Available online at www.sciencedirect.com
ScienceDirect
Chemical Physics Letters 444 (2007) 209–214
CHEMICAL PHYSICS LETTERS
www.elsevier.com/locate/cplett

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

Andrey A. Toropov *, Bakhtiyor F. Rasulev, Danuta Leszczynska, Jerzy Leszczynski

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

Received 16 May 2007; in final form 27 June 2007
Available online 12 July 2007

Abstract

Optimal descriptors, calculated with simplified molecular input line entry system have been used for modeling solubility of fullerene C_{60} in organic solvents. Statistical characteristics of the model are $n = 92$, $R^2 = 0.8612$, $Q^2 = 0.8537$, $s = 0.401$, $F = 558$ (training set) $n = 30$, $R^2 = 0.8908$, $R^2_{\text{pred}} = 0.8748$, $s = 0.435$, $F = 228$ (test set). The applied approach entirely based on topological data provides a reliable model for solubility of C_{60}.

1. Introduction

Data on solubility of the fullerene C_{60} in organic solvents can provide important information that would be of interest in chemistry and biochemistry. Quantitative structure–property (solubility) relationships (QSPR) for this characteristic are not only able to predict numerical data but also to evaluate features of solubility of fullerene C60 as a function of a complex physicochemical phenomenon.

At present the only method to construct predictive model of the fullerene C_{60} solubility in different organic solvents involves QSPR is based on molecular descriptors calculated with molecular graphs of the solvents [1]. Taking into account gradual increase of a number of internet database on physicochemical parameters and biological activity with elucidation of molecular structure by simplified molecular input line entry system (SMILES) there is a need to develop SMILES based predictive models [2–5].

The present study is aimed to estimate predictive potential of the SMILES based optimal descriptors in QSPR modeling of the fullerene C_{60} solubility in organic solvents as a function of the molecular structure of the solvents represented by the SMILES notation. Experimental values of fullerene solubility ($\log S$) are taken from [1].

2. Method

SMILES notations [6] of organic solvents for this study have been built with ACD/ChemSketch software [7].

Optimal descriptors have been defined as

$$\text{DCW(SMILES)} = \sum_{k=1}^{n} \text{CW}(\text{SF}_k), \tag{1}$$

where SF_k is a fragment of SMILES. Majority of the SMILES fragments used in the present study contain one character, except 'Cl', 'Br', [O^-], and [N^+]; $\text{CW}(\text{SF}_k)$ represents correlation weights of the SF_k. Numerical values of the $\text{CW}(\text{SF}_k)$ have been calculated by Monte Carlo method [3–5]. The $\text{CW}(\text{SF}_k)$ produce as large as possible correlation coefficient between the solubility of fullerene C_{60} in organic solvents for the training set. In other words the correlation coefficient is using in role of the target function for the optimization procedure. Having numerical data on the $\text{CW}(\text{SF}_k)$ one can calculate with Eq. (1) DCW(SMILES).

* 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.2007.07.024

An application of the least squares method yields following equation:

$$\log S = C_0 + C_1 \mathrm{DCW(SMILES)}. \quad (2)$$

Predictive potential of the model calculated with Eq. (2) can be estimated using data on organic solvents of the external test set. The same split into training and test sets that has been used in [1] is used in the present study.

3. Results

Statistical characteristics of QSPR for solubility, $\log(S)$, (S is expressed in molar fraction) of fullerene C_{60} in organic solvents are presented in Table 1. One can see from the Table 1 that a statistical quality of these models is quite satisfactory.

Correlation weights for calculation of the DCW(SMILES) values with Eq. (1) obtained in three runs of the Monte Carlo optimization are presented in Table 2. Example of DCW(SMILES) calculation is provided in Table 3.

The one variable model obtained in the first run of the Monte Carlo optimization is given below (see Table 4):

$$\log S = -6.6196(\pm 0.0154) + 0.5014(\pm 0.00266)\ \mathrm{DCW(SMILES)}, \quad (3)$$

$$n = 92,\ R^2 = 0.8612,\ Q^2 = 0.8537,\ s = 0.401,\ F = 558\ \text{(training set)}, \quad (4)$$

$$n = 30,\ R^2 = 0.8908,\ R^2_{\mathrm{pred}} = 0.8748,\ s = 0.435,\ F = 228\ \text{(test set)}. \quad (5)$$

Table 1
Statistical characteristics of the models for three runs of the Monte Carlo optimizations

Run	Training set, $n = 92$				Test set, $n = 30$			
	R^2	Q^2	s	F	R^2	R^2_{pred}	s	F
1	0.8612	0.8537	0.401	558	0.8908	0.8748	0.435	228
2	0.8612	0.8538	0.401	558	0.8905	0.8747	0.435	228
3	0.8612	0.8537	0.401	559	0.8911	0.8753	0.433	229

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

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

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

Table 2
Correlation weights obtained for three runs of the Monte Carlo optimizations

SF_k	$CW(SF_k)$ in run 1	$CW(SF_k)$ in run 2	$CW(SF_k)$ in run 3	k	N_{TRN}[a]	N_{TST}	dR_{TRN}[b] (run 1)	dR_{TST} (run 1)
#	−1.7513027	−1.9763694	−1.9363560	1	1	1	0.00713	0.02592
(	−0.0636205	−0.0764917	−0.0582624	2	94	26	0.00184	−0.00212
/	−0.2224238	−0.2195869	−0.2266901	3	11	1	0.00252	−0.00233
1	0.7333020	0.8328653	0.8232880	4	106	32	0.09384	0.07609
2	0.2772195	0.3115025	0.3167455	5	14	4	0.00446	0.00103
3	−1.3688641	−1.5422654	−1.5207150	6	2	0	0.01551	0.0
=	0.2367907	0.2224600	0.2178262	7	18	5	0.00311	0.00979
C	0.4300723	0.4930871	0.4751774	8	283	108	0.19261	0.10257
Br	1.8947357	2.1645957	2.0997835	9	26	9	0.28179	0.18802
Cl	1.2254199	1.3933766	1.3476833	10	43	11	0.23601	0.14863
F	-0.2913806	-0.3096783	−0.3103693	11	1	0	0.00019	0.0
I	2.9882335	3.3870007	3.3146112	12	8	2	0.15820	0.04119
N	0.8279941	0.9514299	0.9182467	13	7	1	0.01045	−0.00245
O	−0.3217103	−0.3274301	−0.3323367	14	11	7	0.00224	−0.00838
S	0.5774412	0.6548567	0.6338071	15	2	0	0.00154	0.0
\	0.4626149	0.5323039	0.5491613	16	9	1	0.00960	0.00637
c	0.6408136	0.7275105	0.7082618	17	250	80	0.77852	0.77009
n	0.0822050	0.1204241	0.0986911	18	2	0	0.00004	0.0
s	1.5723750	1.8045279	1.7543516	19	2	0	0.01154	0.0
[N^+]	0.4329171	0.2983848	0.4696330	20	2	1	0.00087	0.00424
[O^-]	0.2397301	0.5001536	0.2529812	21	2	1	0.00027	0.00212

[a] N_{TRN} and N_{TST} are numbers of the SF_k in training set and test set, respectively.
[b] $dR = R - R_k$, where R_k is correlation coefficient after assignation $CW(SF_k) = 0$; dR_{TRN} and dR_{TST} are values for training set and test set, respectively.

Table 3
Example of calculation of DCW(SMILES) with $CW(SF_k)$ obtained in the first run of the Monte Carlo optimization

SF_k	$CW(SF_k)$ in run 1	k
C	0.4300723	8
C	0.4300723	8
=	0.2367907	7
1	0.7333020	4
C	0.4300723	8
C	0.4300723	8
C	0.4300723	8
C	0.4300723	8
C	0.4300723	8
=	0.2367907	7
1	0.7333020	4

SMILES = [CC=1CCCCC=1] CAS = 59-49-1 DCW = 4.9506915.

A graphical representation of this model is displayed for the training and the test sets in Figs. 1 and 2, respectively.

Statistical characteristics of nonlinear model for the fullerene C_{60} solubility described in [1] are $n = 122$, $R^2 = 0.892$, $s = 366$. This model is based on quantum chemical descriptors. Attempts to build nonlinear SMILES based models similar to described in [3] have shown that using of the DCW^2 or $DCW^{0.5}$ is not able to improve statistical quality of the models for the log S.

From the Table 2 one can see that some of the $CW(SF_k)$ have stable positive contributions ('Cl', 'Br', 'I', etc.) whereas the others have stable negative contributions, for instance, triple bonds '#', number of cycles '3', branching '('.

Table 4
Training and test sets; experimental and calculated using Eq. (3) values of the solubility, log S, of fullerene C_{60} in organic solvents

CAS No.	SMILES	DCW(SMILES)	$\log S_{expr}$	$\log S_{calc}$	$\log S_{expr} - \log S_{calc}$
	Training set, n = 92				
109-66-0	CCCCC	2.1503615	−6.1	−5.541	−0.559
110-54-3	CCCCCC	2.5804338	−5.1	−5.326	0.226
111-65-9	CCCCCCCC	3.4405784	−5.2	−4.894	−0.306
26635-64-3	CC(C)CCCCC	3.3133374	−5.2	−4.958	−0.242
124-18-5	CCCCCCCCCC	4.3007230	−4.7	−4.463	−0.237
112-40-3	CCCCCCCCCCCC	5.1608676	−3.5	−4.032	0.532
493-01-6	C1CCC2CCCCC2C1	6.3217660	−3.3	−3.450	0.150
493-02-7	C1CCC2CCCCC2C1	6.3217660	−3.5	−3.450	−0.050
137-43-9	BrC1CCCC1	5.5117012	−4.2	−3.856	−0.344
542-18-7	ClC1CCCCC1	5.2724577	−4.1	−3.976	−0.124
108-85-0	BrC1CCCCC1	5.9417735	−3.4	−3.640	0.240
626-62-0	IC1CCCCC1	7.0352713	−2.8	−3.092	0.292
5401-62-7	BrC1CCCCC1Br	7.8365092	−2.6	−2.690	0.090
110-83-8	C1\C=C/CCC1	4.5240196	−3.8	−4.351	0.551
108-87-2	CC1CCCCC1	4.4771101	−4.5	−4.375	−0.125
6876-23-9	CC1CCCCC1C	4.9071824	−4.6	−4.159	−0.441
75-09-2	ClCCl	2.8809121	−4.6	−5.175	0.575
56-23-5	ClC(Cl)(Cl)Cl	5.0772699	−4.4	−4.074	−0.326
74-95-3	BrCBr	4.2195437	−4.5	−4.504	0.004
75-25-2	BrC(Br)Br	5.9870384	−3.2	−3.618	0.418
74-88-4	CI	3.4183058	−4.2	−4.906	0.706
74-97-5	BrCCl	3.5502279	−4.2	−4.840	0.640
74-96-4	BrCC	2.7548803	−5.2	−5.238	0.038
75-03-6	CCI	3.8483781	−4.5	−4.690	0.190
79-34-5	ClC(Cl)C(Cl)Cl	5.5073422	−3.1	−3.858	0.758
107-06-2	ClCCCl	3.3109844	−5.0	−4.959	−0.041
71-55-6	CC(Cl)(Cl)Cl	4.2819223	−4.7	−4.473	−0.227
540-54-5	CCCCl	2.5156368	−5.6	−5.358	−0.242
107-08-4	CCCI	4.2784504	−4.6	−4.474	−0.126
75-29-6	CC(C)Cl	2.3883958	−5.9	−5.422	−0.478
75-26-3	BrC(C)C	3.0577116	−5.4	−5.086	−0.314
75-30-9	CC(C)I	4.1512094	−4.8	−4.538	−0.262
78-87-5	CC(Cl)CCl	3.6138157	−4.9	−4.808	−0.092
142-28-9	ClCCCCl	3.7410567	−4.8	−4.744	−0.056
78-75-1	BrC(C)CBr	4.9524473	−4.3	−4.136	−0.164
627-31-6	ICCCI	7.2666839	−3.4	−2.976	−0.424
96-11-7	BrC(CBr)CBr	6.8471830	−2.9	−3.186	0.286
96-18-4	ClCC(Cl)CCl	4.8392356	−4.0	−4.193	0.193
513-36-0	CC(C)CCl	2.8184681	−5.4	−5.206	−0.194
513-38-2	CC(C)CI	4.5812817	−4.3	−4.323	0.023
507-19-7	BrC(C)(C)C	3.3605429	−5.0	−4.935	−0.065
540-49-8	Br\C=C\Br	5.8116365	−3.7	−3.706	0.006

(*continued on next page*)

Table 4 (*continued*)

CAS No.	SMILES	DCW(SMILES)	$\log S_{expr}$	$\log S_{calc}$	$\log S_{expr} - \log S_{calc}$
127-18-4	Cl/C(Cl)=C(/Cl)Cl	5.2992853	−3.8	−3.963	0.163
513-37-1	C/C(C)=C\Cl	3.2954499	−4.5	−4.967	0.467
71-43-2	c1ccccc1	5.3114856	−4.0	−3.956	−0.044
95-47-6	Cc1ccccc1C	6.1716302	−2.9	−3.525	0.625
108-38-3	Cc1cccc(C)c1	6.0443892	−3.3	−3.589	0.289
526-73-8	Cc1cccc(C)c1C	6.4744615	−3.1	−3.373	0.273
95-63-6	Cc1cc(C)c(C)cc1	6.3472205	−2.5	−3.437	0.937
108-67-8	Cc1cc(C)cc(C)c1	6.3472205	−3.5	−3.437	−0.063
527-53-7	Cc1cc(C)c(C)c(C)c1	6.6500518	−2.4	−3.285	0.885
119-64-2	c1ccc2CCCCc2c1	7.5862138	−2.5	−2.816	0.316
103-65-1	CCCc1ccccc1	6.6017025	−3.5	−3.310	−0.190
98-82-8	CC(C)c1ccccc1	6.4744615	−3.6	−3.373	−0.227
104-51-8	CCCCc1ccccc1	7.0317748	−3.4	−3.094	−0.306
98-06-6	CC(C)(C)c1ccccc1	6.7772928	−3.7	−3.221	−0.479
462-06-6	Fc1ccccc1	5.0201050	−4.1	−4.103	0.003
108-90-7	Clc1ccccc1	6.5369055	−3.0	−3.342	0.342
108-86-1	Brc1ccccc1	7.2062213	−3.3	−3.006	−0.294
95-50-1	Clc1ccccc1Cl	7.7623254	−2.4	−2.728	0.328
108-36-1	Brc1cccc(Br)c1	8.9737160	−2.6	−2.120	−0.480
694-80-4	Clc1ccccc1Br	8.4316412	−2.4	−2.392	−0.008
108-37-2	Clc1cc(Br)ccc1	8.3044002	−3.0	−2.456	−0.544
120-82-1	Cl/C/1=C/C(Cl)C(Cl)\C=C\1	8.2003553	−2.8	−2.508	−0.292
100-42-5	C=Cc1ccccc1	6.4084209	−3.2	−3.406	0.206
98-95-3	[O$^-$][N$^+$](=O)c1ccccc1	5.7719722	−3.9	−3.726	−0.174
100-47-0	N#Cc1ccccc1	4.8182493	−4.2	−4.204	0.004
100-66-3	COc1ccccc1	5.4198476	−3.1	−3.902	0.802
100-52-7	O=Cc1ccccc1	5.6566383	−4.2	−3.783	−0.417
103-71-9	O=C=N/c1ccccc1	6.4989993	−3.4	−3.361	−0.039
99-08-1	O=[N$^+$]([O$^-$])c1cccc(C)c1	6.0748035	−3.4	−3.574	0.174
108-98-5	Sc1ccccc1	5.8889268	−3.0	−3.667	0.667
100-39-0	BrCc1ccccc1	7.6362936	−3.1	−2.791	−0.309
30583-33-6	ClC(Cl)(Cl)c1ccccc1	9.1633356	−3.0	−2.025	−0.975
90-12-0	Cc2cccc1ccccc12	8.8592513	−2.2	−2.178	−0.022
28804-88-8	Cc1c2ccccc2ccc1C	9.2893236	−2.1	−1.962	−0.138
605-02-7	c1cc(c2ccccc2c1)c3ccccc3	9.4090914	−1.9	−1.902	0.002
64-17-5	CCO	0.5384343	−7.1	−6.350	−0.750
71-36-3	CCCCO	1.3985789	−5.9	−5.918	0.018
71-41-0	CCCCCO	1.8286512	−5.3	−5.703	0.403
67-64-1	CC(C)=O	1.0780563	−7.0	−6.079	−0.921
68-12-2	CN(C)C=O	1.9060504	−5.3	−5.664	0.364
110-01-0	C1CCCS1	3.7643344	−5.4	−4.732	−0.668
110-02-1	c1cccs1	5.6022334	−4.4	−3.811	−0.589
554-14-3	Cc1cccs1	6.0323057	−3.0	−3.595	0.595
872-50-4	O=C1CCCN1C	4.3600400	−3.9	−4.433	0.533
110-86-1	c1ccccn1	4.7528770	−4.0	−4.237	0.237
91-22-5	c1cccc2cccnc12	7.8705704	−2.9	−2.673	−0.227
62-53-3	Nc1ccccc1	6.1394797	−3.9	−3.541	−0.359
100-61-8	CNc1ccccc1	6.5695520	−3.8	−3.326	−0.474
121-69-7	CN(C)c1ccccc1	6.8723833	−3.2	−3.174	−0.026
4904-61-4	C1\C=C/CC\C=C/CC/C=C\C1	8.0584170	−2.7	−2.579	−0.121
	Test set, n = 30				
629-59-4	CCCCCCCCCCCCCC	6.0210122	−4.3	−3.601	−0.699
110-82-7	C1CCCCC1	4.0470378	−5.3	−4.590	−0.710
591-49-1	CC=1CCCCC=1	4.9506915	−3.8	−4.137	0.337
2207-01-4	CC1CCCCC1C	4.9071824	−4.6	−4.159	−0.441
1678-91-7	CCC1CCCCC1	4.9071824	−4.3	−4.159	−0.141
67-66-3	ClC(Cl)Cl	3.9790910	−4.8	−4.624	−0.176
106-93-4	BrCCBr	4.6496160	−4.2	−4.288	0.088
106-94-5	BrCCC	3.1849526	−5.2	−5.023	−0.177
109-64-8	BrCCCBr	5.0796883	−4.2	−4.073	−0.127
78-77-3	CC(C)CBr	3.4877839	−4.9	−4.871	−0.029
507-20-0	CC(C)(C)Cl	2.6912271	−5.7	−5.270	−0.430

Table 4 (*continued*)

CAS No.	SMILES	DCW(SMILES)	$\log S_{expr}$	$\log S_{calc}$	$\log S_{expr} - \log S_{calc}$
558-17-8	CC(C)(C)I	4.4540407	−4.4	−4.386	−0.014
79-01-6	Cl\C=C(/Cl)Cl	4.8861451	−3.8	−4.170	0.370
108-88-3	Cc1ccccc1	5.7415579	−3.4	−3.741	0.341
106-42-3	Cc1ccc(C)cc1	6.0443892	−3.3	−3.589	0.289
488-23-3	Cc1ccc(C)c(C)c1C	6.7772928	−2.9	−3.221	0.321
100-41-4	CCc1ccccc1	6.1716302	−3.4	−3.525	0.125
135-98-8	CCC(C)c1ccccc1	6.9045338	−3.6	−3.158	−0.442
591-50-4	Ic1ccccc1	8.2997191	−3.5	−2.458	−1.042
541-73-1	Clc1cccc(Cl)c1	7.6350844	−3.4	−2.791	−0.609
583-53-9	Brc1ccccc1Br	9.1009570	−2.6	−2.056	−0.544
88-72-2	O=[N+]([O−])c1ccccc1C	6.2020445	−3.4	−3.510	0.110
100-44-7	ClCc1ccccc1	6.9669778	−3.4	−3.126	−0.274
90-13-1	Clc2cccc1ccccc12	9.6545989	−2.0	−1.779	−0.221
2586-62-1	Cc2ccc1ccccc1c2Br	10.7539870	−2.1	−1.228	−0.872
71-23-8	CCCO	0.9685066	−6.4	−6.134	−0.266
111-27-3	CCCCCCO	2.2587235	−5.1	−5.487	0.387
111-87-5	CCCCCCCCO	3.1188681	−5.0	−5.056	0.056
107-13-1	C=CC#N	0.6036990	−6.4	−6.317	−0.083
111-96-6	COCCOCCOC	1.6153029	−5.2	−5.810	0.610

The S is expressed in molar fraction.

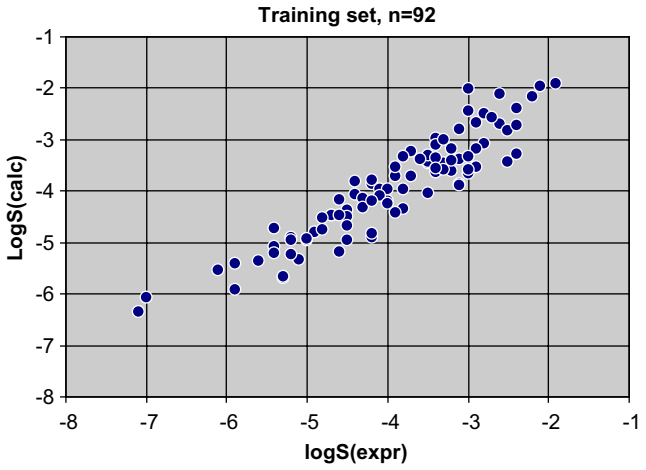

Fig. 1. Plot of experimental versus calculated solubility of fullerene C_{60} ($\log S$, S is expressed in molar fraction) for the training set.

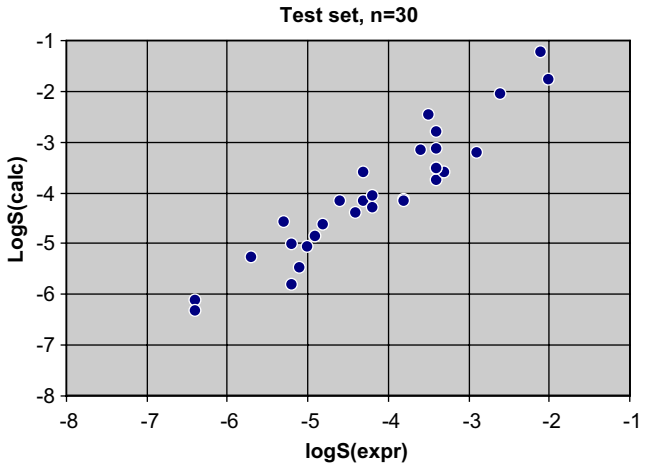

Fig. 2. Plot of experimental versus calculated solubility of fullerene C_{60} ($\log S$, S is expressed in molar fraction) for the test set.

It is also to be noted that the number of given SF_k in the training and test sets provides an important information. For example, 'C' (sp^3) appears 282 times in the training set, and 108 times in the test set, 'Cl' appears 43 times in training set and seven times in test set, whereas '#' appears only one time in both the training and the test sets. '[N^+]' and '[O^-]' appear two times in the training and one time in the test set. Apparently, one can conclude that larger numbers of SF_k in the training set are related to higher statistical significance of the SF_k for a given model.

Thus, the suggested approach allows us to perform logical and transparent analysis of relationships between the molecular structure of solvents (encoded by SMILES) and solubility of fullerene C_{60} in organic solvents ($\log S$).

4. Conclusions

In this study, the applied approach based on solely topological data provides a reliable model for fullerene C_{60} solubility. The accuracy of this model is compared with model based on quantum chemical descriptors from work [1].

Acknowledgements

The authors 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 Nano-Materials 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] H. Liu, X. Yao, R. Zhang, M. Liu, Z. Hu, B. Fan, J. Phys. Chem. B 109 (2005) 20565.
[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, E. Benfenati, Comput. Biol. Chem. 31 (2007) 57.
[6] http://www.daylight.com/.
[7] http://www.acdlabs.com/.