptane	CCCCCCC	1.0661239	98.5	92.995	1.0764521	4.660	4.466
589-34-4	3-Methyl-hexane	CC(CC)CCC	1.0647790	91.0	89.821	1.0607402	3.710	3.817
565-59-3	Pentane, 2,3-dimethyl-	CC(C)C(C)CC	1.0625395	89.7	84.536	1.0537095	3.630	3.527
111-65-9	n-Octane	CCCCCCCC	1.0775478	125.6	119.954	1.0893655	5.180	4.999
589-81-1	3-Methyl-heptane	CC(CC)CCCC	1.0761885	118.0	116.746	1.0734652	4.200	4.343
592-13-2	2,5-Dimethylhexane	CC(C)CCC(C)C	1.0764439	109.1	117.349	1.0666424	4.120	4.061
584-94-1	2,3-Dimethylhexane	CC(C)C(C)CCC	1.0739250	115.6	111.405	1.0663500	4.120	4.049
564-02-3	2,2,3-Trimethylpentane	CC(CC)C(C)(C)C	1.0745306	110.0	112.834	1.0675769	4.090	4.099
1067-08-9	3-Ethyl-3-methylpentane	CC(CC)(CC)CC	1.0745778	118.2	112.945	1.0715253	4.160	4.262
3221-61-2	2-Methyloctane	CC(C)CCCCCC	1.0874146	143.2	143.239	1.0892629	4.690	4.995
2216-34-4	4-Methyloctane	CC(CCC)CCCC	1.0877203	142.4	143.960	1.0863428	4.690	4.874
71-43-2	Benzene	c1ccccc1	1.0597388	80.0	77.927	1.0220093	2.130	2.219
100-41-4	Ethylbenzene	CCc1ccccc1	1.0812561	136.1	128.705	1.0490460	3.150	3.335
108-38-3	1,3-Dimethylbenzene	Cc1cccc(C)c1	1.0850204	139.1	137.589	1.0472894	3.200	3.262
103-65-1	n-Propylbenzene	CCCc1ccccc1	1.0928421	159.2	156.047	1.0616307	3.690	3.854
526-73-8	1,2,3-Trimethylbenzene	Cc1cccc(C)c1C	1.0990962	176.1	170.806	1.0618477	3.660	3.863
108-67-8	1,3,5-Trimethylbenzene	Cc1cc(C)cc(C)c1	1.0955586	164.7	162.457	1.0610796	3.420	3.831
1074-17-5	1-Methyl-2-propylbenzene	CCCc1ccccc1C	1.1070193	185.0	189.503	1.0763883	4.500	4.463
1074-55-1	1-Methyl-4-propylbenzene	CCCc1ccc(C)cc1	1.1034563	183.4	181.095	1.0756097	4.600	4.431
141-93-5	1,3-Diethylbenzene	CCc1cccc(CC)c1	1.1031794	181.1	180.442	1.0693550	4.570	4.173
99-87-6	1-Methyl-4-isopropylbenzene	Cc1ccc(cc1)C(C)C	1.1029258	177.1	179.843	1.0657560	4.100	4.024
874-41-9	1-Ethyl-2,4-dimethylbenzene	Cc1cc(C)ccc1CC	1.1069700	188.4	189.387	1.0738485	4.470	4.358
934-74-7	1-Ethyl-3,5-dimethylbenzene	Cc1cc(C)cc(CC)c1	1.1076092	183.6	190.895	1.0709299	4.550	4.238
527-53-7	1,3,4,5-Tetramethylbenzene	Cc1cc(C)c(C)c(C)c1	1.1102678	198.0	197.169	1.0673161	4.100	4.089
538-68-1	1-Phenylpentane	CCCCCc1ccccc1	1.1163880	205.4	211.612	1.0872547	4.900	4.912
2050-24-0	1-Methyl-3,5-diethylbenzene	CCc1cc(C)cc(CC)c1	1.1138940	205.0	205.727	1.0834357	4.620	4.754
2049-95-8	2-Phenyl-2-methyl butane	CCC(C)(C)c1ccccc1	1.1093225	192.4	194.939	1.0718821	4.390	4.277
120-12-7	Anthracene	c1ccc2cc3ccccc3cc2c1	1.1707520	339.9	339.905	1.0746416	4.450	4.391
91-57-6	2-Methyl-naphthalene	Cc1ccc2ccccc2c1	1.1275063	241.1	237.850	1.0615886	3.860	3.852
2765-18-6	1-Propyl-naphthalene	CCCc2cccc1ccccc12	1.1437996	274.5	276.300	1.0847127	4.700	4.807
575-41-7	1,3-Dimethyl-naphthalene	Cc2cc(C)cc1ccccc12	1.1362175	263.0	258.407	1.0734330	4.420	4.341
571-61-9	1,5-Dimethyl-naphthalene	Cc2cccc1c2cccc1C	1.1377033	265.0	261.914	1.0739181	4.380	4.361
575-37-1	1,7-Dimethyl-naphthalene	Cc1cccc2ccc(C)cc12	1.1406417	263.0	268.848	1.0758595	4.440	4.441
581-40-8	2,3-Dimethyl-naphthalene	Cc1cc2ccccc2cc1C	1.1414168	268.0	270.677	1.0741823	4.400	4.372
581-42-0	2,6-Dimethyl-naphthalene	Cc1ccc2cc(C)ccc2c1	1.1384572	262.0	263.693	1.0755671	4.310	4.429
767-60-2	3-Methyl-1H-indene	C\C2=C\Cc1ccccc12	1.1110874	198.0	199.103	1.0586191	3.800	3.730
643-58-3	2-Methyl-biphenyl	Cc2ccccc2c1ccccc1	1.1369436	255.5	260.121	1.0711996	4.300	4.249
644-08-6	4-Methyl-1,1'-biphenyl	Cc1ccc(cc1)c2ccccc2	1.1389509	267.5	264.858	1.0785178	4.630	4.551
613-33-2	4,4'-Dimethyl-1,1'-biphenyl	Cc1ccc(cc1)c2ccc(C)cc2	1.1500128	295.0	290.963	1.0927192	5.090	5.137
612-94-2	2-Phenylnaphthalene	c1cc(cc2ccccc12)c3ccccc3	1.1713540	345.5	341.325	1.0882965	4.930	4.955
92-06-8	1,3-Diphenylbenzene	c1cc(ccc1)c2cccc(c2)c3ccccc3	1.1822998	363.0	367.156	1.1029683	5.520	5.560
	Test set							
75-28-5	2-Methylpropane	CC(C)C	1.0331090	-11.7	15.084	1.0292550	2.760	2.518
78-78-4	2-Methylbutane	CC(C)CC	1.0420287	27.8	36.133	1.0385252	2.720	2.900
107-83-5	2-Methylpentane	CC(C)CCC	1.0531944	60.2	62.483	1.0509837	3.210	3.415
79-29-8	2,3-Dimethylbutane	CC(C)C(C)C	1.0534443	57.9	63.073	1.0443037	3.420	3.139
591-76-4	2-Methylhexane	CC(C)CCCC	1.0644798	90.0	89.115	1.0635915	3.710	3.935
108-08-7	2,4-Dimethylpentane	CC(C)CC(C)C	1.0650317	80.4	90.418	1.0539983	3.630	3.539
464-06-2	2,2,3-Trimethylbutane	CC(C)C(C)(C)C	1.0628399	80.8	85.245	1.0577574	3.590	3.694
592-27-8	2-Methylheptane	CC(C)CCCCC	1.0758861	117.6	116.033	1.0763507	4.200	4.462
589-53-7	4-Methylheptane	CC(CCC)CCC	1.0761885	117.7	116.746	1.0734652	4.200	4.343
589-43-5	2,4-Dimethylhexane	CC(C)CC(C)CC	1.0742269	109.5	112.117	1.0634913	4.120	3.931
565-75-3	2,3,4-Trimethylpentane	CC(C)C(C)C(C)C	1.0741798	113.5	112.006	1.0595725	4.050	3.769
540-84-1	2,2,4-Trimethylpentane	CC(C)CC(C)(C)C	1.0745306	99.2	112.834	1.0675769	4.090	4.099
111-84-2	Nonane	CCCCCCCCC	1.0890941	150.8	147.202	1.1024338	4.760	5.538
2216-33-3	3-Methyloctane	CC(CC)CCCCC	1.0877203	144.2	143.960	1.0863428	4.690	4.874
1072-05-5	2,6-Dimethylheptane	CC(C)CCCC(C)C	1.0879784	135.2	144.569	1.0794381	4.610	4.589
108-88-3	Methylbenzene	Cc1ccccc1	1.0751555	110.6	114.308	1.0369372	2.730	2.835
95-47-6	1,2-Dimethylbenzene	Cc1ccccc1C	1.0891032	144.5	147.223	1.0513516	3.120	3.430
106-42-3	1,4-Dimethylbenzene	Cc1ccc(C)cc1	1.0855978	138.3	138.951	1.0505910	3.150	3.398
611-14-3	1-Ethyl-2-methylbenzene	Cc1ccccc1CC	1.0963220	165.2	164.259	1.0598924	3.530	3.782
95-63-6	1,2,4-Trimethylbenzene	Cc1cc(C)c(C)cc1	1.0981556	169.3	168.586	1.0605884	3.630	3.811
104-51-8	n-Butylbenzene	CCCCc1ccccc1	1.1045523	183.3	183.681	1.0743663	4.380	4.380
1074-43-7	1-Methyl-3-n-propylbenzene	CCCc1cccc(C)c1	1.1028694	182.0	179.710	1.0722294	4.670	4.292
135-01-3	1,2-Diethylbenzene	CCc1ccccc1CC	1.1025427	184.0	178.939	1.0722693	3.720	4.293
105-05-5	1,4-Diethylbenzene	CCc1ccc(CC)cc1	1.1037665	183.7	181.827	1.0727261	4.580	4.312
933-98-2	1,2-Dimethyl-3-ethylbenzene	CCc1cccc(C)c1C	1.1053326	194.0	185.523	1.0742475	4.340	4.375
934-80-5	1,2-Dimethyl-4-ethylbenzene	Cc1ccc(CC)cc1C	1.1117770	189.5	200.731	1.0750837	4.500	4.409
488-23-3	1,2,3,4-Tetramethylbenzene	Cc1ccc(C)c(C)c1C	1.1118101	205.0	200.809	1.0719522	4.000	4.280
95-93-2	1,2,4,5-Tetramethylbenzene	Cc1cc(C)c(C)cc1C	1.1124017	196.8	202.205	1.0753315	4.000	4.420
2049-94-7	Isopentylbenzene	CC(C)CCc1ccccc1	1.1146664	195.0	207.549	1.0742651	4.430	4.376

Continued **Table 3.** Experimental and values calculated with Eqs. 5 and 6 for normal boiling points and octanol/water partition coefficients (logP)

CAS	Name Test set	SMILES	Normal boiling points, °C			Octanol/water partition coefficient		
			DCW	Exp	Calc	DCW	Exp	Calc
700-12-9	Pentamethylbenzene	Cc1cc(C)c(C)c(C)c1C	1.1246711	232.0	231.159	1.0821528	4.560	4.701
91-20-3	Naphthalene	c1cccc2ccccc12	1.1134715	217.9	204.730	1.0465904	3.300	3.233
85-01-8	Phenanthrene	c2cc3ccc1ccccc1c3cc2	1.1648177	340.0	325.900	1.0759686	4.460	4.446
939-27-5	2-Ethylnaphthalene	CCc1ccc2ccccc2c1	1.1339040	258.0	252.948	1.0739853	4.380	4.364
573-98-8	1,2-dimethylnaphthalene	Cc1c2ccccc2ccc1C	1.1421333	266.5	272.368	1.0763457	4.310	4.461
571-58-4	1,4-Dimethylnaphthalene	Cc1ccc(C)c2ccccc12	1.1419188	268.0	271.862	1.0735403	4.370	4.346
575-43-9	1,6-Dimethylnaphthalene	Cc2cccc1cc(C)ccc12	1.1362175	264.0	258.407	1.0734330	4.260	4.341
569-41-5	1,8-Dimethylnaphthalene	Cc1cccc2cccc(C)c12	1.1400350	270.0	267.416	1.0724785	4.260	4.302
582-16-1	2,7-Dimethylnaphthalene	Cc1cc2cc(C)ccc2cc1	1.1377430	265.0	262.007	1.0734052	4.260	4.340
496-11-7	Indan	c1cccc2CCCc12	1.1172435	177.9	213.631	1.0659955	3.180	4.034
92-52-4	Biphenyl	c1cc(ccc1)c2ccccc2	1.1226195	256.1	226.318	1.0629913	3.980	3.910
643-93-6	3-Methylbiphenyl	Cc1cccc(c1)c2ccccc2	1.1383451	272.7	263.428	1.0751284	4.300	4.411
605-39-0	2,2'-Dimethylbiphenyl	Cc2ccccc2c1ccccc1C	1.1516930	256.0	294.928	1.0860903	4.850	4.864
7383-90-6	3,4'-Dimethyl-1,1'-biphenyl	Cc1cccc(c1)c2ccc(C)cc2	1.1494012	289.0	289.519	1.0892852	4.850	4.996
605-02-7	1-Phenylnaphthalene	c1cc(c2ccccc2c1)c3ccccc3	1.1704197	334.0	339.120	1.0856553	4.930	4.846
92-94-4	1,4-Diphenylbenzene	c1cc(ccc1)c2ccc(cc2)c3ccccc3	1.1809775	376.0	364.035	1.1053511	6.030	5.659

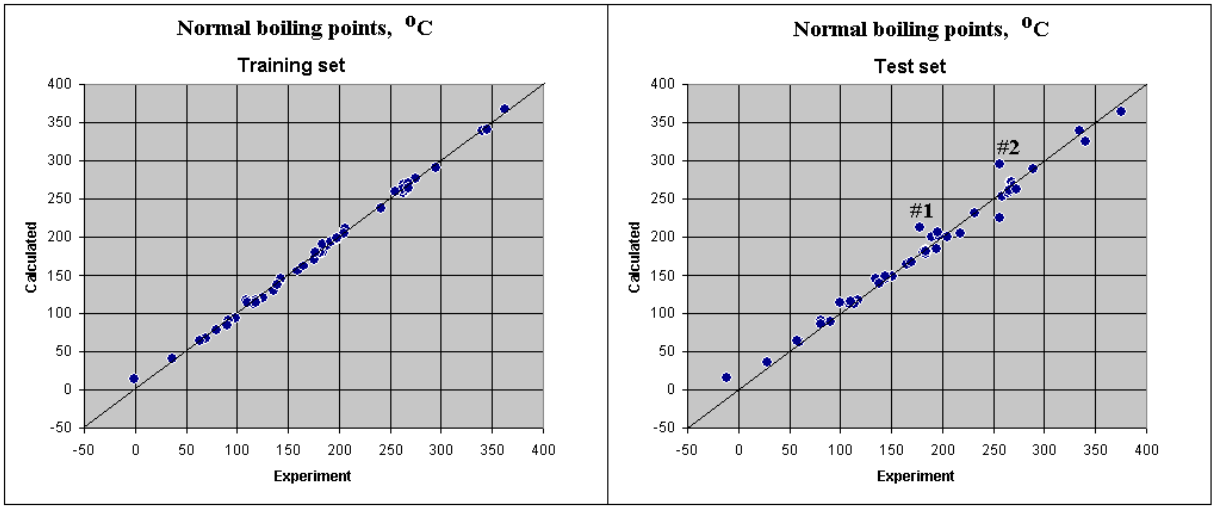


Figure 4. Noraml bolilng points: experimental versus calculated with Eq. 5 values.

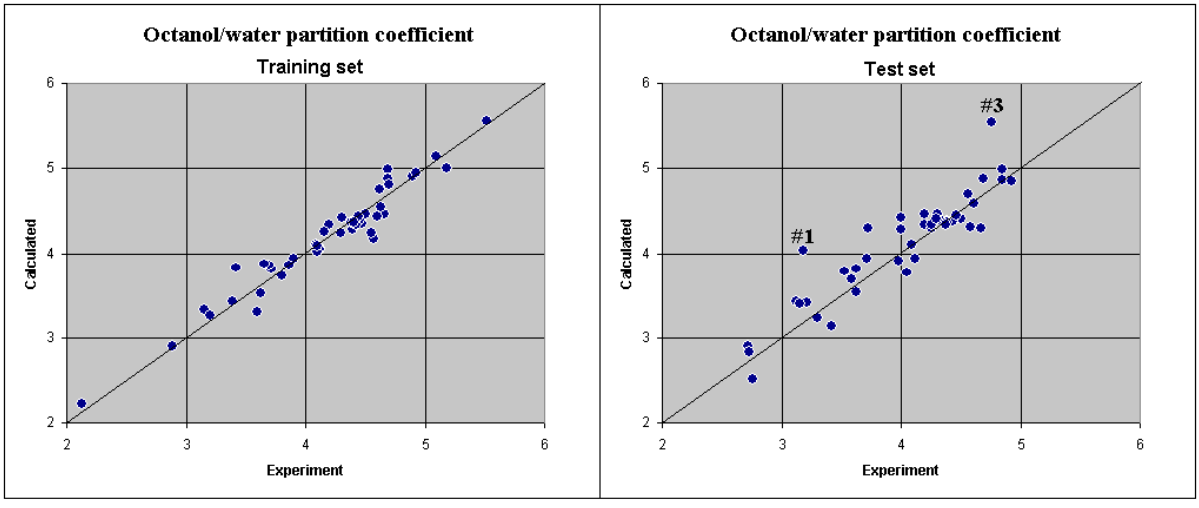
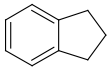
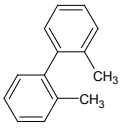
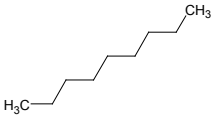


Figure 5. Octanol/water partition coefficient: experimental versus calculated with Eq. 6 values

Table 4. Structures of the outliers indicated in Figs. 4 and 5.

#1	#2	#3
		
496-11-7 Indan	605-39-0 2,2'-Dimethylbiphenyl	111-84-2 Nonane

n=45, r^2=0.982, r^2_{pred}=0.9805, R_m^2=0.9465, s=11.8, F=2394 (test set)

logP = -39.9668(±0.2317) +
+41.2769(±0.2163) * DCW (6)

n=45, r^2=0.943, q^2=0.9381, s=0.153, F=710 (training set)

n=45, r^2=0.855, r^2_{pred}=0.8350, R_m^2=0.8339, s=0.269, F=253 (test set)

Fig. 4 shows the normal boiling points model and Fig. 5 shows the model for octanol/water partition coefficient. There are three outliers (#1, #2, and #3) indicated in Table 4 and Figs. 4 and 5.

We have expected that SMILES attributes [6,7] can give a good predictive model for the normal boiling points and octanol/water partition coefficient. One can see that statistical characteristics of the models for normal boiling points and octanol/water partition coefficient are quite good.

In order to additionally check the predictability of the models which are calculated with Eqs. 5 and 6 we have calculated the R_m^2 metric [10,11]. According to [9,10] the R_m^2 should be greater than 0.5. Thus the above-mentioned criterion confirms predictability of these models.

The number of databases containing SMILES gradually increases so SMILES-based descriptors become convenient for QSPR/QSAR analyses. However, it is preferable to base this modelling on SMILES generated by the same software [9,12].

Applications of the LimN and the $\overline{\Delta P}$ need validation with other properties (activity) and other classes of compounds. If the utility of the plot of LimN versus $\overline{\Delta P}$ as an indicator of the quality of the split into training and test sets is confirmed for other properties and other substances, these plots will be useful for the rational definition of the domain applicability.

4. Conclusions

SMILES-based optimal descriptors are reasonably good predictors of the normal boiling points (°C) and octanol/water partition coefficient (logP) of acyclic and cyclic hydrocarbons. The average relative prevalence of these attributes is a tool for rational definition of the LimN before constructing these models.

Acknowledgements

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

References

[1] P.R. Duchowicz, E.A. Castro, F.M. Fernandez, M.P. Gonzalez, Chem. Phys. Lett. 412, 376 (2005)
[2] S.C. Basak, D. Mills, B.D. Gute, R. Natarajan, Top. Heterocycl. Chem. 3, 39 (2006)
[3] Q-N. Hu, Y.-Z. Liang, K.-T. Fang, J. Data Sci. 1, 361 (2003)
[4] B.D. Gute, S.C. Basak, J. Mol. Graphics Model 20, 95 (2001)
[5] D. Vidal, M. Thormann, M. Pons, J. Chem. Inf. Model. 45, 386 (2005)
[6] A.A. Toropov, A.P. Toropova, D.V. Mukhamedzhanova, I. Gutman, Indian J. Chem. Sec. A, 44, 1545 (2005)
[7] A.A. Toropov, E. Benfenati, Comput. Biol. Chem. 31, 57 (2007)
[8] D.R. Lide, CRC Handbook of Chemistry and Physics, 76th edition (CRC Press, New York, 1995/1996)
[9] ACD/ChemSketch Freeware, version 11.00 (Advanced Chemistry Development, Inc., Toronto, ON, Canada, 2007) www.acdlabs.com
[10] K. Roy, P.P. Roy, Eur. J. Med. Chem. 44, 2913 (2009)
[11] P.P. Roy, S. Paul, I. Mitra, K. Roy, Molecules 14, 1660 (2009)
[12] MDL QSAR version 2.2. (MDL Information Systems Inc., San Leandro, CA, 2003)