VERSITA

Cent. Eur. J. Chem. • 7(4) • 2009 • 846–856
DOI: 10.2478/s11532-009-0095-y

Central European Journal of **Chemistry**

QSPR modelling of the octanol/water partition coefficient of organometallic substances by optimal SMILES-based descriptors

Research Article

Andrey A. Toropov[1,2,*], Alla P. Toropova[1,2], Emilio Benfenati[2]

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

[2)] *Mario Negri Institute of Pharmacological Research, 20156 Milan, ITALY*

Received 19 January 2009; Accepted 15 May 2009

Abstract: Usually, QSPR is not used to model organometallic compounds. We have modeled the octanol/water partition coefficient for organometallic compounds of Na, K, Ca, Cu, Fe, Zn, Ni, As, and Hg by optimal descriptors calculated with simplified molecular input line entry system (SMILES) notations. The best model is characterized by the following statistics: n=54, r^2=0.9807, s=0.677, F=2636 (training set); n=26, r^2=0.9693, s=0.969, F=759 (test set). Empirical criteria for the definition of the applicability domain for these models are discussed.

Keywords: *QSPR • SMILES • Organometallic compound • Octanol/water partition coefficient • Applicability domain*

1. Introduction

Quantitative structure – property/activity relationships (QSPR) predict physicochemical parameters for substances that have not been examined experimentally [1-3]. The octanol/water partition coefficient, called K_{ow} (its logarithm is called logP), is an important parameter for many environmental and ecotoxicological properties [4-6] and has been related to properties, such as adsorption, bio concentration factor, and aquatic toxicity [7-10]. Considering the influence of the organometallic substances in biochemistry, medicine, and ecology [11], QSPR analysis could be useful for both theory and practice. However, studies of organometallic compounds are rare or absent. Classical QSPR analysis is based on molecular graphs [1,2], but an alternative to the molecular graph is the simplified molecular input line entry system (SMILES) [12-16].

The present study was designed to estimate the predictive potential of SMILES-based optimal descriptors for the QSPR of the K_{ow} of organometallic substances. The increasing number of databases with representation of molecular structures by the SMILES, and the convenience of the SMILES for chemical interpretations were reasons for the study.

2. Experimental Procedures

Experimental data for 80 organometallic compounds with their SMILES were taken from The United States Library of Medicine database [17]. Our software had prepared series of random splits into the training set and test set. We have selected for this study one split that is characterized by similar ranges of logP for the training and test sets.

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

Table 1. Statistical characteristics of the SMILES-based models for different limN. The Nact is the number of active SMILES attributes. The best statistical characteristics are indicated by bold.

		Training set, n=54			Test set, n=26		
limN	Nact	R^2	s	F	R^2	s	F
9	105	0.9930	0.407	7396	0.9440	1.21	404
10	101	0.9919	0.438	6369	0.9516	1.08	472
11	94	0.9870	0.554	3952	0.9512	1.11	468
12	93	0.9834	0.628	3074	0.9551	1.05	511
13	**89**	**0.9807**	**0.677**	**2636**	**0.9693**	**0.969**	**759**
14	88	0.9820	0.652	2845	0.9499	1.08	455
15	85	0.9827	0.641	2948	0.9290	1.25	314
16*	81	0.9818	0.656	2812	0.9281	1.26	310
17*	81	0.9808	0.675	2653	0.9233	1.27	289
18	77	0.9801	0.686	2564	0.9490	1.06	446
19	75	0.9793	0.700	2459	0.9105	1.46	244
20	69	0.9741	0.783	1956	0.9098	1.36	242
21	65	0.9735	0.792	1911	0.9084	1.36	238

*) *these models are the same in both cases, if the limN is 16 or 17, because in both these cases Nact is 81*

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

$$DCW(limN) = \Pi\, CW(^1s_k)\, \Pi\, CW(^2s_k)\, \Pi\, CW(^3s_k) \quad (1)$$

where $^1\mathbf{s}_k$, $^2\mathbf{s}_k$, and $^3\mathbf{s}_k$ are SMILES attributes of one, two, and three elements, respectively. A similar approach has been tested in QSPR analysis for the normal boiling points of organic compounds [16]. 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, @@, *etc.*); CW(x) is the correlation weight for the SMILES attribute x (*i.e.*, the $^1\mathbf{s}_k$, $^2\mathbf{s}_k$, and $^3\mathbf{s}_k$). The CWs are calculated by the Monte Carlo method optimization procedure that provides CW values, which when used in Eq. 1, gives the maximum for the correlation coefficient between the descriptor and K_{ow}.

There is an analogy between the three-level separation of the SMILES notation in the $^1\mathbf{s}_k$, $^2\mathbf{s}_k$, and $^3\mathbf{s}_k$ attributes and the extended connectivity of zero- (vertex), first- (edge), and second order (path of length 2) defined in molecular graph [2]. Instead of ')' the symbol '(' was used in the optimization procedure because both indicate the same phenomenon (branching). The similar situation is for the symbols '@' and '@@': each '@' was replaced by '@@', because these symbols are indicators of chiral centers.

The SMILES attributes of the $^2\mathbf{s}_k$ and $^3\mathbf{s}_k$ types are organized according to ASCII codes of their elements. In other words, for each AB and ABC there is only one representation in the list of attributes (*i.e.*, the presence of both ABC and CBA or both AB and BA is impossible).

For the compounds used in this study, there are 509 SMILES attributes (including all types: $^1\mathbf{s}_k$, $^2\mathbf{s}_k$, and $^3\mathbf{s}_k$ over both the training and test sets). Some appear many times in the training set. There are also some that are rare in the training set, or even absent. These SMILES attributes can lead to over-fitting. The LimN index has been suggested as a tool for distinguishing the SMILES attributes according to their numbers in the training set [18]. All SMILES attributes whose presence in the training is less than LimN are assigned with correlation weights equal to 1. These attributes are blocked and do not influence modelling. The number of active (non blocked) attributes is also a mathematical function of the LimN. Only active attributes take part in the modelling, which is an optimization by the Monte Carlo method [19]. The target function of the optimization is a correlation coefficient between the toxicity and the DCW that is calculated with Eq. 1. A maximal value should be obtained for the correlation coefficient over the training set. However, good statistics for the training set may be accompanied by poor statistics for the test set.

3. Results and Discussion

Table 1 contains the statistical parameters of the models based on optimal descriptors obtained with different limN. Examination of the models from limN=9 to 21 showed

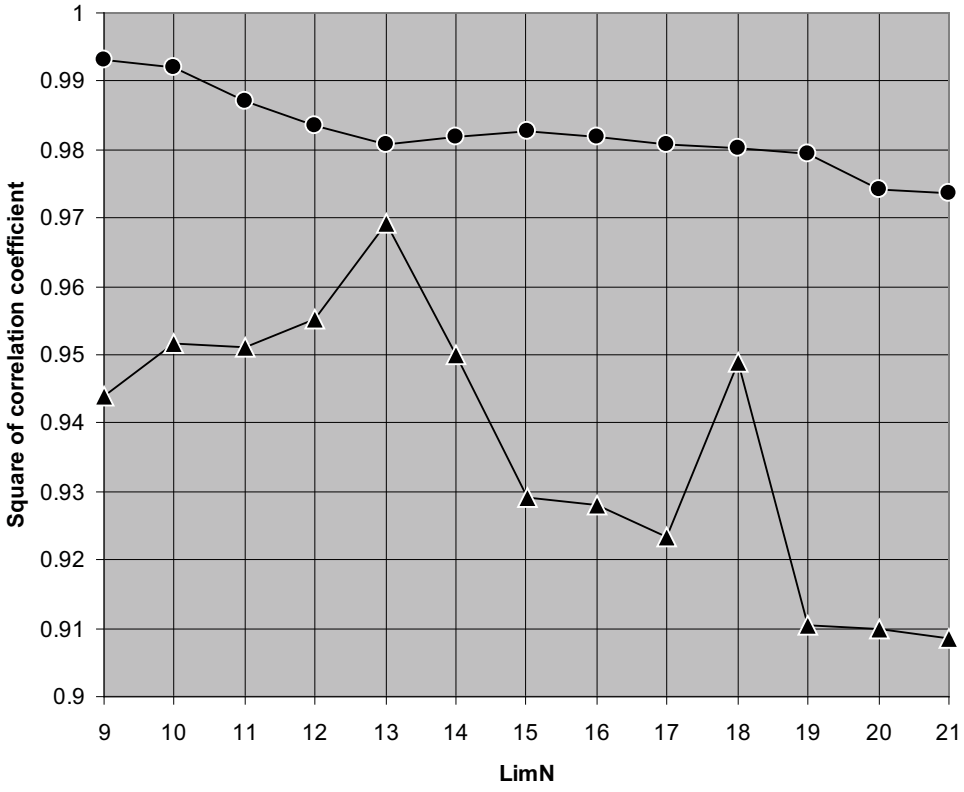

Figure 1. The square of correlation coefficient versus the limN for the training (circles) and test (triangles) sets.

the best results, also considering the external test set when limN=13. Fig. 1 shows the influence of the limN. SMILES attributes that are relatively rare do not produce robust models, and can only increase overfitting.

An empirical formulation of the applicability domain of the models might be the following: 1. the training set should contain as many active SMILES attributes as possible; 2. the test set should be selected so as to provide the largest number of active attributes.

Table 2 compares the statistical characteristics of the models obtained with limN=13 in three probes using the Monte Carlo method. The correlation weights of the active (non blocked) SMILES attributes are listed in Table 3. Only Na is presented in this list, because all other metals have scarce influence on the correlation coefficient between octanol/water partition coefficient and the DCW(limN). An example of the DCW(13) calculation is presented in Table 4.

The best model obtained with limN=13 is:

$$\log P = -54.3191(\pm 0.1207) + 54.3596(\pm 0.1166) * DCW(13) \quad (2)$$

$n=54$, $r^2=0.9807$, $s=0.677$, $F=2636$ (training set)
$n=26$, $r^2=0.9693$, $s=0.969$, $F=759$ (test set)

Table 5 contains the experimental and calculated data on the octanol/water partition coefficient. Graphically, the model is shown in Figs. 2 and 3 for the training and test sets, respectively. The range of K_{ow} is about 30 log units.

The influence of molecular fragments can be detected via SMILES attributes (Table 3). The number of attributes in the training set (as well as in the test set) is an important characteristic of the attribute. If the numerical values of the correlation weight of the given attribute in three probes of the Monte Carlo optimization are more than 1, the attribute is a promoter of the increase of logP. *Vice versa*, if numerical values of the correlation weight are less than 1, the attribute is a promoter of the decrease of logP. However, the number of the attributes in the training set (as well as in the test set) should be taken into account. The abovementioned data are collected in Table 3. If a SMILES attribute has both more than 1 and less than 1 values of its correlation weight in three probes of the Monte Carlo optimization, then the role of the attribute (*i.e.*, is the attribute a promoter of increase/decrease of the logP?) cannot be defined. For instance, branching and the presence of double bonds, *i.e.*, $^3\mathbf{s}_k$ ='___(_____' are promoter of logP increase, whereas, combining oxygen with double bonds, *i.e.*,

Table 2. Statistical characteristics of the models with limN=13 over three probes of the optimization

Training set, n=54			Test set, n=26		
R^2	s	F	R^2	s	F
0.9807	0.677	2636	0.9693	0.969	759
0.9806	0.677	2633	0.9691	0.907	752
0.9803	0.682	2593	0.9695	0.986	763

Table 3. Numerical data on the correlation weights of the active SMILES attributes for models obtained in three probes of the Monte Carlo optimization with limN=13. Blocked SMILES attributes are omitted.

SAk	CW(SAk) Probe 1	CW(SAk) Probe 2	CW(SAk) Probe 3	ID	Number	N(Train)	N(Test)
^1S_k							
+2	1.0240178	1.0172526	1.0046796	2	1	19	5
(	1.0041864	1.0028139	0.9962731	3	2	528	198
+	0.9924988	0.9970115	0.9953166	5	3	43	23
.	0.9919806	0.9945558	0.9935212	7	4	75	33
1	1.0110613	0.9987069	1.0003295	9	5	60	32
2	0.9984910	0.9996822	0.9994069	10	6	23	14
@@	1.0012352	0.9973549	1.0005610	14	7	53	14
=	1.0028243	0.9977411	0.9964018	15	8	128	51
C	1.0029450	1.0074530	1.0003795	17	9	358	144
H	1.0002009	0.9948358	1.0007533	24	10	35	12
O-	1.0100773	0.9916586	0.9952393	28	11	72	28
N	1.0024840	1.0008011	0.9949318	31	12	29	13
Na	0.9945934	1.0000607	0.9997693	32	13	34	11
O	0.9917818	0.9950840	1.0007376	34	14	174	62
S	1.0032377	0.9990276	1.0070088	36	15	18	11
[	0.9986365	0.9942015	0.9968391	38	16	404	152
c	0.9979226	0.9989182	1.0018035	41	17	175	102
^2S_k							
(__(	1.0016391	1.0064491	1.0061856	46	18	49	22
1__(	0.9951599	1.0012260	0.9994873	47	19	27	13
=__(	1.0021010	1.0062788	1.0055619	54	20	113	40
C__(	1.0004813	0.9990165	1.0035785	57	21	231	92
C__.	0.9921109	0.9954258	1.0047288	58	22	19	9
C__1	0.9782890	0.9975258	0.9983715	60	23	16	8
C__@@	1.0035064	1.0016998	1.0021164	64	24	51	12
C__=	1.0111653	1.0013892	1.0086411	65	25	15	9
C__C	1.0047468	1.0004092	0.9983726	66	26	122	53
H__@@	1.0003536	1.0004017	0.9975005	80	27	31	10
N__(	0.9907976	0.9917459	0.9955332	88	28	20	7
Na__+	0.9981391	1.0000692	0.9998579	97	29	34	11
O__(	0.9995603	0.9991684	0.9974302	99	30	179	66
O__.	0.9880339	0.9987916	0.9953709	100	31	16	6
O__=	0.9948522	0.9915269	0.9952596	103	32	113	42
O__C	1.0095190	1.0023973	1.0039611	104	33	17	5
S__(	1.0126817	1.0113263	1.0042925	107	34	29	14
[__+2	1.0146563	1.0104951	1.0146892	111	35	19	5
[__(	0.9962466	0.9984294	1.0004028	112	36	198	64
[__+	0.9928037	0.9872750	0.9906049	114	37	43	23
[__.	0.9937615	0.9988969	1.0034552	116	38	108	42
[__@@	0.9963746	0.9994832	0.9999693	121	39	22	4
[__C	1.0012259	1.0057791	1.0027404	122	40	71	22
[__H	1.0021909	1.0029463	1.0006518	128	41	31	10
[__O-	0.9938148	0.9966070	0.9950979	131	42	144	56

Continued **Table 3.** Numerical data on the correlation weights of the active SMILES attributes for models obtained in three probes of the Monte Carlo optimization with limN=13. Blocked SMILES attributes are omitted.

SAk ^1S_k	CW(SAk) Probe 1	CW(SAk) Probe 2	CW(SAk) Probe 3	ID	Number	N(Train)	N(Test)
[___Na_____	1.0031930	1.0020253	1.0033008	135	43	34	11
c___(_____	1.0018450	1.0006803	1.0044311	146	44	105	40
c___1_____	1.0002416	0.9996090	0.9973689	148	45	46	27
c___2_____	1.0035927	1.0011144	1.0002183	149	46	17	15
c___c_____	1.0052260	1.0052598	1.0032022	157	47	68	50
^3S_k				167	48	43	23
(___C___(___	0.9998794	0.9962172	0.9984299	167	48	43	23
(___O___(___	0.9956881	0.9965249	1.0007529	173	49	31	10
(___c___(___	1.0073026	0.9991919	1.0032067	176	50	19	4
.___C___(___	1.0017003	0.9994211	0.9960860	177	51	15	5
.___[___+___	0.9924408	1.0004596	0.9993427	183	52	13	2
@@__[___(___	0.9896065	0.9999384	0.9932741	212	53	15	4
=___(___(___	1.0103258	1.0041371	1.0008209	214	54	26	13
=___O___(___	0.9968989	0.9938565	1.0019057	222	55	97	37
C___(___(___	0.9960047	0.9946459	1.0005518	226	56	15	8
C___(___C___	0.9998634	0.9994869	1.0024274	227	57	34	16
C___(___=___	0.9945712	0.9967404	1.0013041	232	58	47	19
C___C___(___	1.0021355	0.9992853	1.0033433	247	59	72	18
C___C___C___	1.0003134	0.9987774	1.0064089	253	60	79	40
C___[___(___	0.9981182	0.9987845	1.0005862	266	61	35	8
H___@@__C___	0.9990672	0.9991666	1.0051060	285	62	31	10
H___[___(___	0.9932698	1.0011924	0.9977167	286	63	28	8
O-__[___.___	1.0026439	1.0041633	0.9958940	295	64	30	9
O-__[___(___	0.9983937	1.0024776	0.9978153	297	65	107	41
Na__[___.___	1.0003515	1.0040320	1.0013303	319	66	34	11
O___(___(___	0.9980004	0.9992450	1.0025267	322	67	29	13
O___(___O___	0.9977308	0.9967112	0.9988725	323	68	20	2
O___(___C___	0.9850406	0.9935842	0.9887073	325	69	21	9
O___=___(___	1.0010507	1.0019137	1.0029979	332	70	108	38
O___C___(___	0.9866258	0.9924655	0.9941355	335	71	14	1
[___(___[___	0.9958202	0.9995613	0.9974446	369	72	19	3
[___(___(___	0.9845943	0.9952417	0.9961164	370	73	19	6
[___(___C___	1.0047928	1.0022206	1.0046575	371	74	49	16
[___(___O___	0.9999981	1.0009604	0.9982204	373	75	56	25
[___(___=___	1.0002679	1.0049609	1.0084127	375	76	14	0
[___+___Na__	0.9994835	0.9947135	0.9979094	377	77	34	11
[___.___[___	0.9878934	0.9979178	0.9978481	384	78	39	11
[___.___O___	1.0085172	0.9989691	0.9979203	386	79	14	6
[___@@__C___	1.0036287	1.0025658	1.0008146	397	80	20	2
[___C___@@__	1.0010546	0.9966873	1.0037517	402	81	51	12
[___H___@@__	1.0021551	1.0049360	0.9995110	415	82	31	10
[___O-__[___	0.9955031	1.0032160	1.0041506	418	83	72	28
[___Na__+___	0.9960605	0.9982729	0.9964395	424	84	34	11
c___(___c___	1.0039070	1.0094166	0.9999902	449	85	28	7
c___1___(___	1.0053052	1.0053827	1.0042559	463	86	20	13
c___c___(___	0.9987315	1.0023491	1.0043696	484	87	45	26
c___c___1___	0.9871337	0.9956356	0.9987647	485	88	18	13
c___c___c___	0.9966911	1.0017692	0.9972101	488	89	28	24

*) *The ID is the number in the total list of the SMILES attributes;*
**) *N(Train) and N(test) are numbers of SMILES attributes (^1s_k, ^2s_k and ^3s_k) in the training and test sets, respectively.*

Table 4. Example of the DCW(13) calculation: SMILES="C(=O)([O-])C.C(C)(=O)[O-].[Ni+2]"; CAS= 373-02-4; DCW(13)= 0.9668803

SAk	CW(SAk) Probe 1	ID	N(Train)	N(Test)
C_____	1.0029450	17	358	144
(_____	1.0041864	3	528	198
=_____	1.0028243	15	128	51
O_____	0.9917818	34	174	62
(_____	1.0041864	3	528	198
(_____	1.0041864	3	528	198
[_____	0.9986365	38	404	152
O-_____	1.0100773	28	72	28
[_____	0.9986365	38	404	152
(_____	1.0041864	3	528	198
C_____	1.0029450	17	358	144
._____	0.9919806	7	75	33
C_____	1.0029450	17	358	144
(_____	1.0041864	3	528	198
C_____	1.0029450	17	358	144
(_____	1.0041864	3	528	198
(_____	1.0041864	3	528	198
=_____	1.0028243	15	128	51
O_____	0.9917818	34	174	62
(_____	1.0041864	3	528	198
[_____	0.9986365	38	404	152
O-_____	1.0100773	28	72	28
[_____	0.9986365	38	404	152
._____	0.9919806	7	75	33
[_____	0.9986365	38	404	152
Ni_____	1.0	33	0	1
+2_____	1.0240178	2	19	5
[_____	0.9986365	38	404	152
C___(_____	1.0004813	57	231	92
=___(_____	1.0021010	54	113	40
O___=_____	0.9948522	103	113	42
O___(_____	0.9995603	99	179	66
(___(_____	1.0016391	46	49	22
[___(_____	0.9962466	112	198	64
[___O-_____	0.9938148	131	144	56
[___O-_____	0.9938148	131	144	56
[___(_____	0.9962466	112	198	64
C___(_____	1.0004813	57	231	92
C___._____	0.9921109	58	19	9
C___._____	0.9921109	58	19	9
C___(_____	1.0004813	57	231	92
C___(_____	1.0004813	57	231	92
C___(_____	1.0004813	57	231	92
(___(_____	1.0016391	46	49	22
=___(_____	1.0021010	54	113	40
O___=_____	0.9948522	103	113	42
O___(_____	0.9995603	99	179	66
[___(_____	0.9962466	112	198	64
[___O-_____	0.9938148	131	144	56
[___O-_____	0.9938148	131	144	56
[___._____	0.9937615	116	108	42
[___._____	0.9937615	116	108	42
[___Ni_____	1.0	137	0	1

Continued **Table 4.** Example of the DCW(13) calculation: SMILES="C(=O)([O-])C.C(C)(=O)[O-].[Ni+2]"; CAS= 373-02-4; DCW(13)= 0.9668803

SAk	CW(SAk) Probe 1	ID	N(Train)	N(Test)
Ni__+2_____	1.0	98	0	1
[___+2_____	1.0146563	111	19	5
C___(___=___	0.9945712	232	47	19
O___=___(___	1.0010507	332	108	38
=___O___(___	0.9968989	222	97	37
O___(___(___	0.9980004	322	29	13
[___(___(___	0.9845943	370	19	6
O-__[___(___	0.9983937	297	107	41
[___O-__[___	0.9955031	418	72	28
O-__[___(___	0.9983937	297	107	41
[___(___C___	1.0047928	371	49	16
.___C___(___	1.0017003	177	15	5
C___.___C___	1.0	235	3	1
.___C___(___	1.0017003	177	15	5
C___(___C___	0.9998634	227	34	16
(___C___(___	0.9998794	167	43	23
C___(___(___	0.9960047	226	15	8
=___(___(___	1.0103258	214	26	13
O___=___(___	1.0010507	332	108	38
=___O___(___	0.9968989	222	97	37
[___(___O___	0.9999981	373	56	25
O-__[___(___	0.9983937	297	107	41
[___O-__[___	0.9955031	418	72	28
O-__[___.___	1.0026439	295	30	9
[___.___[___	0.9878934	384	39	11
Ni__[___.___	1.0	340	0	1
[___Ni__+2__	1.0	427	0	1
[___+2__Ni__	1.0	364	0	1

Table 5. Experimental and calculated data on the octanol/water partition coefficient (logP) for the model with LimN=13 (Probe 1)

CAS	SMILES	DCW(13)	logP Expr	logP Calc	Expr-Calc
	Training set				
137-30-4	C(=S)(N(C)C)[S-].[S-]C(N(C)C)=S.[Zn+2]	1.0370702	1.230	2.056	-0.826
557-05-1	[Zn+2].C(=O)([O-])CCCCCCCCCCCCCCCCCC.C(CCCCCCCCCCCCCCCCC)(=O)[O-]	1.2658676	14.440	14.493	-0.053
557-34-6	C(=O)([O-])C.C(C)(=O)[O-].[Zn+2]	0.9668803	-1.280	-1.760	0.480
54-47-7	c1(c(c(c(C)nc1)O)C=O)COP(O)(O)=O	1.0189972	0.370	1.073	-0.703
156-62-7	C(#N)[NH2-2].[Ca+2]	1.0025053	-0.200	0.177	-0.377
544-17-2	C(=O)[O-].C(=O)[O-].[Ca+2]	0.9495837	-2.470	-2.700	0.230
591-64-0	C(CCC(=O)[O-])(C)=O.C(CCC(=O)[O-])(C)=O.[Ca+2]	0.9686797	-2.530	-1.662	-0.868
1592-23-0	C(CCCCCCCCCCC)CCCCCC(=O)[O-].C(CCCCCCCCCCC)CCCCCC(=O)[O-].[Ca+2]	1.2655982	14.340	14.478	-0.138
4075-81-4	C([O-])(CC)=O.C(CC)(=O)[O-].[Ca+2]	0.9866242	-0.400	-0.687	0.287
299-27-4	O[C@@H]([C@@H]([C@@H](CO)O)O)[C@H](C(=O)[O-])O.[K+]	0.9003692	-5.990	-5.375	-0.615
562-54-9	S([O-])(OC)(=O)=O.[K+]	0.9507753	-3.710	-2.635	-1.075
578-36-9	c1(c(cccc1)O)C(=O)[O-].[K+]	0.9452495	-1.490	-2.936	1.446
583-52-8	C(C(=O)[O-])(=O)[O-].[K+].[K+]	0.8696596	-7.000	-7.045	0.045

Continued **Table 5.** Experimental and calculated data on the octanol/water partition coefficient (logP) for the model with LimN=13 (Probe 1)

CAS	SMILES	DCW(13)	logP Expr	logP Calc	Expr-Calc
	Training set				
590-29-4	C(=O)[O-].[K+]	0.9349778	-4.270	-3.494	-0.776
593-29-3	C(CCCCCCCCCCC)CCCCCC(=O)[O-].[K+]	1.0794005	4.130	4.357	-0.227
866-83-1	[K+].OC(=O)C[C@@](O)(CC([O-])=O)C(O)=O	0.8874874	-5.780	-6.076	0.296
53-10-1	C1[C@H]2[C@H]3[C@@H]([C@@]4([C@@H](CC(=O)CC4)CC3)C)CC[C@@]2([C@H](C(COC(CCC(=O)[O-])=O)=O)C1)C.[Na+]	0.9939584	-0.560	-0.288	-0.272
55-03-8	c1(Oc2c(cc(C[C@@H](C([O-])=O)N)cc2I)I)cc(c(O)c(c1)I)I.[Na+]	1.0389855	2.340	2.160	0.180
58-71-9	N12[C@@H]([C@@H](NC(Cc3cccs3)=O)C2=O)SCC(=C1C(=O)[O-])COC(C)=O.[Na+]	0.9351755	-2.200	-3.483	1.283
58-90-2	c1(c(c(cc(c1Cl)Cl)Cl)Cl)O)Cl	1.0749152	4.450	4.113	0.337
62-33-9	C1[N@@](CC[N@@](CC(=O)O[Ca]OC1=O)CC(=O)[O-])CC(=O)[O-].[Na+].[Na+]	0.8354716	-10.420	-8.903	-1.517
62-74-8	C(CF)(=O)[O-].[Na+]	0.9546160	-3.780	-2.427	-1.353
62-76-0	C(C(=O)[O-])(=O)[O-].[Na+].[Na+]	0.8554990	-7.000	-7.815	0.815
64-02-8	N(CCN(CC(=O)[O-])CC(=O)[O-])(CC(=O)[O-])CC(=O)[O-].[Na+].[Na+].[Na+].[Na+]	0.7539595	-13.170	-13.334	0.164
69-57-8	N12[C@@H]([C@@H](NC(Cc3ccccc3)=O)C2=O)SC([C@@H]1C(=O)[O-])(C)C.[Na+]	0.9435785	-3.010	-3.027	0.017
71-67-0	C1(c2c(c(c(Br)c(c2Br)Br)Br)C(O1)=O)(c1cc(c(O)cc1)S(=O)(=O)[O-])c1cc(c(O)cc1)S(=O)(=O)[O-].[Na+].[Na+]	1.0074801	-0.070	0.447	-0.517
71-73-8	C1([C@@](C(N=C(N1)[S-])=O)([C@@H](CCC)C)CC)=O.[Na+]	0.9928877	0.360	-0.346	0.706
72-17-3	C([C@@H](C)O)(=O)[O-].[Na+]	0.9384486	-4.770	-3.305	-1.465
72-57-1	c12c(cc(S(=O)(=O)[O-])c(c1O)\N=N\c1c(cc(c3cc(c(\N=N\c4c(cc5cc(S(=O)(=O)[O-])cc(c5c4O)N)S(=O)(=O)[O-])cc3)C)cc1)C)cc(S(=O)(=O)[O-])cc2N.[Na+].[Na+].[Na+].[Na+]	0.9877562	-0.120	-0.625	0.505
124-41-4	C[O-].[Na+]	0.9405890	-3.180	-3.189	0.009
124-65-2	[As](C)(C)(=O)[O-].[Na+]	0.9518830	-2.180	-2.575	0.395
125-02-0	C1[C@@]2([C@@H]([C@@H]3[C@@H]([C@@]4(C(=CC(=O)C=C4)CC3)C)[C@@H]1O)CC[C@@]2(C(COP(=O)([O-])[O-])=O)O)C.[Na+].[Na+]	0.9144936	-4.840	-4.608	-0.232
125-44-0	C1(C(NC(=O)[NH-]C1=O)=O)(\C(=C\CC)C)CC.[Na+]	0.9761212	-1.030	-1.258	0.228
127-52-6	c1(S([N-]Cl)(=O)=O)ccccc1.[Na+]	1.0084385	0.290	0.499	-0.209
127-65-1	S(c1ccc(C)cc1)([N-]Cl)(=O)=O.[Na+]	0.9852746	0.840	-0.760	1.600
127-68-4	c1(cc(ccc1)[N+](=O)[O-])S(=O)(=O)[O-].[Na+]	0.9626266	-2.610	-1.991	-0.619
127-85-5	[As](c1ccc(N)cc1)(O)(=O)[O-].[Na+]	0.9280616	-3.840	-3.870	0.030
141-01-5	[Fe+2].[O-]C(\C=C\C(=O)[O-])=O	1.0075806	0.620	0.453	0.167
149-45-1	c1(cc(cc(c1O)O)S(=O)(=O)[O-])S(=O)(=O)[O-].[Na+].[Na+]	0.9440706	-3.070	-3.000	-0.070
299-29-6	O[C@@H]([C@H](C(O[Fe]OC([C@@H]([C@@H]([C@H]([C@H](CO)O)O)O)O)=O)=O)O)[C@H]([C@H](CO)O)O	0.8383589	-7.710	-8.746	1.036
516-03-0	C(C([O-])=O)([O-])=O.[Fe+2]	0.9803043	-1.170	-1.030	-0.140
577-11-7	C([C@@H](CC(OC[C@@H](CCCC)CC)=O)S(=O)(=O)[O-])(OC[C@@H](CCCC)CC)=O.[Na+]	1.1037914	6.100	5.683	0.417
1271-19-8	C=1C=C[CH-]C1.C=1C=C[CH-]C1.[Ti+2](Cl)Cl	1.0786100	4.640	4.314	0.326
5905-52-2	[O-]C(=O)[C@@H](C)O.[Fe+2].[O-]C([C@@H](C)O)=O	0.9482484	-3.270	-2.773	-0.497
142-71-2	C(=O)([O-])C.C(C)(=O)[O-].[Cu+2]	0.9668803	-1.380	-1.760	0.380
147-14-8	n12c3c4ccccc4c1nc1c4c(cccc4)c(n1)nc1n([Cu]2)c(c2ccccc12)nc1nc(n3)c2c1cccc2	1.1251144	6.600	6.842	-0.242
527-09-3	O[C@H]([C@H](C([O-])=O)O)[C@@H]([C@@H](CO)O)O.[O-]C([C@@H]([C@H]([C@@H]([C@@H](CO)O)O)O)O)=O.[Cu+2]	0.8694202	-7.510	-7.058	-0.452

Continued **Table 5.** Experimental and calculated data on the octanol/water partition coefficient (logP) for the model with LimN=13 (Probe 1)

CAS	SMILES	DCW(13)	logP Expr	logP Calc	Expr-Calc
	Training set				
540-16-9	C(=O)(CCC)[O-].C([O-])(=O)CCC.[Cu+2]	1.0045106	0.590	0.286	0.304
544-19-4	C(=O)[O-].C(=O)[O-].[Cu+2]	0.9495837	-2.470	-2.700	0.230
598-54-9	CC([O-])=O.[Cu+]	0.9641488	-0.970	-1.908	0.938
814-91-5	C(C([O-])=O)([O-])=O.[Cu+2]	0.9803043	-0.970	-1.030	0.060
815-82-7	C([C@@H]([C@H](C(=O)[O-])O)O)(=O)[O-].[Cu+2]	0.9609423	-1.460	-2.083	0.623
5328-04-1	c1(c2ccccc2)c(c(ccc1)C([O-])=O)[O-].[Cu+2]	1.0457389	3.020	2.527	0.493
547-58-0	c1(ccc(S([O-])(=O)=O)cc1)\N=N\c1ccc(N(C)C)cc1.[Na+]	1.0016240	-0.660	0.129	-0.789
	Test set				
373-02-4	C(=O)([O-])C.C(C)(=O)[O-].[Ni+2]	0.9668803	-1.380	-1.760	0.380
127-82-2	S(c1ccc(O)cc1)(=O)(=O)[O-].S(c1ccc(O)cc1)(=O)(=O)[O-].[Zn+2]	1.0359229	1.370	1.993	-0.623
62-33-9	C1[N@@](CC[N@@](CC(=O)O[Ca]OC1=O)CC(=O)[O-])CC(=O)[O-].[Na+].[Na+]	0.8354716	-10.420	-8.903	-1.517
100-67-4	c1cccc(c1)O.[K+]	0.9776329	-1.170	-1.175	0.005
113-98-4	N12[C@@H]([C@@H](NC(Cc3ccccc3)=O)C2=O)SC([C@@H]1C(=O)[O-])(C)C.[K+]	0.9513557	-3.010	-2.604	-0.406
127-08-2	C(C)(=O)[O-].[K+]	0.9589845	-3.720	-2.189	-1.531
127-95-7	C(C(=O)[O-])(O)=O.[K+]	0.9260848	-4.960	-3.978	-0.982
132-93-4	N12[C@@H]([C@@H](NC([C@H](Oc3ccccc3)C)=O)C2=O)SC([C@@H]1C(=O)[O-])(C)C.[K+]	0.9389365	-2.570	-3.279	0.709
138-84-1	c1(C(=O)[O-])ccc(N)cc1.[K+]	0.9567635	-2.790	-2.310	-0.480
140-89-6	O(C(=S)[S-])CC.[K+]	0.9802876	-2.240	-1.031	-1.209
143-19-1	C(C\C=C/CCCCCCCC)CCCCCC(=O)[O-].[Na+]	1.0789027	3.920	4.330	-0.410
868-14-4	C([C@@H]([C@H](C(=O)[O-])O)O)(O)=O.[K+]	0.9159170	-5.140	-4.530	-0.610
877-24-7	c1(c(cccc1)C(=O)[O-])C(O)=O.[K+]	0.9608295	-2.730	-2.089	-0.641
1319-69-3	[O-]P(=O)(OC[C@@H](O)CO)[O-].[K+].[K+]	0.8603074	-6.960	-7.553	0.593
1929-86-8	c1(c(cc(Cl)cc1)C)O[C@@H](C(=O)[O-])C.[K+]	0.9894361	-0.870	-0.534	-0.336
54-64-8	c1(c(cccc1)C(=O)[O-])S[Hg]CC.[Na+]	0.9915041	-1.880	-0.421	-1.459
126-31-8	S(Cl)(=O)(=O)[O-].[Na+]	0.9737289	-3.450	-1.388	-2.062
127-09-3	C(C)(=O)[O-].[Na+]	0.9511449	-3.720	-2.615	-1.105
127-20-8	C(C(=O)[O-])(C)(Cl)Cl.[Na+]	0.9485950	-2.130	-2.754	0.624
127-39-9	C([C@@H](CC(OCC(C)C)=O)S(=O)(=O)[O-])(OCC(C)C)=O.[Na+]	1.0258125	0.020	1.444	-1.424
128-04-1	C(N(C)C)(=S)[S-].[Na+]	0.9837742	-2.410	-0.842	-1.568
102-54-5	C1=C[CH-]C=C1.C=1C=C[CH-]C1.[Fe+2]	1.0736653	3.280	4.045	-0.765
660-60-6	[O-]C(=O)CCCCCCCCCCCCCCCCC.[Cu+2].O=C(CCCCCCCCCCCCCCCCC)[O-]	1.2687701	14.340	14.651	-0.311
10380-28-6	c1c2cccc([O-])c2ncc1.[Cu+2].c1c2c(ncc1)c(ccc2)[O-]	1.0417877	2.460	2.312	0.148
547-57-9	c1(\N=N\c2c(cc(O)cc2)O)ccc(S(=O)(=O)[O-])cc1.[Na+]	1.0117801	0.690	0.681	0.009
587-98-4	c1(Nc2ccccc2)ccc(\N=N\c2cc(ccc2)S(=O)(=O)[O-])cc1.[Na+]	1.0295749	2.250	1.648	0.602

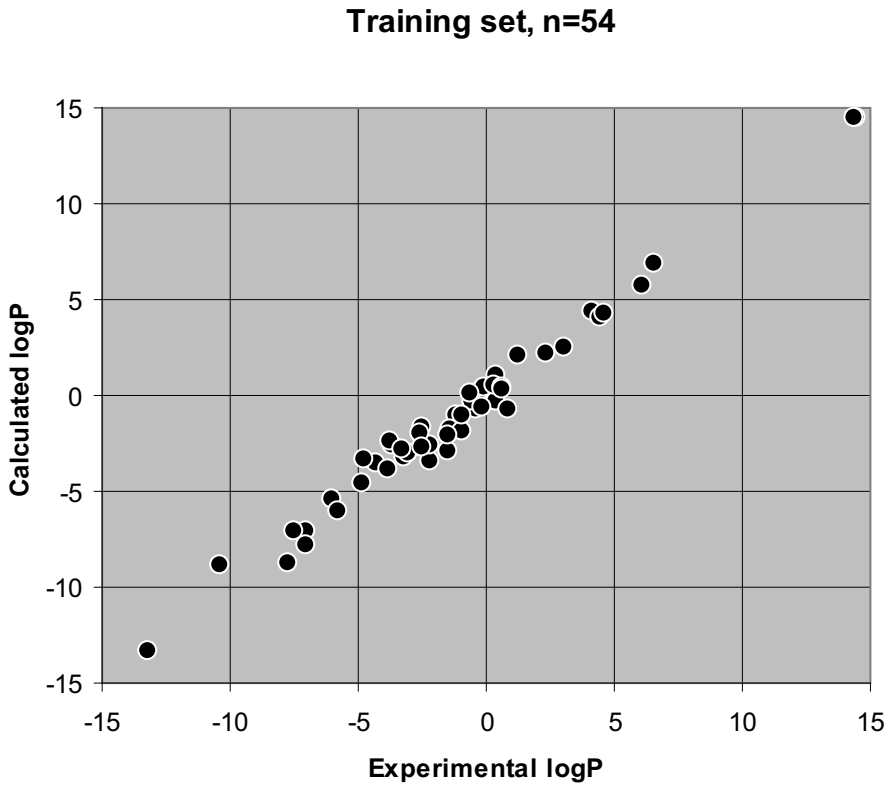

Figure 2. Experimental versus calculated octanol/water partition coefficients for the training set (limN=13, probe 1).

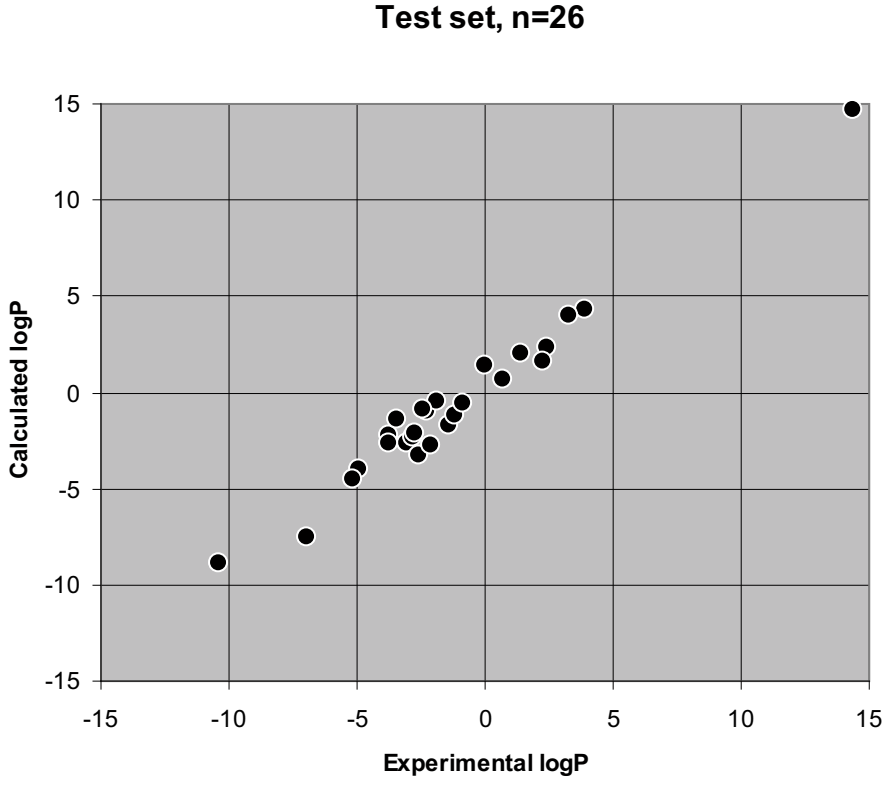

Figure 3. Experimental versus calculated octanol/water partition coefficients for the test set (limN=13, probe 1).

^3s_k ='O___=_____' is a promoter of logP decrease (Table 3).

Our QSPR model shows that it is possible to predict the K_{ow} of organometallic compounds [8]. A number of commercial programs exist but cannot be applied to this class, so our study covers a current gap.

4. Conclusions

SMILES-based optimal descriptors are reasonable predictors for the octanol/water partition coefficient of organometallic compounds. The statistical quality of the models (for the test set) is best when the limN index is equal to 13.

Acknowledgements

The authors thank the Marie Curie fellowship (the contract ID 39036, CHEMPREDICT) and the Federchemica AISPEC for financial support.

References

[1] S.C. Basak, D. Mills, B.D. Gute, R. Natarajan, Top Heterocycl. Chem. 3, 39 (2006)
[2] P.R. Duchowicz, E.A. Castro, A.A. Toropov, E. Benfenati, Top. Heterocycl. Chem. 3, 1 (2006)
[3] A.A. Toropov, E. Benfenati, Current Drug Discovery Technologies 4, 77 (2007)
[4] P.R. Duchowicz, M.G. Vitale, E.A. Castro, J. Math. Chem. 44, 541 (2008)
[5] T. Puzyn, N. Suzuki, M. Haranczyk, Environ. Sci. Tech. 42, 5189 (2008)
[6] P.C.M. van Noort, Chemosphere 74, 1024 (2009)
[7] J. Padmanabhan, R. Parthasarathi, V. Subramanian, P.K. Chattaraj, Bioorg. Med. Chem. 14, 1021 (2006)
[8] E.A. Tehrany, F. Fournier, S. Desobry, J. Food Eng. 64, 315 (2004)
[9] F.A. de Lima Ribeiro, M.M.C. Ferriera, J. Mol. Struct. (Theochem.) 663, 109 (2003)
[10] M.M.C. Ferreira, Chemosphere 44, 125 (2001)
[11] V.V. Pavlishchuk, Coordin. Chem. Rev. 181, 1 (1999)
[12] D. Weininger, J. Chem. Inf. Comput. Sci. 28, 31 (1988)
[13] D. Weininger, A. Weininger, J.L. Weininger, J. Chem. Inf. Comput. Sci. 29, 97 (1989)
[14] D. Weininger, J. Chem. Inf. Comput. Sci. 30, 237 (1990)
[15] D. Vidal, M. Thormann, M. Pons, J. Chem. Inf. Model. 46, 836 (2006)
[16] A.A.Toropov, A.P. Toropova, D.V. Mukhamedzhanova, I. Gutman, Indian J. Chem. A 44, 1545 (2005)
[17] http://toxnet.nlm.nih.gov/
[18] A.A. Toropov, E. Benfenati, Eur. J. Med. Chem. 42, 606 (2007)
[19] A.A. Toropov, T.W. Schultz, J. Chem. Inf. Comput. Sci. 43, 560 (2003)