J Math Chem (2009) 46:1232–1251
DOI 10.1007/s10910-008-9514-0

ORIGINAL PAPER

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

Andrey A. Toropov · Alla P. Toropova · Emilio Benfenati · Danuta Leszczynska · Jerzy Leszczynski

Received: 13 November 2008 / Accepted: 4 December 2008 / Published online: 6 January 2009

Abstract Optimal descriptors calculated with International Chemical Identifier (InChI) have been used to construct one-variable model of the solubility of fullerene C_{60} in organic solvents . Attempts to calculate the model for three splits into training and test sets gave stable results. Statistical quality of the model is $n = 92$, $r^2 = 0.9447$, $Q^2 = 0.9418$, $s = 0.253$, standard deviation of error of prediction (SDEP) $= 0.258$, $F = 1,538$ (training set) and $n = 30, r^2 = 0.9398$, $R^2_{\text{pred}} = 0.9315$, $s = 0.348$, $F = 437$ (test set).

Keywords QSPR · InChI · Fullerene C_{60} solubility

1 Introduction

The selection of solvent for fullerene C_{60} can be a first step to design of new nanomaterials or new nanotechnologies. It is encourages search for theoretical methods to predict of the C_{60} solubility as a mathematical function of structure of solvent [1].

Simplified molecular input line entry system (SMILES) [2–4] and International Chemical Identifier (InChI) [5–7] are an elucidations of the molecular structure by special sequences of symbols. The number of available via Internet databases which are using mentioned representations of the molecular structure is gradually

A. A. Toropov (✉) · A. P. Toropova · E. Benfenati
Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, Milan 20156, Italy
e-mail: aatoropov@yahoo.com

D. Leszczynska · J. Leszczynski
Interdisciplinary Nanotoxicity Center, Department of Chemistry, Jackson State University, 1400 J.R. Lynch Street, P.O. Box 17910, Jackson, MS 39217, USA

increasing [8,9]. Under such circumstances the construct of the quantitative structure—property/activity relationships (QSPR/QSAR) which are based on the SMILES or InChI (i.e., directly from Internet databases) becomes very tempting.

This study has been aimed to estimate the ability of the InChI as a tool for the QSPR/QSAR analyses. It is to be noted that there are experience of the use the SMILES as the tool of the QSPR/QSAR analyses in general [10–14] and for QSPR modeling of the fullerene C_{60} in particular [15,16]. However, a description of the using the InChI in the QSPR/QSAR is still absent.

2 Method

Figure 1 shows an example of the InChI. There are three basic layers for the InChI: formula, connectivity, and hydrogen atoms.

The formula is provider of the following InChI attributes (I_k):

Br2; Br3; Br (bromine atom); C10; C11; C12; C14; C16; C2; C3; C4; C5; C6; C7; C8; C9; C; (carbon atom); Cl2; Cl3; Cl4; Cl (chlorine atom); F (fluorine atom); I (iodine atom); N (nitrogen atom); O2; O3; O (oxygen atom); S (sulphur atom);

The connectivity layer gives the following group of the I_k:

(10; (11; (14; (2; (3; (4; (5; (6; (7; (8; (9; (; ,10; ,11; ,1; ,2; ,3; ,4; ,5; ,6; ,7; ,8; ,9; −10; −11;

−12; −13; −14; −15; −1; −2; −3; −4; −5; −6; −7; −8; −9; 0; 1; 2; 3; 4; 5; 6; 7; 8; 9; c11; c1; c2; c3; c4; c6; c7; c8; c9;

Hydrogen atoms layer gives:

h1; h2; h3; h4; h5; h6; h7; h8; h9; H10; H11; H12; H14; H16; H18; H22; H26; H30; H2; H3; H4; H5; H6; H7; H8; H9; H;

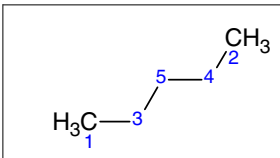
Next additional group is indicators of electronic charge and double bonds

+; −; b2;

and finally the symbol of /.

The algorithm for building of the InChI-based model is the following.

1. Definition of the split (i.e., list of the training set and list of the test set);
2. Definition of the total list of the InChI attributes;
3. Calculation (by the Monte Carlo optimization) of the correlation weights $W(I_k)$ for the InChI attributes which give maximum of the correlation coefficient (for the training set) between the DCW(InChI) and logS, C_{60} fullerene solubility. The DCW(InChI) is calculated as



InChI=1/C5H12/c1-3-5-4-2/h3-5H2,1-2H3

Description	Layers
Formula	C5H12
Connectivity	c1-3-5-4-2
Hydrogen atoms	h3-5H2,1-2H3

Fig. 1 Example of the InChI layers (pentane, C_5H_{12}, CAS 109-66-0)

$$\mathrm{DCW\ (InChI)} = \Sigma \mathrm{W\ (I_k)} \tag{1}$$

where the I_k is InChI attribute and $W(I_k)$ is the correlation weight for the I_k.

4. Having the $W(I_k)$ which produce maximum of the correlation coefficient for the training set, one can using data on the training set to calculate model of the fullerene C_{60} solubility

$$\mathrm{logS} = \mathrm{C_0} + \mathrm{C_1}\ ^{*}\ \mathrm{DCW\ (InChI)} \tag{2}$$

5. Predictive potential of the Eq. 2 one can estimate with data on the test set.

3 Results

Three splits into training and test sets have been examined (Table 1). Table 2 shows that statistical characteristics of the models are satisfactory for all splits. Table 3 shows the distribution of the InChI attributes (between the training set and test set) for three splits. One can see from Table 3 that there are variations in the distributions of the InChI attributes for these splits. However, Fig. 2 shows that models for all three splits are satisfactory ones.

Table 4 contains correlation weights $W(I_k)$ obtained by the Monte Carlo optimization. Table 5 shows an example of the DCW(InChI) calculation with Eq. 1. Model that was obtained in the first run of the Monte Carlo optimization (split 1) is the following

$$\mathrm{logs} = -7.9824\ (\pm 0.1397) + 0.3250\ (\pm 0.0010)\ ^{*}\ \mathrm{DCW\ (InChI)} \tag{3}$$

$n = 92$, $r^2 = 0.9447$, $Q^2 = 0.9418$, $s = 0.253$, SDEP=0.258, $F = 1538$ (training set)
$n = 30$, $r^2 = 0.9398$, $R^2_{\mathrm{pred}} = 0.9315$, $s = 0.348$, $F = 437$, $k = 0.9931$, $k' = 1.0031$ (test set)

The model (calculated with Eq. 3) is characterized by correlation coefficient (r), the crossvalidated coefficient Q^2 [16], predictive r^2 value (R^2_{pred}) [17], standard error of estimation (s), standard deviation of error of prediction (SDEP) [18] and slopes k and k', which according to [19] must satisfy the following conditions: $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$ Table 6 shows experimental and calculated with Eq. 3 values of fullerence C_{60} solubility.

4 Discussion

Majority of the $W(I_k)$ have similar values of the correlation weights for the three splits and for three runs of the Monte Carlo optimization (Table 4). But there are also attributes with unstable weights (i.e., there are both negative and positive values of the weights, in the Monte Carlo optimization runs). For instance, '(9', ',10', 'C4', 'h5', 'h7' (Table 4). Probably, this unstability is indicator of weak influence of relevant

Table 1 Splits into training set and test set

List of test set for split 1	List of test set for split 2	List of test set for split 3
493-01-6	111-65-9	111-65-9
542-18-7	124-18-5	124-18-5
5401-62-7	493-02-7	493-01-6
74-96-4	110-83-8	137-43-9
142-28-9	108-87-2	108-85-0
13-36-0	74-95-3	626-62-0
108-38-3	74-97-5	74-95-3
103-65-1	75-03-6	74-88-4
462-06-6	107-08-4	79-34-5
629-59-4	13-36-0	142-28-9
110-82-7	507-19-7	627-31-6
2207-01-4	108-38-3	96-11-7
106-93-4	104-51-8	13-36-0
106-94-5	108-86-1	108-38-3
109-64-8	108-37-2	526-73-8
78-77-3	120-82-1	527-53-7
507-20-0	98-95-3	103-65-1
558-17-8	100-66-3	100-47-0
108-88-3	100-39-0	90-12-0
106-42-3	71-36-3	605-02-7
488-23-3	872-50-4	71-36-3
100-41-4	91-22-5	110-02-1
541-73-1	2207-01-4	110-82-7
583-53-9	106-94-5	507-20-0
88-72-2	109-64-8	558-17-8
2586-62-1	78-77-3	79-01-6
71-23-8	507-20-0	541-73-1
111-27-3	100-41-4	583-53-9
111-87-5	111-27-3	100-44-7
107-13-1	107-13-1	71-23-8

molecular phenomenon (i.e., InChI attributes, such as, chemical element, connectivity, kind of bond, etc.) upon the C_{60} solubility.

Most probably the models obtained in the runs of the Monte Carlo optimization are based on InChI attributes, which are obeying two conditions: the first, InChI attributes which are images of molecular phenomena of considerable influence to the solubility, the second, InChI attributes which are not rare in the training set. Hence, if majority of the InChI attributes is absent in the training set (unsuccessful split) the satisfactory QSPR model becomes impossible. Thus probabilistic analysis that is represented in Table 3 can be used as a tool for definition of the applicability domain for this model:

Table 2 Statistical characteristics of the models for fullerene C_{60} solubility

Run	Training set, $N = 92$			Test set, $N = 30$		
	r^2	s	F	r^2	s	F
Split 1						
1	0.9447	0.253	1538	0.9398	0.348	437
2	0.9465	0.249	1591	0.9428	0.348	461
3	0.9477	0.246	1631	0.9390	0.349	431
Average	0.9463	0.249	1587	0.9405	0.348	443
Split 2						
1	0.9493	0.250	1687	0.8851	0.338	216
2	0.9486	0.252	1661	0.8868	0.333	219
3	0.9488	0.252	1667	0.8840	0.333	213
Average	0.9489	0.251	1671	0.8853	0.334	216
Split 3						
1	0.9503	0.239	1722	0.9393	0.290	433
2	0.9493	0.241	1686	0.9398	0.287	437
3	0.9507	0.238	1736	0.9420	0.286	455
Average	0.9501	0.239	1715	0.9404	0.288	442

substances for which should be done prediction must have InChI without of rare (in the training set) attributes.

In fact described approach is based on groups contributions. However, the groups which are extracted from the InChI have not the transparent interpretations which are typical for the Free Wilson [20] scheme, Fujita approach [21], and models based on the semi empirical topological indices [22,23]. In spite of absence of the simple interpretations of the molecular fragments encoded by InChI attributes, one can, using described algorithm, obtain reasonable prediction of the fullerene C_{60} solubility. Taking into account, clear genesis of the InChI attributes (chemical element, connectivity, bonds, charges, etc) one can extract from the model (Table 4) robust heuristic information, which probably is not less important than groups contributions.

For instance, one can see from Table 3, that InChI attributes of the '−3' and the '−7' (components of the connectivity layer, Fig. 1) have considerable prevalence in both training set and test set. From Table 4 one can detect that '−7' has stable correlation weight, whereas '−3' have both positive and negative values of the correlation weight in different runs of the Monte Carlo optimization. Under such circumstances, one can formulate the following conclusions: 1. the '−7' has an apparent influence on the C_{60} solubility (promoter of increase), and 2. the '−3' has not the influence on this parameter. Similar analysis is available for any InChI attributes from Tables 3 and 4.

Thus, the InChI-based optimal descriptors can hint that some mechanistic interpretations for the fullerene C_{60} solubility exists. In particular, presence of the '−7' in connectivity layer of InChI is indicator of increase of the solubility, whereas the presence of the '−4' in connectivity layer of InChI is indicator of decrease of the

Table 3 Distributions of the InChI attributes (between training set and test set)

I_k	Split 1		Split 2		Split 3	
	Number of I_k in training set	Number of I_k in test set	Number of I_k in training set	Number of I_k in test set	Number of I_k in training set	Number of I_k in test set
(10	2	1	2	1	2	1
(11	1	0	1	0	1	0
(14	1	0	1	0	0	1
(2	**17**	**7**	**19**	**5**	**18**	**6**
(3	**7**	**0**	**7**	**0**	**6**	**1**
(4	**5**	**1**	**5**	**1**	**3**	**3**
(5	**7**	**2**	**9**	**0**	**7**	**2**
(6	3	1	3	1	3	1
(7	**3**	**2**	**4**	**1**	**4**	**1**
(8	5	2	5	2	5	2
(9	**6**	**2**	**6**	**2**	**7**	**1**
(	**46**	**14**	**48**	**12**	**46**	**14**
+	1	0	1	0	1	0
,10	2	0	2	0	2	0
,11	1	0	1	0	1	0
,1	**49**	**21**	**52**	**18**	**55**	**15**
,2	**9**	**5**	**9**	**5**	**11**	**3**
,3	**8**	**5**	**7**	**6**	**10**	**3**
,4	3	0	3	0	3	0
,5	1	0	1	0	1	0
,6	**3**	**1**	**3**	**1**	**2**	**2**
,7	5	1	5	1	5	1
,8	**2**	**0**	**1**	**1**	**2**	**0**
,9	1	0	1	0	1	0
−10	**12**	**3**	**11**	**4**	**11**	**4**
−11	**5**	**1**	**6**	**0**	**4**	**2**
−12	**4**	**1**	**5**	**0**	**4**	**1**
−13	**1**	**1**	**2**	**0**	**1**	**1**
−14	**1**	**1**	**2**	**0**	**1**	**1**
−15	1	0	1	0	0	1
−1	**45**	**10**	**44**	**11**	**38**	**17**
−2	**76**	**27**	**77**	**26**	**78**	**25**
−3	**78**	**26**	**76**	**28**	**79**	**25**
−4	**71**	**27**	**73**	**25**	**75**	**23**
−5	**64**	**22**	**66**	**20**	**64**	**22**
−6	**56**	**18**	**57**	**17**	**58**	**16**
−7	**38**	**13**	**38**	**13**	**40**	**11**
−8	**28**	**10**	**29**	**9**	**30**	**8**
−9	**17**	**5**	**18**	**4**	**17**	**5**

Table 3 continued

I_k	Split 1		Split 2		Split 3	
	Number of I_k in training set	Number of I_k in test set	Number of I_k in training set	Number of I_k in test set	Number of I_k in training set	Number of I_k in test set
–	1	0	1	0	1	0
/	92	30	92	30	92	30
0	**6**	**2**	**8**	**0**	**6**	**2**
1	**7**	**3**	**10**	**0**	**7**	**3**
2	**10**	**2**	**11**	**1**	**9**	**3**
3	**6**	**3**	**6**	**3**	**7**	**2**
4	8	1	8	1	8	1
5	**8**	**3**	**7**	**4**	**7**	**4**
6	**9**	**2**	**8**	**3**	**7**	**4**
7	1	0	0	1	1	0
8	**4**	**3**	**7**	**0**	**6**	**1**
9	**5**	**1**	**6**	**0**	**4**	**2**
Br2	**4**	**4**	**6**	**2**	**6**	**2**
Br3	**2**	**0**	**2**	**0**	**1**	**1**
Br	**9**	**4**	**6**	**7**	**11**	**2**
C10	**8**	**2**	**7**	**3**	**7**	**3**
C11	**1**	**1**	**2**	**0**	**1**	**1**
C12	3	0	3	0	3	0
C14	0	1	1	0	1	0
C16	**1**	**0**	**1**	**0**	**0**	**1**
C2	**8**	**2**	**9**	**1**	**8**	**2**
C3	**12**	**5**	**13**	**4**	**13**	**4**
C4	**6**	**4**	**5**	**5**	**5**	**5**
C5	6	1	6	1	6	1
C6	**17**	**6**	**17**	**6**	**18**	**5**
C7	**11**	**2**	**10**	**3**	**11**	**2**
C8	**7**	**5**	**8**	**4**	**10**	**2**
C9	**5**	**1**	**5**	**1**	**4**	**2**
Cl2	**4**	**2**	**6**	**0**	**4**	**2**
Cl3	**6**	**0**	**5**	**1**	**5**	**1**
Cl4	**3**	**0**	**3**	**0**	**2**	**1**
Cl	**9**	**3**	**8**	**4**	**9**	**3**
C	**7**	**0**	**5**	**2**	**5**	**2**
F	**0**	**1**	**1**	**0**	**1**	**0**
H10	**4**	**4**	**4**	**4**	**5**	**3**
H11	**3**	**0**	**3**	**0**	**1**	**2**
H12	**10**	**2**	**12**	**0**	**8**	**4**
H14	**7**	**2**	**6**	**3**	**8**	**1**
H16	**2**	**1**	**2**	**1**	**3**	**0**

Table 3 continued

I_k	Split 1		Split 2		Split 3	
	Number of I_k in training set	Number of I_k in test set	Number of I_k in training set	Number of I_k in test set	Number of I_k in training set	Number of I_k in test set
H18	4	2	4	2	4	2
H22	**1**	**0**	**0**	**1**	**0**	**1**
H26	1	0	1	0	1	0
H30	0	1	1	0	1	0
H2	**44**	**19**	**42**	**21**	**46**	**17**
H3	**47**	**20**	**47**	**20**	**54**	**13**
H4	**6**	**3**	**8**	**1**	**6**	**3**
H5	**11**	**2**	**10**	**3**	**11**	**2**
H6	**9**	**2**	**10**	**1**	**9**	**2**
H7	**13**	**2**	**11**	**4**	**14**	**1**
H8	**3**	**2**	**4**	**1**	**4**	**1**
H9	**5**	**6**	**6**	**5**	**7**	**4**
H	**70**	**21**	**72**	**19**	**70**	**21**
I	**8**	**1**	**7**	**2**	**5**	**4**
N	**11**	**2**	**9**	**4**	**12**	**1**
O2	**2**	**1**	**2**	**1**	**3**	**0**
O3	1	0	1	0	1	0
O	**9**	**3**	**8**	**4**	**10**	**2**
S	**4**	**0**	**4**	**0**	**3**	**1**
b2	2	0	2	0	2	0
c11	1	0	1	0	1	0
c1	**57**	**22**	**57**	**22**	**62**	**17**
c2	**6**	**0**	**4**	**2**	**5**	**1**
c3	**5**	**1**	**6**	**0**	**4**	**2**
c4	**3**	**2**	**4**	**1**	**2**	**3**
c6	**1**	**1**	**2**	**0**	**1**	**1**
c7	**12**	**4**	**13**	**3**	**12**	**4**
c8	**6**	**0**	**4**	**2**	**4**	**2**
c9	1	0	1	0	1	0
h1	**41**	**9**	**38**	**12**	**35**	**15**
h2	**10**	**6**	**10**	**6**	**15**	**1**
h3	**20**	**5**	**21**	**4**	**20**	**5**
h4	7	3	7	3	7	3
h5	**4**	**3**	**6**	**1**	**4**	**3**
h6	**3**	**0**	**3**	**0**	**1**	**2**
h7	**2**	**2**	**1**	**3**	**4**	**0**
h8	2	0	2	0	2	0
h9	**1**	**2**	**2**	**1**	**2**	**1**

Attributes with the different distribution are indicated by bold

Table 4 Correlation weights for the calculation of DCW(InChI) with Eq. 1. Three runs of the Monte Carlo optimization have been carried out for each split into training set and test set

I_k	Split 1			Split 2			Split 3		
	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3
(10	1.5270741	2.1788037	1.8833852	1.5270741	2.1788037	1.8833852	1.5270741	2.1788037	1.8833852
(11	3.2919160	3.2780505	3.8743949	3.2919160	3.2780505	3.8743949	3.2919160	3.2780505	3.8743949
(14	0.8252322	1.2458653	0.9967601	0.8252322	1.2458653	0.9967601	0.8252322	1.2458653	0.9967601
(2	0.3879617	0.7426252	0.7685262	0.3879617	0.7426252	0.7685262	0.3879617	0.7426252	0.7685262
(3	1.0026027	0.9962688	1.0975908	1.0026027	0.9962688	1.0975908	1.0026027	0.9962688	1.0975908
(4	2.2770323	0.9395249	1.9467570	2.2770323	0.9395249	1.9467570	2.2770323	0.9395249	1.9467570
(5	0.5990894	1.4525841	0.9401716	0.5990894	1.4525841	0.9401716	0.5990894	1.4525841	0.9401716
(6	1.2528528	0.4390974	0.3919963	1.2528528	0.4390974	0.3919963	1.2528528	0.4390974	0.3919963
(7	−0.3037739	−0.4981416	−0.6484539	−0.3037739	−0.4981416	−0.6484539	−0.3037739	−0.4981416	−0.6484539
(8	0.9542017	0.5221816	0.9372641	0.9542017	0.5221816	0.9372641	0.9542017	0.5221816	0.9372641
(9	0.3487618	−0.0983907	0.1748857	0.3487618	−0.0983907	0.1748857	0.3487618	−0.0983907	0.1748857
(	0.3481656	0.3989177	0.2515877	0.3481656	0.3989177	0.2515877	0.3481656	0.3989177	0.2515877
+	2.1391551	1.4384265	1.8402422	2.1391551	1.4384265	1.8402422	2.1391551	1.4384265	1.8402422
,10	1.2962282	1.2988873	0.9726131	1.2962282	1.2988873	0.9726131	1.2962282	1.2988873	0.9726131
,11	0.0210853	−0.4256827	−0.0456591	0.0210853	−0.4256827	−0.0456591	0.0210853	−0.4256827	−0.0456591
,1	0.4421542	0.3610609	0.1436358	0.4421542	0.3610609	0.1436358	0.4421542	0.3610609	0.1436358
,2	0.5530336	0.1903704	0.5798204	0.5530336	0.1903704	0.5798204	0.5530336	0.1903704	0.5798204
,3	−0.2543010	−0.3470018	0.0502112	−0.2543010	−0.3470018	0.0502112	−0.2543010	−0.3470018	0.0502112
,4	−0.5041659	−0.4953500	−0.6534183	−0.5041659	−0.4953500	−0.6534183	−0.5041659	−0.4953500	−0.6534183
,5	1.0278815	0.1422890	−0.0984711	1.0278815	0.1422890	−0.0984711	1.0278815	0.1422890	−0.0984711
,6	2.6991359	2.5714537	2.4930555	2.6991359	2.5714537	2.4930555	2.6991359	2.5714537	2.4930555
,7	0.7112837	0.3924745	1.0799710	0.7112837	0.3924745	1.0799710	0.7112837	0.3924745	1.0799710

Table 4 continued

I_k	Split 1			Split 2			Split 3		
	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3
,8	0.4028429	0.3028112	0.3026085	0.4028429	0.3028112	0.3026085	0.4028429	0.3028112	0.3026085
,9	−0.3255445	−0.4531144	−0.4737900	−0.3255445	−0.4531144	−0.4737900	−0.3255445	−0.4531144	−0.4737900
−10	0.3820470	0.5122282	0.4156119	0.3820470	0.5122282	0.4156119	0.3820470	0.5122282	0.4156119
−11	1.3842584	2.1538679	1.3549628	1.3842584	2.1538679	1.3549628	1.3842584	2.1538679	1.3549628
−12	0.5227524	0.1170791	0.0540155	0.5227524	0.1170791	0.0540155	0.5227524	0.1170791	0.0540155
−13	0.5545050	0.3838421	0.4715506	0.5545050	0.3838421	0.4715506	0.5545050	0.3838421	0.4715506
−14	0.5670429	0.3212755	0.3816642	0.5670429	0.3212755	0.3816642	0.5670429	0.3212755	0.3816642
−15	0.3110355	1.0219631	1.0029557	0.3110355	1.0219631	1.0029557	0.3110355	1.0219631	1.0029557
−1	1.1005957	1.4826400	1.4238212	1.1005957	1.4826400	1.4238212	1.1005957	1.4826400	1.4238212
−2	0.6093029	0.4650875	0.4350668	0.6093029	0.4650875	0.4350668	0.6093029	0.4650875	0.4350668
−3	0.0975796	−0.1905156	0.3143514	0.0975796	−0.1905156	0.3143514	0.0975796	−0.1905156	0.3143514
−4	0.7174808	0.6283553	0.8030665	0.7174808	0.6283553	0.8030665	0.7174808	0.6283553	0.8030665
−5	0.7976968	1.3066616	0.9031482	0.7976968	1.3066616	0.9031482	0.7976968	1.3066616	0.9031482
−6	1.1041909	1.0514836	0.8644455	1.1041909	1.0514836	0.8644455	1.1041909	1.0514836	0.8644455
−7	0.7717947	0.7903351	0.9184101	0.7717947	0.7903351	0.9184101	0.7717947	0.7903351	0.9184101
−8	0.4326739	0.4772444	0.6054839	0.4326739	0.4772444	0.6054839	0.4326739	0.4772444	0.6054839
−9	0.3285966	0.2019516	0.0732852	0.3285966	0.2019516	0.0732852	0.3285966	0.2019516	0.0732852
−	−0.2705353	−0.2129493	−0.2721493	−0.2705353	−0.2129493	−0.2721493	−0.2705353	−0.2129493	−0.2721493
/	−0.5043203	−0.4495929	−0.4201719	−0.5043203	−0.4495929	−0.4201719	−0.5043203	−0.4495929	−0.4201719
0	0.5360181	0.3992264	0.2022387	0.5360181	0.3992264	0.2022387	0.5360181	0.3992264	0.2022387
1	−0.1540216	−0.1031449	0.3120346	−0.1540216	−0.1031449	0.3120346	−0.1540216	−0.1031449	0.3120346
2	1.9668728	2.0337713	1.4253518	1.9668728	2.0337713	1.4253518	1.9668728	2.0337713	1.4253518
3	1.5718281	0.7396442	1.3417383	1.5718281	0.7396442	1.3417383	1.5718281	0.7396442	1.3417383

Table 4 continued

I_k	Split 1			Split 2			Split 3		
	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3
4	−0.4958935	−0.4986342	−0.7030132	−0.4958935	−0.4986342	−0.7030132	−0.4958935	−0.4986342	−0.7030132
5	1.7659343	2.2652996	1.7982545	1.7659343	2.2652996	1.7982545	1.7659343	2.2652996	1.7982545
6	1.3737633	1.1099373	1.0669446	1.3737633	1.1099373	1.0669446	1.3737633	1.1099373	1.0669446
7	1.0326661	1.0034053	2.5295223	1.0326661	1.0034053	2.5295223	1.0326661	1.0034053	2.5295223
8	0.9474995	0.3993267	0.4239624	0.9474995	0.3993267	0.4239624	0.9474995	0.3993267	0.4239624
9	2.3759176	2.2490171	1.9147497	2.3759176	2.2490171	1.9147497	2.3759176	2.2490171	1.9147497
Br2	3.8019073	3.2275120	3.7088822	3.8019073	3.2275120	3.7088822	3.8019073	3.2275120	3.7088822
Br3	3.7245714	3.9958343	4.4043409	3.7245714	3.9958343	4.4043409	3.7245714	3.9958343	4.4043409
Br	2.0913214	1.9960972	1.9100078	2.0913214	1.9960972	1.9100078	2.0913214	1.9960972	1.9100078
C10	1.1477245	1.8952252	1.3034189	1.1477245	1.8952252	1.3034189	1.1477245	1.8952252	1.3034189
C11	1.5780321	1.4089731	0.3219107	1.5780321	1.4089731	0.3219107	1.5780321	1.4089731	0.3219107
C12	1.1711630	1.5134547	1.5836869	1.1711630	1.5134547	1.5836869	1.1711630	1.5134547	1.5836869
C14	1.0004565	0.6024608	0.5888972	1.0004565	0.6024608	0.5888972	1.0004565	0.6024608	0.5888972
C16	0.8497236	1.3649242	1.0015640	0.8497236	1.3649242	1.0015640	0.8497236	1.3649242	1.0015640
C2	0.3955939	0.6716089	0.8193777	0.3955939	0.6716089	0.8193777	0.3955939	0.6716089	0.8193777
C3	0.1496663	0.0247695	−0.0874478	0.1496663	0.0247695	−0.0874478	0.1496663	0.0247695	−0.0874478
C4	−0.4036247	0.1343727	−0.3009498	−0.4036247	0.1343727	−0.3009498	−0.4036247	0.1343727	−0.3009498
C5	2.0516145	1.6207144	2.1357674	2.0516145	1.6207144	2.1357674	2.0516145	1.6207144	2.1357674
C6	1.3298226	1.3609822	1.7744324	1.3298226	1.3609822	1.7744324	1.3298226	1.3609822	1.7744324
C7	1.1455106	0.9966486	1.3656799	1.1455106	0.9966486	1.3656799	1.1455106	0.9966486	1.3656799
C8	1.9751242	2.2086462	1.6572894	1.9751242	2.2086462	1.6572894	1.9751242	2.2086462	1.6572894
C9	1.6399979	1.8658547	1.6284945	1.6399979	1.8658547	1.6284945	1.6399979	1.8658547	1.6284945
Cl2	2.7468193	2.0490346	2.7842867	2.7468193	2.0490346	2.7842867	2.7468193	2.0490346	2.7842867

Table 4 continued

I_k	Split 1			Split 2			Split 3		
	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3
Cl3	0.7253160	0.8507849	1.3844930	0.7253160	0.8507849	1.3844930	0.7253160	0.8507849	1.3844930
Cl4	2.5466320	2.1983564	2.9944248	2.5466320	2.1983564	2.9944248	2.5466320	2.1983564	2.9944248
Cl	1.5699113	1.6369047	1.8594514	1.5699113	1.6369047	1.8594514	1.5699113	1.6369047	1.8594514
C	3.3849868	3.6285473	2.9510677	3.3849868	3.6285473	2.9510677	3.3849868	3.6285473	2.9510677
F	1.0009245	−0.1990012	−0.1014541	1.0009245	−0.1990012	−0.1014541	1.0009245	−0.1990012	−0.1014541
H10	2.6329922	2.4506166	2.6328200	2.6329922	2.4506166	2.6328200	2.6329922	2.4506166	2.6328200
H11	3.0799885	2.7513574	2.9643554	3.0799885	2.7513574	2.9643554	3.0799885	2.7513574	2.9643554
H12	−0.1385480	0.0000535	−0.3490188	−0.1385480	0.0000535	−0.3490188	−0.1385480	0.0000535	−0.3490188
H14	0.7671328	0.5171750	0.6822775	0.7671328	0.5171750	0.6822775	0.7671328	0.5171750	0.6822775
H16	0.1458755	0.3268843	0.3098475	0.1458755	0.3268843	0.3098475	0.1458755	0.3268843	0.3098475
H18	−0.5036017	−0.5020824	−0.7037268	−0.5036017	−0.5020824	−0.7037268	−0.5036017	−0.5020824	−0.7037268
H22	1.3480085	1.0029202	1.0022537	1.3480085	1.0029202	1.0022537	1.3480085	1.0029202	1.0022537
H26	3.0207818	2.9541108	3.2672237	3.0207818	2.9541108	3.2672237	3.0207818	2.9541108	3.2672237
H30	1.0044798	0.3953269	0.6964517	1.0044798	0.3953269	0.6964517	1.0044798	0.3953269	0.6964517
H2	−0.4992814	−0.4029158	−0.5290658	−0.4992814	−0.4029158	−0.5290658	−0.4992814	−0.4029158	−0.5290658
H3	1.1073621	1.2535694	1.0435735	1.1073621	1.2535694	1.0435735	1.1073621	1.2535694	1.0435735
H4	1.3804232	1.8301197	1.6023455	1.3804232	1.8301197	1.6023455	1.3804232	1.8301197	1.6023455
H5	1.3904234	1.3647806	1.6965995	1.3904234	1.3647806	1.6965995	1.3904234	1.3647806	1.6965995
H6	0.5869020	1.0084404	0.9499241	0.5869020	1.0084404	0.9499241	0.5869020	1.0084404	0.9499241
H7	1.0354609	1.1463063	1.4766225	1.0354609	1.1463063	1.4766225	1.0354609	1.1463063	1.4766225
H8	2.0070304	1.6402355	1.8982529	2.0070304	1.6402355	1.8982529	2.0070304	1.6402355	1.8982529
H9	0.6203625	0.6979560	0.4412148	0.6203625	0.6979560	0.4412148	0.6203625	0.6979560	0.4412148
H	1.0593655	1.0991846	0.8394167	1.0593655	1.0991846	0.8394167	1.0593655	1.0991846	0.8394167
I	3.5669401	3.5878484	3.3416072	3.5669401	3.5878484	3.3416072	3.5669401	3.5878484	3.3416072
N	−0.3569019	−0.1530156	−0.4217054	−0.3569019	−0.1530156	−0.4217054	−0.3569019	−0.1530156	−0.4217054

Table 4 continued

I_k	Split 1			Split 2			Split 3		
	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3	$CW(I_k)$ in run 1	$CW(I_k)$ in run 2	$CW(I_k)$ in run 3
O2	−0.4953848	−0.5034510	−0.6116875	−0.4953848	−0.5034510	−0.6116875	−0.4953848	−0.5034510	−0.6116875
O3	−0.5007671	−0.4990670	−0.7029023	−0.5007671	−0.4990670	−0.7029023	−0.5007671	−0.4990670	−0.7029023
O	−0.4955365	−0.5030371	−0.5965571	−0.4955365	−0.5030371	−0.5965571	−0.4955365	−0.5030371	−0.5965571
S	2.1774747	1.9079441	1.6544812	2.1774747	1.9079441	1.6544812	2.1774747	1.9079441	1.6544812
b2	−0.3299279	−0.1290885	−0.4526957	−0.3299279	−0.1290885	−0.4526957	−0.3299279	−0.1290885	−0.4526957
c11	3.2951396	3.2955746	2.5008763	3.2951396	3.2955746	2.5008763	3.2951396	3.2955746	2.5008763
c1	0.9127424	0.7843666	1.4549991	0.9127424	0.7843666	1.4549991	0.9127424	0.7843666	1.4549991
c2	1.9876453	1.8024339	2.8112117	1.9876453	1.8024339	2.8112117	1.9876453	1.8024339	2.8112117
c3	1.8313005	2.6269393	2.2891913	1.8313005	2.6269393	2.2891913	1.8313005	2.6269393	2.2891913
c4	3.7993774	3.7984603	3.8798531	3.7993774	3.7984603	3.8798531	3.7993774	3.7984603	3.8798531
c6	2.3155141	1.6330466	2.5003741	2.3155141	1.6330466	2.5003741	2.3155141	1.6330466	2.5003741
c7	1.3361855	1.2496794	1.0997710	1.3361855	1.2496794	1.0997710	1.3361855	1.2496794	1.0997710
c8	0.0362480	0.2028330	0.2949905	0.0362480	0.2028330	0.2949905	0.0362480	0.2028330	0.2949905
c9	2.7839135	2.5034572	2.3355114	2.7839135	2.5034572	2.3355114	2.7839135	2.5034572	2.3355114
h1	2.3256350	2.1160321	2.0002300	2.3256350	2.1160321	2.0002300	2.3256350	2.1160321	2.0002300
h2	3.0168525	3.3995456	2.9815680	3.0168525	3.3995456	2.9815680	3.0168525	3.3995456	2.9815680
h3	0.4292022	0.8102634	1.0284866	0.4292022	0.8102634	1.0284866	0.4292022	0.8102634	1.0284866
h4	1.5358137	1.3768992	1.6861034	1.5358137	1.3768992	1.6861034	1.5358137	1.3768992	1.6861034
h5	−0.1915543	0.6461483	0.3268447	−0.1915543	0.6461483	0.3268447	−0.1915543	0.6461483	0.3268447
h6	0.0382853	0.3619316	1.0424572	0.0382853	0.3619316	1.0424572	0.0382853	0.3619316	1.0424572
h7	−0.4963374	−0.4961533	0.0281538	−0.4963374	−0.4961533	0.0281538	−0.4963374	−0.4961533	0.0281538
h8	−0.0921495	0.0774660	0.1659468	−0.0921495	0.0774660	0.1659468	−0.0921495	0.0774660	0.1659468
h9	2.0723134	1.0985338	1.6090366	2.0723134	1.0985338	1.6090366	2.0723134	1.0985338	1.6090366

Table 5 Example of the DCW calculation with correlation weights for split 1, which have been obtained in the first run of the Monte Carlo optimization:CAS = 109-66-0; "InChI = 1/C5H12/c1-3-5-4-2/h3-5H2,1-2H3"; DCW(InChI) = 6.9256652

I_k	$CW(I_k)$ in run 1, split 1
C5	2.0516145
H12	−0.1385480
/	−0.5043203
c1	0.9127424
−3	0.0975796
−5	0.7976968
−4	0.7174808
−2	0.6093029
/	−0.5043203
h3	0.4292022
−5	0.7976968
H2	−0.4992814
,1	0.4421542
−2	0.6093029
H3	1.1073621

Table 6 Experimental and calculated with Eq. 3 values of fullerence C_{60} Solubility (Split 1, run 1)

CAS	InChI	DCW(InChI)	logS Expr	logS Calc
Training set				
109-66-0	InChI=1/C5H12/c1-3-5-4-2/h3-5H2,1-2H3	6.9256652	−6.1	−5.732
110-54-3	InChI=1/C6H14/c1-3-5-6-4-2/h3-6H2,1-2H3	8.5202391	−5.1	−5.213
111-65-9	InChI=1/C8H18/c1-3-5-7-8-6-4-2/h3-8H2,1-2H3	8.4277578	−5.2	−5.243
26635-64-3	InChI=1/C8H18/c1-4-5-6-7-8(2)3/h8H,4-7H2,1-3H3	9.8900761	−5.2	−4.768
124-18-5	InChI=1/C10H22/c1-3-5-7-9-10-8-6-4-2/h3-10H2,1-2H3	10.1119850	−4.7	−4.696
112-40-3	InChI=1/C12H26/c1-3-5-7-9-11-12-10-8-6-4-2/h3-12H2,1-2H3	13.8559130	−3.5	−3.479
493-02-7	InChI=1/C10H18/c1-2-6-10-8-4-3-7-9(10)5-1/h9-10H,1-8H2	13.6229333	−3.5	−3.555
137-43-9	InChI=1/C5H9Br/c6-5-3-1-2-4-5/h5H,1-4H2	11.7186893	−4.2	−4.174
108-85-0	InChI=1/C6H11Br/c7-6-4-2-1-3-5-6/h6H,1-5H2	14.1979354	−3.4	−3.368
626-62-0	InChI=1/C6H11I/c7-6-4-2-1-3-5-6/h6H,1-5H2	15.6735541	−2.8	−2.888
110-83-8	InChI=1/C6H10/c1-2-4-6-5-3-1/h1-2H,3-6H2	12.6386752	−3.8	−3.875
108-87-2	InChI=1/C7H14/c1-7-5-3-2-4-6-7/h7H,2-6H2,1H3	9.9570731	−4.5	−4.746
6876-23-9	InChI=1/C8H16/c1-7-5-3-4-6-8(7)2/h7-8H,3-6H2,1-2H3	11.4629124	−4.6	−4.257

Table 6 continued

CAS	InChI	DCW(InChI)	logS Expr	logS Calc
75-09-2	InChI=1/CH2Cl2/c2-1-3/h1H2	9.6360583	−4.6	−4.851
56-23-5	InChI=1/CCl4/c2-1(3,4)5	11.1280762	−4.4	−4.366
74-95-3	InChI=1/CH2Br2/c2-1-3/h1H2	10.6911463	−4.5	−4.508
75-25-2	InChI=1/CHBr3/c2-1(3)4/h1H	14.4883994	−3.2	−3.274
74-88-4	InChI=1/CH3I/c1-2/h1H3	12.0056908	−4.2	−4.081
74-97-5	InChI=1/CH2BrCl/c2-1-3/h1H2	10.5504717	−4.2	−4.553
75-03-6	InChI=1/C2H5I/c1-2-3/h2H2,1H3	10.0310291	−4.5	−4.722
79-34-5	InChI=1/C2H2Cl4/c3-1(4)2(5)6/h1-2H	15.2735925	−3.1	−3.018
107-06-2	InChI=1/C2H4Cl2/c3-1-2-4/h1-2H2	10.2085322	−5.0	−4.665
71-55-6	InChI=1/C2H3Cl3/c1-2(3,4)5/h1H3	8.7872105	−4.7	−5.127
540-54-5	InChI=1/C3H7Cl/c1-2-3-4/h2-3H2,1H3	8.2481706	−5.6	−5.302
107-08-4	InChI=1/C3H7I/c1-2-3-4/h2-3H2,1H3	10.2451994	−4.6	−4.653
75-29-6	InChI=1/C3H7Cl/c1-3(2)4/h3H,1-2H3	6.6443406	−5.9	−5.823
75-26-3	InChI=1/C3H7Br/c1-3(2)4/h3H,1-2H3	7.1657507	−5.4	−5.654
75-30-9	InChI=1/C3H7I/c1-3(2)4/h3H,1-2H3	8.6413694	−4.8	−5.174
78-87-5	InChI=1/C3H6Cl2/c1-3(5)2-4/h3H,2H2,1H3	10.2085138	−4.9	−4.665
78-75-1	InChI=1/C3H6Br2/c1-3(5)2-4/h3H,2H2,1H3	11.2636018	−4.3	−4.322
627-31-6	InChI=1/C3H6I2/c4-2-1-3-5/h1-3H2	13.5902262	−3.4	−3.566
96-11-7	InChI=1/C3H5Br3/c4-1-3(6)2-5/h3H,1-2H2	15.6599046	−2.9	−2.893
96-18-4	InChI=1/C3H5Cl3/c4-1-3(6)2-5/h3H,1-2H2	12.6606492	−4.0	−3.868
513-38-2	InChI=1/C4H9I/c1-4(2)3-5/h4H,3H2,1-2H3	11.5113287	−4.3	−4.241
507-19-7	InChI=1/C4H9Br/c1-4(2,3)5/h1-3H3	8.7079791	−5.0	−5.152
540-49-8	InChI=1/C2H2Br2/c3-1-2-4/h1-2H/b2-1+	13.3480651	−3.7	−3.644
127-18-4	InChI=1/C2Cl4/c3-1(4)2(5)6	12.2828908	−3.8	−3.990
513-37-1	InChI=1/C4H7Cl/c1-4(2)3-5/h3H,1-2H3	9.5763692	−4.5	−4.870
71-43-2	InChI=1/C6H6/c1-2-4-6-5-3-1/h1-6H	10.7368645	−4.0	−4.493
95-47-6	InChI=1/C8H10/c1-7-5-3-4-6-8(7)2/h3-6H,1-2H3	15.1964772	−2.9	−3.044
526-73-8	InChI=1/C9H12/c1-7-5-4-6-8(2)9(7)3/h4-6H,1-3H3	15.3041195	−3.1	−3.009
95-63-6	InChI=1/C9H12/c1-7-4-5-8(2)9(3)6-7/h4-6H,1-3H3	16.0800351	−2.5	−2.756
108-67-8	InChI=1/C9H12/c1-7-4-8(2)6-9(3)5-7/h4-6H,1-3H3	15.0009516	−3.5	−3.107
527-53-7	InChI=1/C10H14/c1-7-5-8(2)10(4)9(3)6-7/h5-6H,1-4H3	17.6756893	−2.4	−2.238
119-64-2	InChI=1/C10H12/c1-2-6-10-8-4-3-7-9(10)5-1/h1-2,5-6H,3-4,7-8H2	17.3329462	−2.5	−2.349

Table 6 continued

CAS	InChI	DCW(InChI)	logS Expr	logS Calc
98-82-8	InChI=1/C9H12/c1-8(2)9-6-4-3-5-7-9/h3-8H,1-2H3	12.8476707	−3.6	−3.807
104-51-8	InChI=1/C10H14/c1-2-3-7-10-8-5-4-6-9-10/h4-6,8-9H,2-3,7H2,1H3	14.2853107	−3.4	−3.340
98-06-6	InChI=1/C10H14/c1-10(2,3)9-7-5-4-6-8-9/h4-8H,1-3H3	13.8861327	−3.7	−3.469
108-90-7	InChI=1/C6H5Cl/c7-6-4-2-1-3-5-6/h1-5H	14.3314371	−3.0	−3.325
108-86-1	InChI=1/C6H5Br/c7-6-4-2-1-3-5-6/h1-5H	14.8528472	−3.3	−3.155
95-50-1	InChI=1/C6H4Cl2/c7-5-3-1-2-4-6(5)8/h1-4H	16.2086925	−2.4	−2.715
108-36-1	InChI=1/C6H4Br2/c7-5-2-1-3-6(8)4-5/h1-4H	16.2557158	−2.6	−2.699
694-80-4	InChI=1/C6H4BrCl/c7-5-3-1-2-4-6(5)8/h1-4H	17.1231059	−2.4	−2.417
108-37-2	InChI=1/C6H4BrCl/c7-5-2-1-3-6(8)4-5/h1-4H	16.1150412	−3.0	−2.745
120-82-1	InChI=1/C6H3Cl3/c7-4-1-2-5(8)6(9)3-4/h1-3H	15.8600688	−2.8	−2.828
100-42-5	InChI=1/C8H8/c1-2-8-6-4-3-5-7-8/h2-7H,1H2	13.6405354	−3.2	−3.549
98-95-3	InChI=1/C6H5NO2/c8-7(9)6-4-2-1-3-5-6/h1-5H	12.3475961	−3.9	−3.969
100-47-0	InChI=1/C7H5N/c8-6-7-4-2-1-3-5-7/h1-5H	11.3597729	−4.2	−4.290
100-66-3	InChI=1/C7H8O/c1-8-7-5-3-2-4-6-7/h2-6H,1H3	14.5935458	−3.1	−3.239
100-52-7	InChI=1/C7H6O/c8-6-7-4-2-1-3-5-7/h1-6H	10.7241110	−4.2	−4.497
103-71-9	InChI=1/C7H5NO/c9-6-8-7-4-2-1-3-5-7/h1-5H	14.0445758	−3.4	−3.418
99-08-1	InChI=1/C7H7NO2/c1-6-3-2-4-7(5-6)8(9)10/h2-5H,1H3	15.0344359	−3.4	−3.096
108-98-5	InChI=1/C6H6S/c7-6-4-2-1-3-5-6/h1-5,7H	14.8467628	−3.0	−3.157
100-39-0	InChI=1/C7H7Br/c8-6-7-4-2-1-3-5-7/h1-5H,6H2	15.6528882	−3.1	−2.895
30583-33-6	InChI=1/C7H5Cl3/c8-7(9,10)6-4-2-1-3-5-6/h1-5H	15.0371150	−3.0	−3.095
90-12-0	InChI=1/C11H10/c1-9-5-4-7-10-6-2-3-8-11(9)10/h2-8H,1H3	17.8780798	−2.2	−2.172
28804-88-8	InChI=1/C12H12/c1-9-7-8-11-5-3-4-6-12(11)10(9)2/h3-8H,1-2H3	17.8596800	−2.1	−2.178
605-02-7	InChI=1/C16H12/c1-2-7-13(8-3-1)16-12-6-10-14-9-4-5-11-15(14)16/h1-12H	18.6871580	−1.9	−1.909
64-17-5	InChI=1/C2H6O/c1-2-3/h3H,2H2,1H3	4.1897799	−7.1	−6.621
71-36-3	InChI=1/C4H10O/c1-2-3-4-5/h5H,2-4H2,1H3	7.0485534	−5.9	−5.692
71-41-0	InChI=1/C5H12O/c1-2-3-4-5-6/h6H,2-5H2,1H3	8.1464989	−5.3	−5.335

Table 6 continued

CAS	InChI	DCW(InChI)	logS Expr	logS Calc
67-64-1	InChI=1/C3H6O/c1-3(2)4/h1-2H3	4.5252470	−7.0	−6.512
68-12-2	InChI=1/C3H7NO/c1-4(2)3-5/h3H,1-2H3	7.7073105	−5.3	−5.478
110-01-0	InChI=1/C4H8S/c1-2-4-5-3-1/h1-4H2	9.5514724	−5.4	−4.878
110-02-1	InChI=1/C4H4S/c1-2-4-5-3-1/h1-4H	10.4835121	−4.4	−4.575
554-14-3	InChI=1/C5H6S/c1-5-3-2-4-6-5/h2-4H,1H3	15.1872559	−3.0	−3.047
872-50-4	InChI=1/C5H9NO/c1-6-4-2-3-5(6)7/h2-4H2,1H3	12.4681441	−3.9	−3.930
110-86-1	InChI=1/C5H5N/c1-2-4-6-5-3-1/h1-5H	11.5987818	−4.0	−4.213
91-22-5	InChI=1/C9H7N/c1-2-6-9-8(4-1)5-3-7-10-9/h1-7H	15.9259640	−2.9	−2.806
62-53-3	InChI=1/C6H7N/c7-6-4-2-1-3-5-6/h1-5H,7H2	12.2616637	−3.9	−3.997
100-61-8	InChI=1/C7H9N/c1-8-7-5-3-2-4-6-7/h2-6,8H,1H3	13.7483554	−3.8	−3.514
121-69-7	InChI=1/C8H11N/c1-9(2)8-6-4-3-5-7-8/h3-7H,1-2H3	14.9551343	−3.2	−3.122
4904-61-4	InChI=1/C12H18/c1-2-4-6-8-10-12-11-9-7-5-3-1/h1-2,7-10H,3-6,11-12H2/b2-1-,9-7-,10-8-	16.4326070	−2.7	−2.642
591-49-1	InChI=1/C7H12/c1-7-5-3-2-4-6-7/h5H,2-4,6H2,1H3	11.6686012	−3.8	−4.190
1678-91-7	InChI=1/C8H16/c1-2-8-6-4-3-5-7-8/h8H,2-7H2,1H3	10.3307742	−4.3	−4.625
67-66-3	InChI=1/CHCl3/c2-1(3)4/h1H	11.4891440	−4.8	−4.248
79-01-6	InChI=1/C2HCl3/c3-1-2(4)5/h1H	12.4889666	−3.8	−3.923
135-98-8	InChI=1/C10H14/c1-3-9(2)10-7-5-4-6-8-10/h4-9H,3H2,1-2H3	11.8981558	−3.6	−4.115
591-50-4	InChI=1/C6H5I/c7-6-4-2-1-3-5-6/h1-5H	16.3284659	−3.5	−2.676
100-44-7	InChI=1/C7H7Cl/c8-6-7-4-2-1-3-5-7/h1-5H,6H2	15.1314781	−3.4	−3.065
90-13-1	InChI=1/C10H7Cl/c11-10-7-3-5-8-4-1-2-6-9(8)10/h1-7H	18.2227136	−2.0	−2.060
111-96-6	InChI=1/C6H14O3/c1-7-3-5-9-6-4-8-2/h3-6H2,1-2H3	9.5525372	−5.2	−4.878
Test set				
493-01-6	InChI=1/C10H18/c1-2-6-10-8-4-3-7-9(10)5-1/h9-10H,1-8H2	13.6229333	−3.3	−3.555
542-18-7	InChI=1/C5H9Cl/c6-5-3-1-2-4-5/h5H,1-4H2	11.1972792	−4.1	−4.343
5401-62-7	InChI=1/C6H10Br2/c7-5-3-1-2-4-6(5)8/h5-6H,1-4H2	17.0462239	−2.6	−2.442
74-96-4	InChI=1/C2H5Br/c1-2-3/h2H2,1H3	8.5554104	−5.2	−5.202
142-28-9	InChI=1/C3H6Cl2/c4-2-1-3-5/h1-3H2	10.8032326	−4.8	−4.471
13-36-0	InChI=1/C4H9Cl/c1-4(2)3-5/h4H,3H2,1-2H3	9.5142999	−5.4	−4.890

Table 6 continued

CAS	InChI	DCW(InChI)	logS Expr	logS Calc
108-38-3	InChI=1/C8H10/c1-7-4-3-5-8(2)6-7/h3-6H,1-2H3	14.9627071	−3.3	−3.120
103-65-1	InChI=1/C9H12/c1-2-6-9-7-4-3-5-8-9/h3-5,7-8H,2,6H2,1H3	14.3260910	−3.5	−3.326
462-06-6	InChI=1/C6H5F/c7-6-4-2-1-3-5-6/h1-5H	13.7624503	−4.1	−3.510
629-59-4	InChI=1/C14H30/c1-3-5-7-9-11-13-14-12-10-8-6-4-2/h3-14H2,1-2H3	12.8347429	−4.3	−3.811
110-82-7	InChI=1/C6H12/c1-2-4-6-5-3-1/h1-6H2	8.4527676	−5.3	−5.235
2207-01-4	InChI=1/C8H16/c1-7-5-3-4-6-8(7)2/h7-8H,3-6H2,1-2H3	11.4629124	−4.6	−4.257
106-93-4	InChI=1/C2H4Br2/c3-1-2-4/h1-2H2	11.2636202	−4.2	−4.322
106-94-5	InChI=1/C3H7Br/c1-2-3-4/h2-3H2,1H3	8.7695807	−5.2	−5.132
109-64-8	InChI=1/C3H6Br2/c4-2-1-3-5/h1-3H2	11.8583206	−4.2	−4.128
78-77-3	InChI=1/C4H9Br/c1-4(2)3-5/h4H,3H2,1-2H3	10.0357100	−4.9	−4.721
507-20-0	InChI=1/C4H9Cl/c1-4(2,3)5/h1-3H3	8.1865690	−5.7	−5.322
558-17-8	InChI=1/C4H9I/c1-4(2,3)5/h1-3H3	10.1835978	−4.4	−4.673
108-88-3	InChI=1/C7H8/c1-7-5-3-2-4-6-7/h2-6H,1H3	14.6564084	−3.4	−3.219
106-42-3	InChI=1/C8H10/c1-7-3-5-8(2)6-4-7/h3-6H,1-2H3	14.9627071	−3.3	−3.120
488-23-3	InChI=1/C10H14/c1-7-5-6-8(2)10(4)9(7)3/h5-6H,1-4H3	16.8997737	−2.9	−2.490
100-41-4	InChI=1/C8H10/c1-2-8-6-4-3-5-7-8/h3-7H,2H2,1H3	13.3392426	−3.4	−3.647
541-73-1	InChI=1/C6H4Cl2/c7-5-2-1-3-6(8)4-5/h1-4H	15.2006278	−3.4	−3.042
583-53-9	InChI=1/C6H4Br2/c7-5-3-1-2-4-6(5)8/h1-4H	17.2637805	−2.6	−2.372
88-72-2	InChI=1/C7H7NO2/c1-6-4-2-3-5-7(6)8(9)10/h2-5H,1H3	15.3817052	−3.4	−2.983
2586-62-1	InChI=1/C11H9Br/c1-8-6-7-9-4-2-3-5-10(9)11(8)12/h2-7H,1H3	19.3368127	−2.1	−1.698
71-23-8	InChI=1/C3H8O/c1-2-3-4/h4H,2-3H2,1H3	7.2856526	−6.4	−5.615
111-27-3	InChI=1/C6H14O/c1-2-3-4-5-6-7/h7H,2-6H2,1H3	8.8740539	−5.1	−5.098
111-87-5	InChI=1/C8H18O/c1-2-3-4-5-6-7-8-9/h9H,2-8H2,1H3	10.9070253	−5.0	−4.438
107-13-1	InChI=1/C3H3N/c1-2-3-4/h2H,1H2	6.2476824	−6.4	−5.952

solubility (Tables 3 and 4). Of course, these hypothesises are probabilistic, but they are formulated on basis of data, that has been reproduced for series of the Monte Carlo optimization runs (Fig. 2).

Finally, described InChI-based models are better than models based on optimal descriptors which are calculating with SMILES [16].

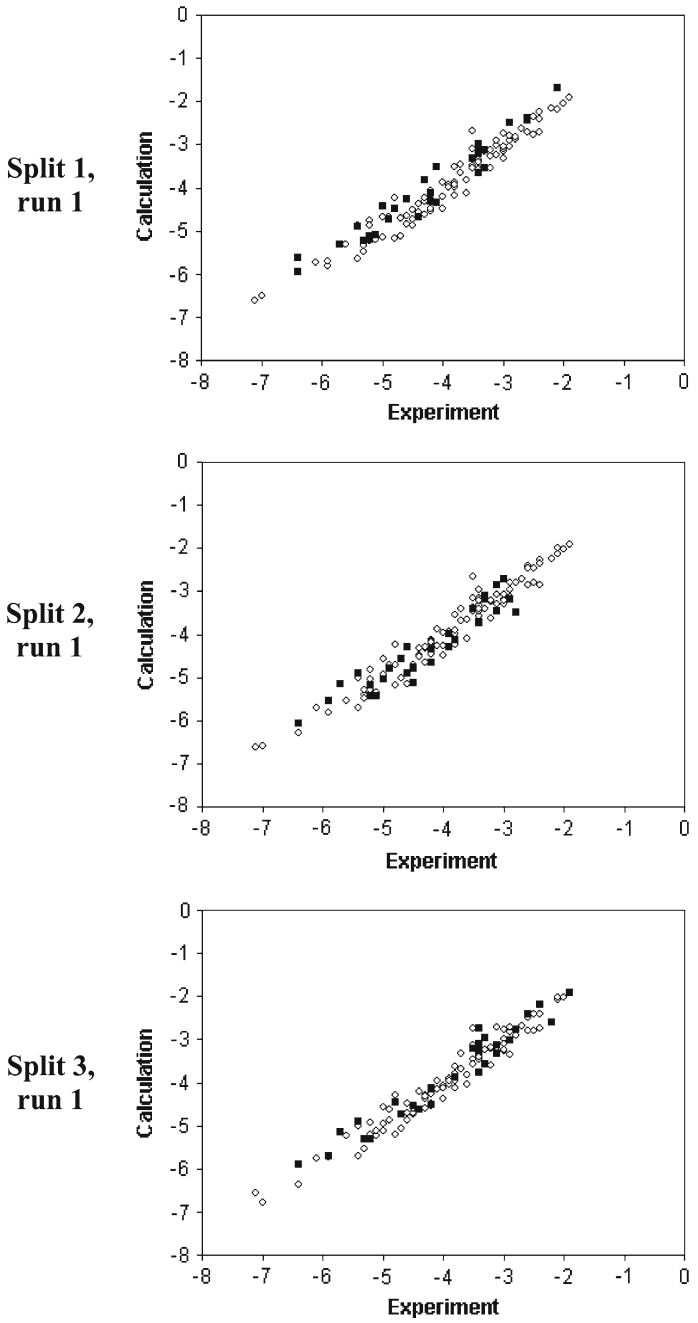


Fig. 2 QSPR models of C_{60} solubility for three splits:○ training set, ■ test set

5 Conclusions

International Chemical Identifier can be used as elucidation of the molecular structure for the QSPR analysis of the solubility of fullerene C_{60} in organic solvents. Distribution of the InChI attributes in the training set and test set is important criterion for definition of the splits: the absence of majority of InChI attributes in training set leads to poor QSPR model. The statistical quality of the InChI-based model for fullerene C_{60} solubility is better than the quality of SMILES-based model [16] of this parameter.

Acknowledgements Authors are grateful to the EC project CAESAR (contract SSP1-022674-CAESAR) and the Marie Curie Fellowship for financial support (contract 39036, CHEMPREDICT).

References

1. H. Liu, X. Yao, R. Zhang, M. Liu, Z. Hu, B. Fan, J. Phys. Chem. B **109,** 20565–20571 (2005)
2. D. Weininger, J. Chem. Inf. Comput. Sci. **28,** 31–36 (1988)
3. D. Weininger, A. Weininger, J.L. Weininger, J. Chem. Inf. Comput. Sci. **29,** 97–101 (1989)
4. D. Weininger, J. Chem. Inf. Comput. Sci. **30,** 237–243 (1990)
5. Unofficial InChI FAQ, http://wwmm.ch.cam.ac.uk/inchifaq/
6. Wendy Warr & Associates, http://www.warr.com/
7. ACD/ChemSketch Freeware, version 11.00, 2007, Advanced Chemistry Development, Inc., Toronto, ON, Canada, www.acdlabs.com
8. United States National Library of Medicine, http://toxnet.nlm.nih.gov/
9. NIST Chemistry WebBook, http://webbook.nist.gov/chemistry/
10. D. Vidal, M. Thormann, M. Pons, J. Chem. Inf. Model. **45,** 386–393 (2005)
11. A.A. Toropov, A.P. Toropova, D.V. Mukhamedzhanova, I. Gutman, Indian J. Chem. A **44,** 1545–1552 (2005)
12. A.A. Toropov, A.P. Toropova, I. Raska Jr., Eur. J. Med. Chem. **43,** 714–740 (2008)
13. A.A. Toropov, E. Benfenati, Bioorg. Med. Chem. **16,** 4801–4809 (2008)
14. A.A. Toropov, A.P. Toropova, E. Benfenati, Chem. Phys. Lett. **461,** 343–347 (2008)
15. A.A. Toropov, D. Leszczynska, J. Leszczynski, Chem. Phys. Lett. **441,** 119–122 (2007)
16. A.A Toropov, B.F. Rasulev, D. Leszczynska, J. Leszczynski, Chem. Phys. Lett. **444,** 209–214 (2007)
17. A. Ashek, S.J. Cho, Biorg. Med. Chem. **14,** 1474–1482, (2006)
18. S. Ray, C. Sengupta, K. Roy, Cent. Eur. J. Chem. **5,** 1094–1113 (2007)
19. A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. **20,** 269–276 (2002)
20. M.-O. Fouchécourt, M. Béliveau, K. Krishnan, Sci. Total Environ. **274,** 125–135 (2001)
21. V.K. Gombar, V.K. Kapoor, Eur. J. Med. Chem. **25,** 689–695 (1990)
22. L.C. Porto, E.S. Souza, J.B. Da Silva, R.A. Yunes, V.E.F. Heinzen, Talanta **76,** 407–412 (2008)
23. E.A.Castro, F. Torrens, A.A. Toropov, I.V. Nesterov, O.M. Nabiev, Mol. Simul. **30,** 691–696 (2004)