Chemical Physics Letters 457 (2008) 332–336

Contents lists available at ScienceDirect

Chemical Physics Letters

journal homepage: www.elsevier.com/locate/cplett

Multiplicative SMILES-based optimal descriptors: QSPR modeling of fullerene C_{60} solubility in organic solvents

Andrey A. Toropov [a,*], Bakhtiyor F. Rasulev [a], Danuta Leszczynska [b], Jerzy Leszczynski [a]

[a] Computational Center for Molecular Structure and Interactions, Department of Chemistry, Jackson State University, 1400 J. R. Lynch Street, P.O. Box 17910, Jackson, MS 39217, USA
[b] Department of Civil and Environmental Engineering, Jackson State University, 1325 Lynch St, Jackson, MS 39217-0510, USA

ARTICLE INFO

Article history:
Received 13 February 2008
In final form 6 April 2008
Available online 10 April 2008

ABSTRACT

Optimal descriptors calculated with simplified molecular input line entry system have been used for modeling solubility of fullerene C_{60} in organic solvents. This approach provides improvement over the previous models. Statistical characteristics of the developed model are: $n = 92$, $R^2 = 0.9372$, $Q^2 = 0.9339$, $s = 0.270$, $F = 1342$ (training set); $n = 28$, $R^2 = 0.9151$, $R^2_{pred} = 0.9032$, $s = 0.334$, $F = 280$ (test set).

1. Introduction

Over last 15 years the basic interest in fullerenes has been surpassed by investigations of their industrial relevance. Among such emerging areas are nanoscience and nanotechnology. Due to applications of fullerenes in many aspects of nanotechnology [1–6] a development of predictive models of their physicochemical properties becomes very urgent task. One of the important characteristics of nanomaterials is their solubility. Solubility of fullerene is not only important technologically, but is also its crucial ecological (environmental) feature [6].

Quantitative structure-property relationship (QSPR) represents a powerful method for analyzing and predicting the properties of various compounds. Such methodology can be also applied for the predictions of different characteristics of fullerene including its solubility in various organic solvents.

Recently SMILES-based optimal descriptors have been used in QSPR modeling of the fullerene C_{60} solubility in organic solvents ($\log S$) [7]. The version of the SMILES-optimal descriptors described recently in [8] that was used for such predictions takes into account combinations of different SMILES attributes.

The aim of the present study is estimation of the optimal descriptors calculated with combinations of the SMILES attributes for the QSPR modeling of the C_{60} solubility.

2. Method

SMILES notations [9–12] for organic solvents examined in this study have been built with ACD/ChemSketch software [13]. The SMILES notation is representation of molecular structure by sequence of characters. These characters are representing presence of chemical elements (e.g., 'C', 'N', 'Cl', 'Br', etc.); a branching is indicated by brackets; a double bonds are indicated by '=', and other features of a molecular structure can be expressed by different symbols [9–12].

The experimental values of the C_{60} solubility $\log S$ (S is expressed in molar fraction) are taken from [1,7].

Optimal descriptors are defined as

$$DCW = CW(Nb) \cdot CW(Ndb) \cdot \prod CW(ss_k) \quad (1)$$

where Nb is a number of brackets in given SMILES, these are indicated below as '(000', '(001', etc.; Ndb is a number of the double covalent bonds indicated in SMILES by '=', these are indicated below as the '=000', '=001', etc.; ss_k represents two SMILES consequent elements in the SMILES strings; and $CW(SA_k)$ is the correlation weight of SMILES attribute of the SA_k, (SA_k = Nb or Ndb or ss_k).

$CW(ss_k)$ is correlation weight of two components' SMILES fragment. The 'Cl', 'Br', 'N+', and 'O-' are SMILES components containing two characters. All other SMILES components contain one character. The ss_k represents an association of the pairs of SMILES components and can be expressed by the following scheme:

'ABCDE...' = 'AB' + 'BC' + 'CD' + 'DE' + ...
or
'CC(N...' = 'CC' + 'C('+'(N'+...

For transparent identification the ss_k are inserted in blank of 12 symbols, for instance 'Cl' and 'C' are represented by following 12 symbols

—- —- —- = > Cl—- C— —- (Courier New)
or
_____ = > Cl__C_____ (Times New Roman).

It is possible that a molecular fragment can be represented by two different forms: i.e., 'AB' and 'BA'. It can lead to the existence two different correlation weights, $CW('AB') \neq CW('BA')$ for the

* Corresponding author.
E-mail address: aatoropov@yahoo.com (A.A. Toropov).

0009-2614/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.cplett.2008.04.013

same molecular fragment. In order to avoid this situation each two components SMILES fragment is arranged according to ASCII codes of its characters.

The ss_k are local SMILES attributes (i.e., they characterize a fragment of molecule). Other SMILES attributes (i.e., Nb and Ndb) are global because they characterize molecules in whole.

The correlation weights were calculated by Monte Carlo method optimization with the target function that represents the correlation coefficient between the DCW and experimental values of the C_{60} solubility ($\log S$) for the training set. The optimization is the following. For each attribute, CW is determined initially by setting the start values of all CWs to 1 ± 0.01*random. The random is 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 correlation coefficient (R1) between solubility and descriptor of Eq. (1) values on the training set is calculated. In a generated random sequence, each attribute correlation weight CWi was modified with the algorithm

1. ΔCWi:=0.001*CWi; Eps:=0.01*ΔCWi;
2. CWi:=CWi + ΔCWi;
3. Calculation of R2, i.e., the correlation coefficient between solubility and descriptor of Eq. (1) after modify CWi;
4. If R2 > R1 then R1:=R2; go to 2
5. CWi:=CWi − ΔCWi;
6. ΔCWi:= − 0.5*ΔCWi;
7. If absolute value (ΔCWi) >Eps then go to 2

Then steps of 1–7 are carried out for all CWs, one can repeat this algorithm from point of generation of random sequence. If the increasing the correlation coefficient becomes less 0.001 the process was stopping.

A correct division into training and test set is an important step for a QSPR analysis. The splits used in this work have been done accordingly to the logic suggested in [14]. In other words we have used random splits, but, every time, we select training and test with $\log S$ ranges as similar as possible.

There is a number of different software for generating SMILES and as a rule different software packages generate various SMILES [15]. It is to be noted that selected SMILES-based optimal descriptors should be calculated with SMILES generated by specific software. The application of a mixture of SMILES generated by different software packages is apparently improper.

3. Results

Before analysis of the obtained results one notices that here are SMILES attributes (SA_k) which are absent in the training set, but appear in the test set (Split 1). These are: Br__2_____ and c___I_____. Correlation weights of these SA_k have been fixed as 1. There are two outliers detected in this model. CAS numbers of these outliers are 111-96-6 and 591-49-1. They have been removed from the analysis.

To estimate ability of the applied approach, we have examined three splits of data into training and test sets. Lists of the test sets for these three splits are provided in Table 1. From the data of Table 2 one can see, that the SMILES-based optimal descriptors provide quite good models for these three splits. Below we demonstrate this approach with data for the split 1, the details on split 2 and split 3 are omitted.

The correlation weights obtained in the three runs (split 1) are displayed in Table 3. An example of the DCW calculation is provided in Table 4 (correlation weights are obtained in the first run). The model obtained in the first run can be expressed as follows: See Table 5

Table 1
Lists of the CAS numbers for the test sets of three splits into training and test sets

Test set for Split 1	Test set for Split 3	Test set for Split 3
629-59-4	110-54-3	124-18-5
110-82-7	26635-64-3	110-82-7
2207-01-4	124-18-5	542-18-7
1678-91-7	493-02-7	5401-62-7
67-66-3	108-85-0	2207-01-4
106-93-4	108-87-2	1678-91-7
106-94-5	540-54-5	74-95-3
109-64-8	75-29-6	74-96-4
78-77-3	78-87-5	106-93-4
507-20-0	142-28-9	540-54-5
558-17-8	627-31-6	106-94-5
79-01-6	513-36-0	107-08-4
108-88-3	507-19-7	78-87-5
106-42-3	558-17-8	142-28-9
488-23-3	103-65-1	78-75-1
100-41-4	98-82-8	78-77-3
135-98-8	98-06-6	513-38-2
591-50-4	462-06-6	108-88-3
541-73-1	99-08-1	106-42-3
583-53-9	30583-33-6	526-73-8
88-72-2	90-12-0	135-98-8
100-44-7	2586-62-1	583-53-9
90-13-1	71-23-8	88-72-2
2586-62-1	71-41-0	100-44-7
71-23-8	111-27-3	2586-62-1
111-27-3	107-13-1	71-23-8
111-87-5	68-12-2	111-87-5
107-13-1	110-86-1	110-01-0

Table 2
Statistical characteristics of the models for three runs of the Monte Carlo optimizations

Run	Training set, n = 92				Test set, n = 28			
	R^2	Q^2	s	F	R^2	R^2_{pred}	s	F
Split 1								
1	0.9372	0.9339	0.270	1342	0.9151	0.9032	0.334	280
2	0.9381	0.9348	0.268	1364	0.9170	0.9058	0.334	287
3	0.9368	0.9335	0.271	1335	0.9157	0.9031	0.333	282
Split 2								
1	0.9393	0.9362	0.257	1392	0.8812	0.8605	0.434	193
2	0.9366	0.9333	0.277	1329	0.8869	0.8705	0.355	204
3	0.9366	0.9334	0.263	1330	0.9008	0.8823	0.401	236
Split 3								
1	0.9301	0.9266	0.291	1197	0.9245	0.9128	0.303	318
2	0.9272	0.9234	0.297	1146	0.9257	0.9148	0.301	324
3	0.9349	0.9316	0.281	1292	0.9175	0.9006	0.299	289

$Q^2 = 1 - \frac{\sum[Y_{pred} - Y]^2}{\sum[Y - \overline{Y}(training)]^2}$ (Y and Y_{pred} on the training set)

$R^2_{pred} = 1 - \frac{\sum[Y_{pred} - Y]^2}{\sum[Y - \overline{Y}(training)]^2}$ (Y and Y_{pred} on the test set)

where Y and Y_{pred} are experimental and predicted values of the solubility, respectively; $\overline{Y}$ (training) is an average of the experimental values of the solubility over the training set.

$$\log S = -111.0059(\pm 0.3578) + 104.4283(\pm 0.3482)\ \mathrm{DCW}$$
$$n = 92, R^2 = 0.9372, Q^2 = 0.9339, s = 0.270, F = 1342\ \text{(training set)}$$
$$n = 28, R^2 = 0.9151, R^2_{pred} = 0.9032, s = 0.334, F = 280\ \text{(test set)} \quad (2)$$

Graphically this model is presented in Figs. 1 and 2 for the training and test sets, respectively.

There are few important findings from our study. From the data presented in Table 3 one can conclude that the number of given SA_k in the training and test sets provides vital information. For example, combination of 'C_C'appears 42 times in the training set, and 16 times in the test set, fragment 'C_(' appears 30 times in training set and eight times in test set, whereas fragment 'C_#' (Carbon atom and double bound) appears only one time in both the training

Table 3
Correlation weights obtained in three runs of the Monte Carlo optimizations

SFk	CW(SFk) in run 1	CW(SFk) in run 2	CW(SFk) in run 3	NS_{TRN}*	NS_{TST}*
Nb					
(000_____	1.0020126	1.0079526	1.0019862	58	18
(001_____	0.9995078	1.0001606	0.9983929	22	7
(002_____	1.0050649	0.9995640	1.0030261	11	3
(003_____	1.0080632	0.9974593	1.0051757	1	0
Ndb					
=000_____	1.0015239	1.0007616	1.0090123	78	25
=001_____	1.0005774	1.0017031	1.0032336	11	3
=002_____	0.9974884	0.9950636	0.9995177	2	0
=003_____	0.9932315	0.9940121	0.9877568	1	0
ss_k					
(___(_____	0.9922395	0.9961462	0.9928144	5	2
/___(_____	0.9973705	1.0065826	1.0007674	1	1
1___(_____	1.0091342	0.9936342	1.0042586	1	0
1___/_____	1.0013804	0.9996943	0.9933452	1	0
2___1_____	1.0072940	1.0013918	1.0068034	2	1
=___(_____	0.9973836	0.9965577	1.0000845	4	0
=___1_____	0.9968491	1.0018566	0.9996277	1	0
C___#_____	0.9966864	0.9972223	0.9946025	1	1
C___(_____	1.0011648	1.0023966	1.0013693	30	8
C___/_____	1.0012240	1.0041943	1.0022873	5	0
C___1_____	1.0027746	1.0023091	1.0029011	16	5
C___2_____	1.0018877	1.0017308	1.0019795	3	0
C___=_____	1.0029415	1.0002960	1.0052434	11	2
C___C_____	1.0017099	1.0013616	1.0015524	42	16
Br__(_____	1.0058510	1.0065463	1.0061871	4	0
Br__1_____	1.0103018	1.0086649	1.0103933	2	1
Br__C_____	1.0083138	1.0066863	1.0080894	12	4
Cl__(_____	1.0030300	1.0038726	1.0032197	9	4
Cl__/_____	1.0036380	0.9991489	1.0014439	2	1
Cl__1_____	1.0101403	1.0082998	1.0102093	1	0
Cl__C_____	1.0057271	1.0047479	1.0056522	12	2
I___(_____	1.0107270	1.0101919	1.0107900	1	1
I___C_____	1.0136363	1.0112341	1.0134472	6	0
N___#_____	0.9977557	0.9973260	0.9994555	1	1
N___(_____	1.0094557	1.0090451	1.0098842	2	0
N___/_____	1.0061706	1.0042150	1.0027233	1	0
N___1_____	1.0134387	1.0109352	1.0135689	1	0
N___=_____	1.0048384	1.0083872	1.0034045	1	0
N___C_____	1.0012249	1.0012649	1.0011540	4	0
O___(_____	1.0109299	1.0044299	1.0041816	1	0
O___=_____	0.9934739	0.9949591	0.9954636	7	1
O___C_____	0.9964546	0.9969737	0.9963004	4	3
S___1_____	0.9978520	0.9971357	1.0030375	1	0
S___C_____	1.0007746	1.0017432	0.9951992	1	0
[___(_____	1.0038873	1.0000338	1.0027011	2	1
[___=_____	1.0061830	1.0024541	0.9992372	1	1
[___N+____	0.9986161	1.0007709	1.0041027	2	1
[___O-____	0.9975272	1.0034463	0.9989813	2	1
[___[_____	1.0082968	1.0033677	1.0028863	1	0
___(_____	0.9955201	0.9983736	1.0008824	1	0
___1_____	1.0057830	1.0020334	1.0049033	3	0
___C_____	1.0003099	0.9995404	0.9989982	5	1
___Br____	1.0090313	1.0095217	1.0102536	1	0
___Cl____	1.0130719	1.0081873	1.0120744	1	1
c___(_____	0.9999402	1.0017036	1.0002127	14	5
c___/_____	1.0009706	1.0040743	1.0044049	1	0
c___1_____	1.0057579	1.0025992	1.0055169	39	12
c___2_____	1.0021895	1.0014607	1.0020729	5	2
c___3_____	1.0011735	1.0039808	1.0027798	1	0
c___C_____	1.0053418	1.0057770	1.0054442	16	6
c___Br____	1.0071535	1.0066603	1.0070073	2	1
c___Cl____	1.0064541	1.0060682	1.0062460	4	2
c___F_____	1.0007425	1.0010998	1.0003870	1	0
c___N_____	1.0024035	1.0023930	1.0022532	2	0
c___O_____	1.0136674	1.0119203	1.0135985	1	0
c___S_____	1.0110341	1.0097199	1.0108673	1	0
c___c_____	1.0004894	1.0018701	1.0006201	39	12
n___1_____	1.0081820	1.0041953	1.0075363	1	0
n___c_____	0.9997553	1.0021272	1.0000484	2	0
s___1_____	1.0013279	1.0077399	1.0088621	2	0
s___c_____	1.0072405	0.9999464	0.9994304	2	0

* N_{TRN} and N_{TST} are numbers of the SMILES which contain given SA_k in training set and test set, respectively.

Table 4
Example of calculation of DCW(SMILES) with $CW(SA_k)$ obtained in the first run of the Monte Carlo optimization

SA_k	$CW(SA_k)$
C___C_____	1.0017099
C___C_____	1.0017099
C___C_____	1.0017099
C___C_____	1.0017099
(000_____	1.0020126
=000_____	1.0015239

SMILES = 'CCCCC'; CAS= 109-66-0.
DCW = 1.0104210.

Table 5
Training and test sets; experimental and calculated using Eq. (2) values of the solubility, log*S*, of fullerene C_{60} in organic solvents. The *S* is expressed in molar fraction

CAS	SMILES	DCW	Expr	Calc	Expr-Calc
	Training set				
109-66-0	CCCCC	1.0104210	−6.100	−5.489	−0.611
110-54-3	CCCCCC	1.0121487	−5.100	−5.309	0.209
111-65-9	CCCCCCCC	1.0156130	−5.200	−4.947	−0.253
26635-64-3	CC(C)CCCCC	1.0143308	−5.200	−5.081	−0.119
124-18-5	CCCCCCCCCC	1.0190892	−4.700	−4.584	−0.116
112-40-3	CCCCCCCCCCCC	1.0225772	−3.500	−4.220	0.720
493-01-6	C1CCC2CCCCC2C1	1.0300835	−3.300	−3.436	0.136
493-02-7	C1CCC2CCCCC2C1	1.0300835	−3.500	−3.436	−0.064
137-43-9	BrC1CCCC1	1.0255718	−4.200	−3.907	−0.293
542-18-7	ClC1CCCCC1	1.0246900	−4.100	−3.999	−0.101
108-85-0	BrC1CCCCC1	1.0273255	−3.400	−3.724	0.324
626-62-0	IC1CCCCC1	1.0327483	−2.800	−3.158	0.358
5401-62-7	BrC1CCCCC1Br	1.0379088	−2.600	−2.619	0.019
110-83-8	C1\C=C/CCC1	1.0262827	−3.800	−3.833	0.033
108-87-2	CC1CCCCC1	1.0205970	−4.500	−4.427	−0.073
6876-23-9	CC1CCCCC1C	1.0234288	−4.600	−4.131	−0.469
75-09-2	ClCCl	1.0150672	−4.600	−5.004	0.404
56-23-5	ClC(Cl)(Cl)Cl	1.0210036	−4.400	−4.384	−0.016
74-95-3	BrCBr	1.0202954	−4.500	−4.458	−0.042
75-25-2	BrC(Br)Br	1.0283708	−3.200	−3.615	0.415
74-88-4	CI	1.0172241	−4.200	−4.779	0.579
74-97-5	BrCCl	1.0176779	−4.200	−4.732	0.532
74-96-4	BrCC	1.0136130	−5.200	−5.156	−0.044
75-03-6	CCI	1.0189635	−4.500	−4.597	0.097
79-34-5	ClC(Cl)C(Cl)Cl	1.0313876	−3.100	−3.300	0.200
107-06-2	ClCCCl	1.0168029	−5.000	−4.823	−0.177
71-55-6	CC(Cl)(Cl)Cl	1.0169254	−4.700	−4.810	0.110
540-54-5	CCCCl	1.0127414	−5.600	−5.247	−0.353
107-08-4	CCCI	1.0207058	−4.600	−4.415	−0.185
75-29-6	CC(C)Cl	1.0092996	−5.900	−5.606	−0.294
75-26-3	BrC(C)C	1.0140643	−5.400	−5.109	−0.291
75-30-9	CC(C)I	1.0170447	−4.800	−4.798	−0.002
78-87-5	CC(Cl)CCl	1.0169711	−4.900	−4.805	−0.095
142-28-9	ClCCCCl	1.0185415	−4.800	−4.641	−0.159
78-75-1	BrC(C)CBr	1.0224951	−4.300	−4.228	−0.072
627-31-6	ICCCl	1.0346245	−3.400	−2.962	−0.438
96-11-7	BrC(CBr)CBr	1.0358217	−2.900	−2.837	−0.063
96-18-4	ClCC(Cl)CCl	1.0227954	−4.000	−4.197	0.197
513-36-0	CC(C)CCl	1.0131924	−5.400	−5.200	−0.200
513-38-2	CC(C)CI	1.0211603	−4.300	−4.368	0.068
507-19-7	BrC(C)(C)C	1.0141474	−5.000	−5.100	0.100
540-49-8	Br\C=C\Br	1.0274329	−3.700	−3.713	0.013
127-18-4	Cl/C(Cl)=C(/Cl)Cl	1.0265767	−3.800	−3.802	0.002
513-37-1	C/C(C)=C\Cl	1.0198323	−4.500	−4.507	0.007
71-43-2	c1ccccc1	1.0229745	−4.000	−4.178	0.178
95-47-6	Cc1ccccc1C	1.0312926	−2.900	−3.310	0.410
108-38-3	Cc1cccc(C)c1	1.0276336	−3.300	−3.692	0.392
526-73-8	Cc1cccc(C)c1C	1.0304848	−3.100	−3.394	0.294
95-63-6	Cc1cc(C)c(C)cc1	1.0351253	−2.500	−2.910	0.410
108-67-8	Cc1cc(C)cc(C)c1	1.0351253	−3.500	−2.910	−0.590
527-53-7	Cc1cc(C)c(C)c(C)c1	1.0399998	−2.400	−2.400	0.000
119-64-2	c1ccc2CCCCc2c1	1.0414731	−2.500	−2.247	−0.253
103-65-1	CCCc1ccccc1	1.0319591	−3.500	−3.240	−0.260
98-82-8	CC(C)c1ccccc1	1.0256768	−3.600	−3.896	0.296

Table 5 (*continued*)

CAS	SMILES	DCW	Expr	Calc	Expr-Calc
104-51-8	CCCCc1ccccc1	1.0337237	−3.400	−3.056	−0.344
98-06-6	CC(C)(C)c1ccccc1	1.0257608	−3.700	−3.887	0.187
462-06-6	Fc1ccccc1	1.0237341	−4.100	−4.099	−0.001
108-90-7	Clc1ccccc1	1.0295769	−3.000	−3.489	0.489
108-86-1	Brc1ccccc1	1.0302924	−3.300	−3.414	0.114
95-50-1	Clc1ccccc1Cl	1.0400171	−2.400	−2.399	−0.001
108-36-1	Brc1cccc(Br)c1	1.0391455	−2.600	−2.490	−0.110
694-80-4	Clc1ccccc1Br	1.0401834	−2.400	−2.381	−0.019
108-37-2	Clc1cc(Br)ccc1	1.0384239	−3.000	−2.565	−0.435
120-82-1	Cl/C/1=C/C(Cl)C(Cl)\C=C\1	1.0362278	−2.800	−2.794	−0.006
100-42-5	C=Cc1ccccc1	1.0335206	−3.200	−3.077	−0.123
98-95-3	[O-][N+](=O)c1ccccc1	1.0256541	−3.900	−3.899	−0.001
100-47-0	N#Cc1ccccc1	1.0227308	−4.200	−4.204	0.004
100-66-3	COc1ccccc1	1.0332795	−3.100	−3.102	0.002
100-52-7	O=Cc1ccccc1	1.0237643	−4.200	−4.096	−0.104
103-71-9	O=C=N/c1ccccc1	1.0304056	−3.400	−3.402	0.002
99-08-1	O=[N+]([O-])c1cccc(C)c1	1.0304592	−3.400	−3.397	−0.003
108-98-5	Sc1ccccc1	1.0342621	−3.000	−3.000	−0.000
100-39-0	BrCc1ccccc1	1.0369893	−3.100	−2.715	−0.385
30583-33-6	ClC(Cl)(Cl)c1ccccc1	1.0375707	−3.000	−2.654	−0.346
90-12-0	Cc2cccc1ccccc12	1.0420102	−2.200	−2.191	−0.009
28804-88-8	Cc1c2ccccc2ccc1C	1.0413729	−2.100	−2.257	0.157
605-02-7	c1cc(c2ccccc2c1)c3ccccc3	1.0448134	−1.900	−1.898	−0.002
64-17-5	CCO	1.0016915	−7.100	−6.401	−0.699
71-36-3	CCCCO	1.0051200	−5.900	−6.043	0.143
71-41-0	CCCCCO	1.0068387	−5.300	−5.863	0.563
67-64-1	CC(C)=O	0.9961259	−7.000	−6.982	−0.018
68-12-2	CN(C)C=O	1.0106588	−5.300	−5.465	0.165
110-01-0	C1CCCS1	1.0111777	−5.400	−5.410	0.010
110-02-1	c1cccs1	1.0248396	−4.400	−3.984	−0.416
554-14-3	Cc1cccs1	1.0303141	−3.000	−3.412	0.412
872-50-4	O=C1CCCN1C	1.0256023	−3.900	−3.904	0.004
110-86-1	c1ccccn1	1.0246877	−4.000	−4.000	−0.000
91-22-5	c1cccc2cccnc12	1.0349531	−2.900	−2.928	0.028
62-53-3	Nc1ccccc1	1.0254333	−3.900	−3.922	0.022
100-61-8	CNc1ccccc1	1.0266893	−3.800	−3.790	−0.010
121-69-7	CN(C)c1ccccc1	1.0336700	−3.200	−3.062	−0.138
4904-61-4	C1\C=C/CC\C=C/CC/C=C\ C1 Test set	1.0371283	−2.700	−2.700	0.000
629-59-4	CCCCCCCCCCCCCC	1.0260772	−4.300	−3.854	−0.446
110-82-7	C1CCCCC1	1.0188549	−5.300	−4.609	−0.691
2207-01-4	CC1CCCCC1C	1.0234288	−4.600	−4.131	−0.469
1678-91-7	CCC1CCCCC1	1.0223422	−4.300	−4.244	−0.056
67-66-3	ClC(Cl)Cl	1.0171266	−4.800	−4.789	−0.011
106-93-4	BrCCBr	1.0220400	−4.200	−4.276	0.076
106-94-5	BrCCC	1.0153462	−5.200	−4.975	−0.225
109-64-8	BrCCCBr	1.0237876	−4.200	−4.094	−0.106
78-77-3	CC(C)CBr	1.0157983	−4.900	−4.928	0.028
507-20-0	CC(C)(C)Cl	1.0093823	−5.700	−5.598	−0.102
558-17-8	CC(C)(C)I	1.0171280	−4.400	−4.789	0.389
79-01-6	Cl\C=C(/Cl)Cl	1.0278505	−3.800	−3.669	−0.131
108-88-3	Cc1ccccc1	1.0284391	−3.400	−3.608	0.208
106-42-3	Cc1ccc(C)cc1	1.0276336	−3.300	−3.692	0.392
488-23-3	Cc1ccc(C)c(C)c1C	1.0379973	−2.900	−2.610	−0.290
100-41-4	CCc1ccccc1	1.0301976	−3.400	−3.424	0.024
135-98-8	CCC(C)c1ccccc1	1.0274306	−3.600	−3.713	0.113
591-50-4	Ic1ccccc1	1.0229745	−3.500	−4.178	0.678
541-73-1	Clc1cccc(Cl)c1	1.0326074	−3.400	−3.172	−0.228
583-53-9	Brc1ccccc1Br	1.0409063	−2.600	−2.306	−0.294
88-72-2	O=[N+]([O-])c1ccccc1C	1.0258397	−3.400	−3.879	0.479
100-44-7	ClCc1ccccc1	1.0343290	−3.400	−2.993	−0.407
90-13-1	Clc2cccc1ccccc12	1.0431631	−2.000	−2.070	0.070
2586-62-1	Cc2ccc1ccccc1c2Br	1.0421891	−2.100	−2.172	0.072
71-23-8	CCCO	1.0034043	−6.400	−6.222	−0.178
111-27-3	CCCCCCO	1.0085602	−5.100	−5.684	0.584
111-87-5	CCCCCCCCO	1.0120123	−5.000	−5.323	0.323
107-13-1	C=CC#N	1.0046153	−6.400	−6.096	−0.304

and test sets. Fragment 'Br_C' appears 12 times in the training set and four times in the test set. We believe that the larger numbers of SA_k in the training set are related to higher statistical significance of the particular SA_k for a given model.

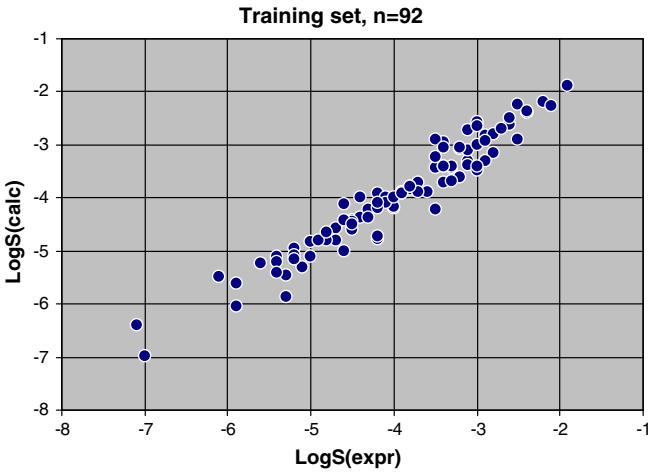

Fig. 1. Plot of experimental versus calculated solubility of fullerene C_{60} (log*S*, *S* is expressed in molar fraction) for the training set.

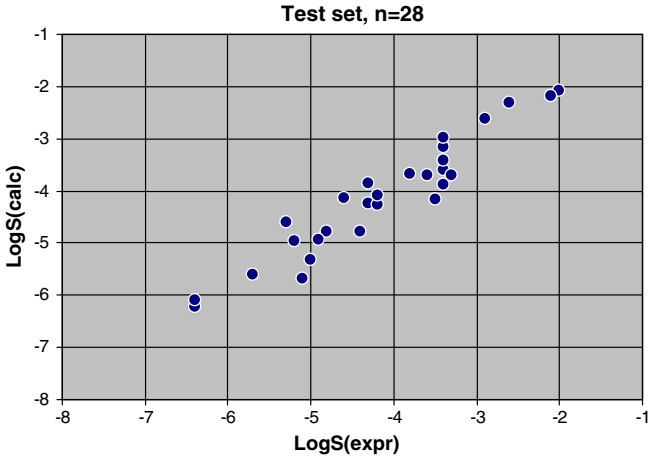

Fig. 2. Plot of experimental versus calculated solubility of fullerene C_{60} (log *S*, *S* is expressed in molar fraction) for the test set.

Optimal descriptors calculated with two components SA_k together with the global SMILES attributes (Nb and Ndb) are based on more information related to the molecular structure than the previous version of the descriptors ($n = 92$, $R^2 = 0.8612$, $Q^2 = 0.8537$, $s = 0.401$, $F = 558$ for training set and $n = 30$, $R^2 = 0.8908$, $R_2\text{pred} = 0.8748$, $s = 0.435$, $F = 228$ for test set) [7]. This improvement is confirmed by three different splits (Tables 1 and 2). However, further increase of the detailing may leads to the overtraining (i.e., a situation when an excellent model for the training set is accompanied by a poor model for the test).

Thus the proposed method has clear interpretations (each SMILES attribute is promoter of increase or decrease of the fullerene C_{60} solubility and this is defined by the correlation weight) and can be used in QSPR analyses of a data obtained directly from Internet databases which contain SMILES (generated by a specific software).

4. Conclusions

Our present study demonstrates that an application of the multiplicative SMILES based optimal descriptors technique for predictions of solubility of fullerene in organic solvents yields reliable

models. These models represent improvement over the initial development step as described in Ref. [7]. The statistical characteristics of the model calculated with Eq. (2) are better than the statistical characteristics of the model described in the previous study [1].

Acknowledgements

The authors would like to thank for support the High Performance Computational Design of Novel Materials (HPCDNM) – Contract #W912HZ-06-C-0057 and the Development of Predictive Techniques for Modeling Properties of NanoMaterials Using New OSPR/ASAR Approach Based on Optimal NanoDescriptors – Contract #W912HZ-06-C-096 Projects funded by the Department of Defense through the US. Army Engineer Research and Development Center, Vicksburg, MS. One of the authors (AAT) expresses gratitude for the Marie Curie fellowship financial support (the contract ID 39036, CHEMPREDICT).

References

[1] Yu.Yu Prylutskyy, et al. , Mater. Sci. Eng. Sect. C 23 (2003) 109.
[2] D.T. Colbert, R. Smalley, Trends Biotechnol. 17 (1999) 46.
[3] J. Wood, Nanotechnol. Mater. Today 7 (2004) 12.
[4] N. Sivaraman, T.G. Srinivasan, P.R. Vasudeva, J. Chem. Inf. Comput. Sci. 41 (2001) 1067.
[5] S.M. Danauskas, P.C. Jurs, J. Chem. Inf. Comput. Sci. 41 (2001) 419.
[6] A. Franco, S.F. Hansen, S.I. Olsen, L. Butti, Regul. Toxicol. Pharm. 48 (2007) 171.
[7] A.A. Toropov, B.F. Rasulev, D. Leszczynska, J. Leszczynski, Chem. Phys. Lett. 444 (2007) 209.
[8] A.A. Toropov, A.P. Toropova, D.V. Mukhamedzhanova, I. Gutman, Indian J. Chem. Sect. A 44 (2005) 1545.
[9] D. Weininger, J. Chem. Inf. Comput. Sci. 28 (1988) 31.
[10] D. Weininger, A. Weininger, J.L. Weininger, J. Chem. Inf. Comput. Sci. 29 (1989) 97.
[11] D. Weininger, J. Chem. Inf. Comput. Sci. 30 (1990) 237.
[12] <http://www.daylight.com/>.
[13] <http://www.acdlabs.com/>.
[14] J.T. Leonard, K. Roy, QSAR Combust. Sci. 25 (2006) 235.
[15] S.J. Coles, N.E. Day, P. Murray-Rust, H.S. Rzepa, Y. Zhang, Org. Biomol. Chem. 3 (2005) 1832.