Available online at www.sciencedirect.com

ScienceDirect

Chemical Physics Letters 441 (2007) 119–122

CHEMICAL PHYSICS LETTERS

www.elsevier.com/locate/cplett

QSPR study on solubility of fullerene C_{60} in organic solvents using optimal descriptors calculated with SMILES

Andrey A. Toropov *, Danuta Leszczynska, Jerzy Leszczynski

Computational Center for Molecular Structure and Interactions, Jackson State University, Jackson, MS, USA

Received 13 March 2007; in final form 13 April 2007
Available online 5 May 2007

Abstract

Theoretical modeling of solubility of C_{60} in various organic solvents (benzene derivatives) has been carried out. The optimal descriptors calculated with simplified molecular input line entry system notation have been applied in this study. The obtained model of fullerene C_{60} solubility (10^4 molar fraction of C_{60} at $T = 298$ K) in organic solvents statistically characterized by: $n = 25$, $r^2 = 0.81$, $s = 3.60$, $F = 101$ (training set) and $n = 11$, $r^2 = 0.79$, $s = 4.67$, $F = 34$ (test set).

1. Introduction

Quantitative structure–property/activity relationships (QSPR/QSAR) are efficient tools for prediction of physicochemical parameters and biological activity of chemicals. Fullerenes are known to be important components of many nanotechnologies. Thus, predictive models for physicochemical properties of these substances and their solutions can be useful in both technological and theoretical applications.

Rational selection of solvents for fullerenes in general and for fullerene C_{60} in particular is fairly important for both basic research and possible innovation applications [1]. QSPR analysis of the fullerene C_{60} solubility in different solvents based on information on molecular structures of solvents is the aim of the present study. Simplified molecular input line entry system (SMILES) has been used in this work as a method of elucidation of molecular structure. The recent QSPR/QSAR analysis [2–4] based on the SMILES notation has shown that such an approach can be reasonable good alternative for molecular graph [3]. The modeling of solubility (10^4 molar fraction of C_{60} at $T = 298$ K) taken from Ref. [1] that is proposed in the present work provides another example of the application of the SMILES approach. A correlation between solubility of the fullerene C_{60} and structure of solvents is quite complex. Taking into account these facts we selected a representative series of benzene derivatives for the QSPR analysis. In the case of several values of solubility reported for one solvent the average solubility has been used.

2. Method

One variable correlations used for prediction of fullerene C_{60} solubility have been obtained with descriptor of correlation weights (DCW) calculated as

$$\mathrm{DCW} = \prod_{k=1}^{N} \mathrm{CW}(S_k) \tag{1}$$

where S_k is as a rule one character fragment of the SMILES notation (except two characters for Cl and Br, as well as four characters for [N+] and [O−]), $\mathrm{CW}(S_k)$ is the so-called correlation weight of the S_k, N is a number of the fragments in the given SMILES.

The correlation weights $\mathrm{CW}(S_k)$ are calculated as coefficients which produce as large as possible correlation coefficient between DCW and the solubility for the training set. The present calculations have been done by the Monte Carlo method described in Ref. [5].

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

0009-2614/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.cplett.2007.04.094

Having numerical data for these CW(S_k) one can calculate DCW for all solvents of the training and test sets. Using data for the training set, one can calculate solubility by Least Squares method following

$$S = C_0 + C_1 \cdot \mathrm{DCW} \tag{2}$$

where the S is the calculated solubility of C_{60}, which can be validated with solvents of the test set. Separation into the training and test sets has been done randomly but interval of the S should be similar for the training and test sets.

3. Results and discussion

Statistical characteristics of Eq. (2) like models for the solubility are collected in Table 1. One can see from Table 1 that statistical quality of the models is reasonable good, and also the statistical characteristics are reproduced well in case of basic split into training and test sets as well as for some additional different splits into training and test sets. These splits have been obtained by means of exchange some solvents from training set into test set and vice versa. Numerical values of the correlation weights for calculation of the DCW obtained in three runs of the Monte Carlo optimization are listed in Table 2. An example of calculation of the DCW is demonstrated in Table 3. The model of the fullerene solubility (S), obtained in the first run of the Monte Carlo optimization is represented as follows:

$$S = -93.9871(\pm 2.4060) + 51.2212(\pm 1.2252) \cdot \mathrm{DCW} \tag{3}$$

Table 1
Statistical characteristics of models for three runs of the Monte Carlo optimization basic split into training and test sets

Probe	Training set, $n = 25$			Test set, $n = 11$			
	R^2	S	F	R^2	R^2_{pred} [a]	S	F
1	0.8140	3.60	101	0.7887	0.7203	4.67	34
2	0.8150	3.59	101	0.7900	0.7203	4.67	34
3	0.8161	3.58	102	0.7903	0.7235	4.65	34
Exchange: 71-43-2 in test set and 95-47-6 in training set							
1	0.7469	4.21	68	0.8937	0.8659	3.32	76
2	0.7478	4.20	68	0.8940	0.8658	3.32	76
3	0.7568	4.13	72	0.8968	0.8743	3.21	78
Exchange: 180-90-7 in test set and 591-50-4 in training set							
1	0.8374	3.38	118	0.7680	0.5935	5.45	30
2	0.8416	3.33	122	0.7678	0.6012	5.40	30
3	0.8393	3.36	120	0.7713	0.5986	5.41	30
Exchange: 71-43-2 in test set and 591-50-4 in training set							
1	0.8160	3.57	102	0.7845	0.7289	4.67	33
2	0.8151	3.57	101	0.7826	0.7290	4.66	32
3	0.8163	3.56	102	0.7870	0.7335	4.63	33
Exchange: 180-90-7 in test set and 95-47-6 in training set							
1	0.7646	4.10	75	0.8670	0.8411	3.48	59
2	0.7654	4.10	75	0.8715	0.8450	3.44	61
3	0.7621	4.12	74	0.8644	0.8378	3.52	57

[a] $R^2_{\mathrm{pred}} = 1 - \sum_{k=1}^{n}(E_{y_k} - C_{y_k})^2 / \sum_{k=1}^{n}(E_{y_k} - A_y)^2$ where E_{y_k}, C_{y_k} are experimental and calculated values of the solubility, respectively; A_y is average value of solubility on given set (i.e., training or test set), n is number of compounds in the set.

Table 2
Correlation weights of the SMILES fragments obtained in three runs of the Monte Carlo optimization for basic split into training and test sets

S_k	CW(S_k) in Probe 1	CW(S_k) in Probe 2	CW(S_k) in Probe 3	N_{TRN}	N_{TST}
#	0.9759488	0.9975168	0.9884857	1	0
(	1.0083450	1.0091311	1.0094479	34	10
/	1.0007953	1.0252143	1.0167532	1	0
1	1.0947374	1.0733150	1.0551286	50	22
2	1.1311368	1.1408825	1.1503892	2	0
=	0.9562439	0.9538631	0.9494339	6	1
C	1.0010415	1.0012220	1.0017311	40	14
Br	1.0881372	1.0951278	1.1030406	5	3
Cl	1.0166561	1.0179934	1.0199889	4	4
F	0.9830704	0.9820306	0.9812587	1	0
I	1.0	1.0	1.0	0	1
N	1.0044482	0.9810229	0.9889116	2	0
O	1.1008799	1.1084716	1.1177687	5	3
S	1.0766008	1.0837499	1.0905982	1	0
c	1.0781328	1.0149657	1.0164879	150	66
[N+]	0.9618867	0.9572778	0.9933863	2	1
[O−]	0.9770318	0.9768140	0.9380622	2	1

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

Table 3
Example of the DCW calculation for benzene with correlation weights of the first run of Monte Carlo optimization: SMILES = (c1ccccc1), DCW = 1.8821467

S_k	CW(S_k)
c	1.0781328
1	1.0947374
c	1.0781328
c	1.0781328
c	1.0781328
c	1.0781328
c	1.0781328
1	1.0947374

Experimental and calculated values of solubility for the training and test sets are shown in Table 4. A visualization of the correlation between experimental and calculated values of the solubility for the training and test sets (basic split) is displayed in Figs. 1 and 2, respectively.

Unexpectedly large differences between the experimental and calculated values of the C_{60} fullerene solubility are found for relative simple structures: 1,2-dimethylbenzene (CAS 95-47-6), bromobenzene (CAS 108-86-1), and benzaldehyde (CAS 100-52-7). One can assume that the possible differences between experimental and predicted values of solubility for simple solvents are due to a formation of a specific associations of solvent clusters near the C_{60} molecules. This possibility warrants separate study and characterization of such solvent–fullerene complexes that could be done by application of the reliable ab initio methods.

Fullerenes as well as other nanostructures can hardly be represented by the ‘classical’ molecular graph. Taking this into account the SMILES-like representation of nanostructures provides an efficient, perspective approach for construction of predictive models for nano substances.

Table 4
Basic split into training and test sets of the organic solvents

CAS	SMILES	DCW	S_{Expr}	S_{Calc}	$S_{\text{Expr}} - S_{\text{Calc}}$
Training set					
71-43-2	c1ccccc1	1.8821467	1.81	2.42411	−0.61411
108-88-3	Cc1ccccc1	1.8841069	3.55	2.52452	1.02548
108-38-3	Cc1cccc(C)c1	1.9176791	3.59	4.24412	−0.65412
106-42-3	Cc1ccc(C)cc1	1.9176791	7.67	4.24412	3.42588
108-67-8	Cc1cc(C)cc(C)c1	1.9518494	2.70	5.99437	−3.29437
488-23-3	Cc1ccc(C)c(C)c1C[a]	1.9538823	12.00	6.09849	5.90151
119-64-2	c1ccc2CCCCc2c1	2.4181991	30.10	29.88136	0.21864
103-65-1	CCCc1ccccc1	1.8880336	2.90	2.72564	0.17436
98-82-8	CC(C)c1ccccc1	1.9196763	2.32	4.34642	−2.02642
135-98-8	CCC(C)c1ccccc1	1.9216757	2.38	4.44883	−2.06883
98-06-6	CC(C)(C)c1ccccc1	1.9538823	1.93	6.09849	−4.16849
462-06-6	Fc1ccccc1	1.8502827	0.77	0.79200	−0.02200
108-90-7	Clc1ccccc1	1.9134959	8.96	4.02986	4.93014
108-86-1	**Brc1ccccc1**	***2.0480338***	***4.45***	***10.92105***	***−6.47105***
108-36-1	Brc1cccc(Br)c1	2.2658913	23.10	22.07997	1.02003
100-42-5	C=Cc1ccccc1	1.8035422	5.97	−1.60210	7.57210
100-47-0	N#Cc1ccccc1	1.8469712	0.58	0.62238	−0.04238
104-92-7	COc1ccc(Br)cc1	2.2948138	29.90	23.56141	6.33859
100-52-7	**O=Cc1ccccc1**	***1.9834176***	***0.59***	***7.61133***	***−7.02133***
103-71-9	O=C=N/c1ccccc1	1.9065827	3.68	3.67575	0.00425
108-98-5	Sc1ccccc1	2.0263206	9.84	9.80887	0.03113
88-72-2	O=[N+]([O−])c1ccccc1C	1.8952438	3.98	3.09496	0.88504
99-08-1	O=[N+]([O−])c1cccc(C)c1	1.9270074	3.88	4.72193	−0.84193
100-39-0	BrCc1ccccc1	2.0501668	8.15	11.03031	−2.88031
98-07-7	ClC(Cl)(Cl)c1ccccc1	2.0467479	9.43	10.85518	−1.42518
Test set					
95-47-6	**Cc1ccccc1C**	***1.8860692***	***14.30***	***2.62503***	***11.67497***
526-73-8	Cc1cccc(C)c1C	1.9196763	8.77	4.34642	4.42358
100-41-4	CCc1ccccc1	1.8860692	3.79	2.62503	1.16497
104-51-8	CCCCc1ccccc1	1.8899999	4.12	2.82637	1.29363
591-50-4	Ic1ccccc1[b]	1.8821467	3.26	2.42411	0.83589
583-53-9	Brc1ccccc1Br	2.2285418	23.10	20.16688	2.93312
120-82-1	Clc1cc(Cl)c(Cl)cc1	2.0446184	13.70	10.74611	2.95389
98-95-3	[O−][N+](=O)c1ccccc1	1.8932720	1.14	2.99396	−1.85396
100-66-3	COc1ccccc1	2.0741754	8.45	12.26006	−3.81006
2398-37-0	COc1cc(Br)ccc1	2.2948138	28.40	23.56141	4.83859
100-44-7	ClCc1ccccc1	1.9154888	3.83	4.13193	−0.30193

Experimental and calculated values of solubility fullerene C_{60} (S, 10^4 molar fraction C_{60} at $T = 298$ K).

[a] Solvents which have large value of the differences between ($S_{\text{Expr}} - S_{\text{Calc}}$).

[b] I (iodine) is absent in the training set, CW(I) = 1.0.

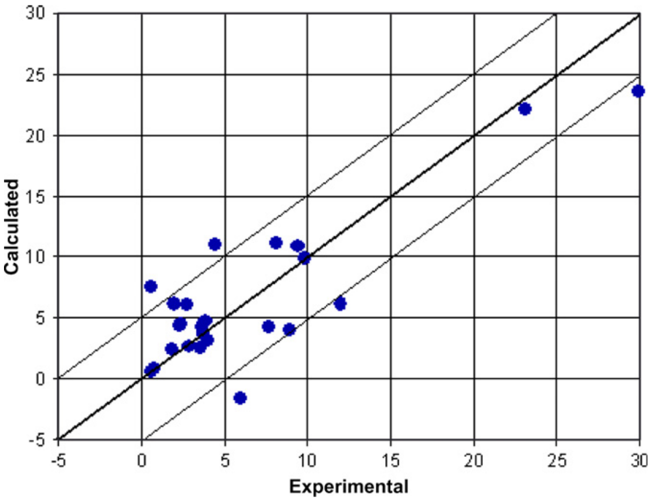

Fig. 1. Plot of experimental versus calculated solubility of fullerene C_{60} in organic solvents for the training set (10^4 molar fraction, at $T = 298$ K).

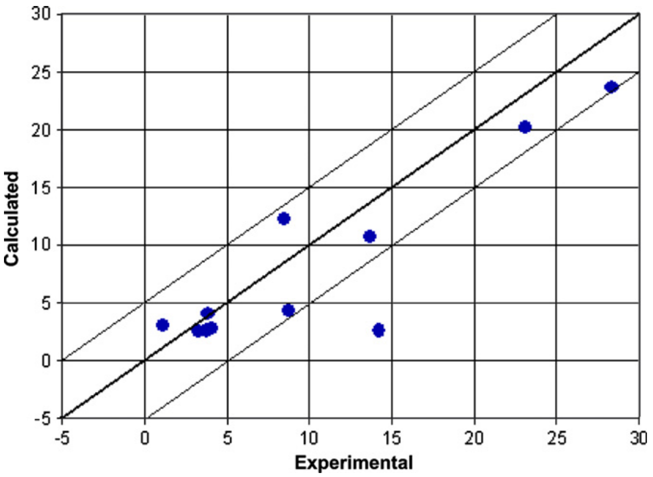

Fig. 2. Plot of experimental versus calculated solubility of fullerene C_{60} in organic solvents for the test set (10^4 molar fraction, at $T = 298$ K).

Recently, reasonable good models of Young's modulus were obtained by the similar approach [6].

4. Conclusions

Optimal descriptor calculated with SMILES notation can be used for prediction of the solubility of fullerene C_{60} in organic solvents. The current work has been limited to the benzene derivatives. However we believe that the similar approach could be applied to a more general class of solvents.

Acknowledgements

The authors would like to thanks 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-0061 Projects funded by the Department of Defense through the US Army Engineer Research and Development Center, Vicksburg, MS.

References

[1] M.V. Korobov, A.L. Smith, Solubility of the fullerenes, in: K.M. Kadish, R.S. Ruoff (Eds.), Fullerenes: Chemistry, Physics, and Technology, Wiley Inter Science, 2000, p. 55 (Chapter 2).

[2] D. Vidal, M. Thormann, M. Pons, J. Chem. Inf. Model. 45 (2005) 386.

[3] A.A. Toropov, A.P. Toropova, D.V. Mukhamedzhanova, I. Gutman, Indian J. Chem. A 44 (2005) 1545.

[4] A. Toropov, K. Nesmerak, I. Raska Jr., K. Waisser, K. Palat, Comput. Biol. Chem. 30 (2006) 434.

[5] A.A. Toropov, T.W. Schultz, J. Chem. Inf. Comput. Sci. 43 (2003) 560.

[6] A.A. Toropov, J. Leszczynski, Chem. Phys. Lett. 433 (2006) 125.