Mol Divers (2011) 15:249–256
DOI 10.1007/s11030-010-9245-6

SHORT COMMUNICATION

CORAL: QSPR models for solubility of [C_{60}] and [C_{70}] fullerene derivatives

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

Received: 7 January 2010 / Accepted: 2 March 2010 / Published online: 27 March 2010

Abstract Quantitative structure–property relationships (QSPRs) between the molecular structure of [C_{60}] and [C_{70}] fullerene derivatives and their solubility in chlorobenzene (mg/mL) have been established by means of CORAL (CORrelations And Logic) freeware. The CORAL models are based on representation of the molecular structure by simplified molecular input line entry system (SMILES). Three random splits into the training and the external validation sets have been examined. The ranges of statistical characteristics of these models are as follows: $n = 18$, $r^2 = 0.748$–0.815, $s = 15.1$ –17.5 (mg/mL), $F = 47$–71 (training set); $n = 9$, $r^2 = 0.806$–0.936, $s = 12.5$–17.5 (mg/mL), $F = 29$–103 (validation set).

Keywords QSPR · SMILES · Fullerene · Solubility · Optimal descriptor

Introduction

Theoretical tools are currently used for predictions of various molecular properties. They could be used with (theoretical/experimental approaches) or without (ab initio methods) input from experiments. Quantitative structure–property/activity relationships (QSPRs/QSARs) are very useful techniques that are applied for the estimation of physicochemical and biological parameters for substances which have not been examined by experiments. In spite of improvement of laboratory equipment, experimental analysis of all newly synthesized substances is impossible. Thus, a QSPR approach provides a necessary compromise that allows for the estimation of physicochemical parameters of large classes of compounds which are important from the point of view of theory or applications in industry.

Fullerene derivatives have been intensively studied by both the experimental and computational chemists. In addition, they are vital for many technological applications. For such industrial applications the knowledge of various physico-chemical parameters of fullerene derivatives, including their solubility, is crucial [1]. Owing to an increase in the number of databases with the molecular structure represented by simplified molecular input line entry system (SMILES) [2–5] available via the Internet, the SMILES-based QSPR models [6–12] become convenient alternative to models based on molecular graphs [13–31] .

QSPR/QSAR analyses of fullerene derivatives which are based on molecular graphs are problematic owing to the complexity of architecture of their molecules (Table 1). However, there is some experience on the QSPR/QSAR analyses of ful-

Electronic supplementary material The online version of this article (doi:10.1007/s11030-010-9245-6) contains supplementary material, which is available to authorized users.

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

G. Gini
Department of Electronics and Information, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy

D. Leszczynska
Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental Engineering, Jackson State University, 1325 Lynch St., Jackson, MS 39217-0510, USA

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

Table 1 Molecular structures of [C_{60}] and [C_{70}] fullerene derivatives

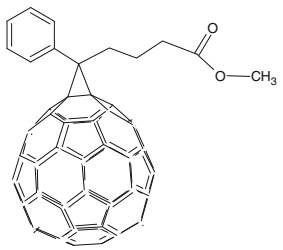
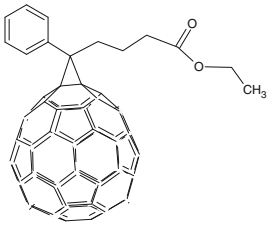
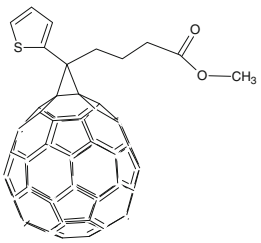
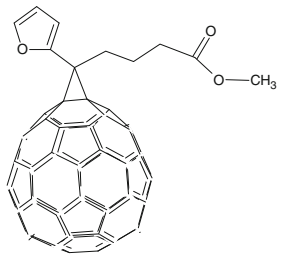
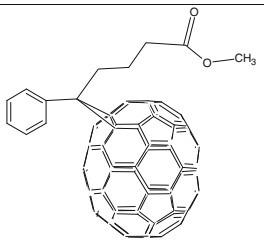
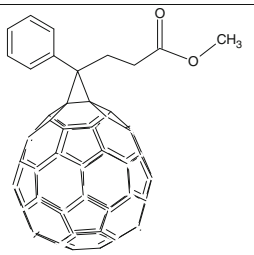
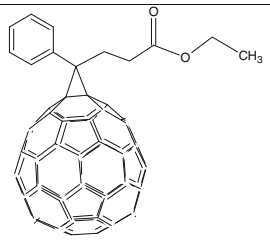
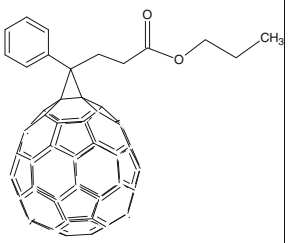
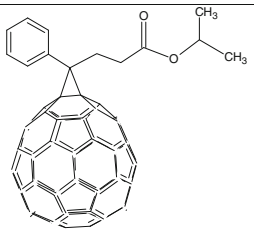
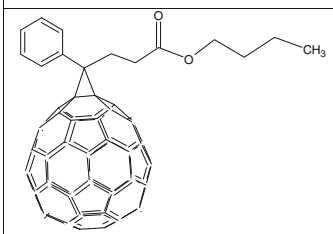
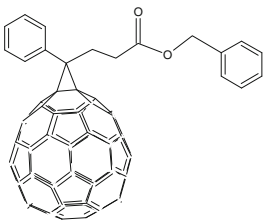
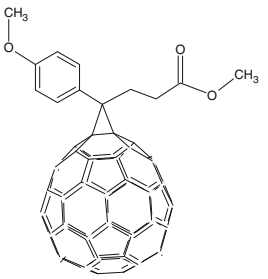
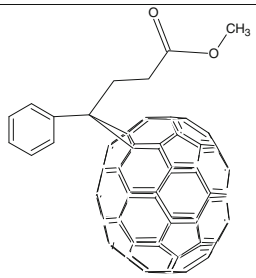
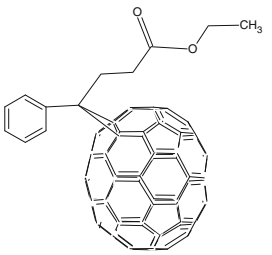
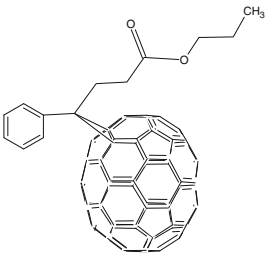
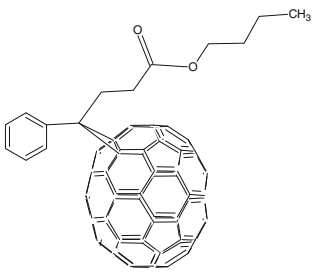
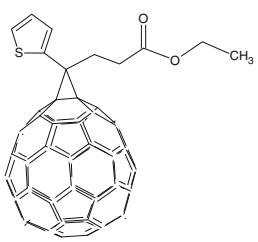
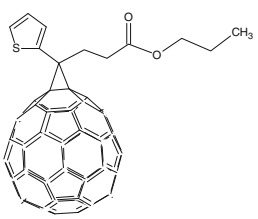
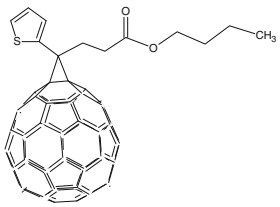
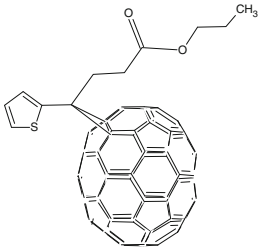
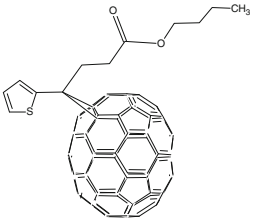
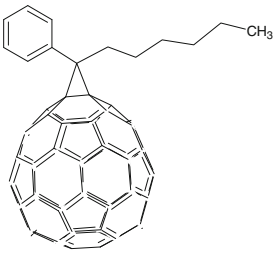
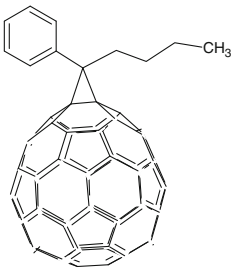
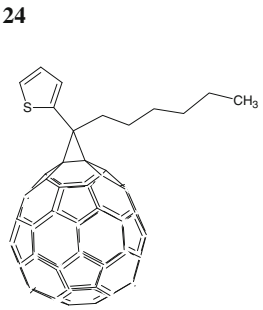
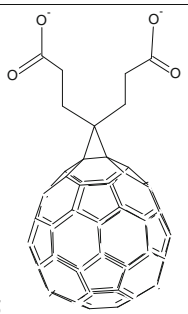
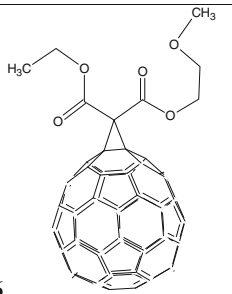
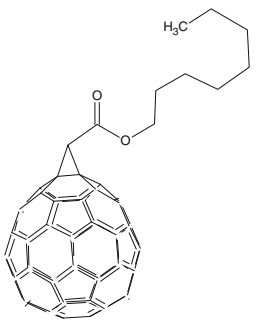
1 	**2** 	**3** 
4 	**5** 	**6** 
7 	**8** 	**9** 
10 	**11** 	**12** 
13 	**14** 	**15** 

Table 1 continued

16 	**17** 	**18** 
19 	**20** 	**21** 
22 	**23** 	**24** 
25 	**26** 	**27** 

lerene derivatives [9,15,16,18,19]. Taking into account the gradual increase in role of these substances in natural sciences and industry, one can expect that QSPR/QSAR models for these substances could be very useful.

The aim of the present study is an evaluation of ability the SMILES-based optimal descriptors as possible tools for efficient QSPR prediction of solubility of [C_{60}] and [C_{70}] fullerene derivatives in chlorobenzene [1].

Method

Data

The numerical data on the solubility of [C_{60}] and [C_{70}] fullerene derivatives in chlorobenzene in mg/mL were taken from [1].

Optimal SMILES-based descriptors

SMILES is a sequence of symbols which are representation of molecular architecture. First of all, SMILES encodes presence of chemical elements, e.g., 'c', 'C', 'N', 'Ni', etc. Also, SMILES encodes presence of different covalent bonds, i.e., '=' and '#'. Finally, SMILES encodes some 3D aspects, such as, rotations near bonds ('@' and '@@'), the branching of molecular skeleton (brackets), presence of cycles (digits), and so on [2–5]. Thus, the SMILES can be an alternative of molecular graph in the QSPR/QSAR analysis. In other words, SMILES can be a basis for calculation of the molecular descriptors.

Optimal SMILES-based descriptors used in the present study are calculated as the follows:

$$\mathrm{DCW\ (Threshold)} = \Sigma\ \mathrm{CW}\ (S_k) \tag{1}$$

where S_k is an element of SMILES and CW(S_k) represents so-called correlation weights for the S_k. The element of SMILES can be one character (e.g., 'c', 'C', '=', '#', etc), two characters that cannot to be examined independently (e.g., 'Br', 'Cl', etc.), and three characters (e.g., %10, %11, etc., these are used for depiction of cycles if the number of cycles is larger than 9). The threshold is the parameter for separation of SMILES elements into two classes: rare and not rare. We have used Threshold = 1. This value indicates that S_k that takes place in the training less than 1 time should be blocked, i.e., its correlation weight should be equal to zero.

Using the Monte Carlo method one can calculate CW(S_k) which for the training set yield as large as possible correlation coefficient between the DCW and the solubility. After evaluation of the CW(S_k) for the compounds of the training set one can calculate the DCW and define a model:

$$S\ (\mathrm{mg/mL}) = \mathrm{C_0} + \mathrm{C_1}\quad \mathrm{DCW(Threshold)} \tag{2}$$

The predictability of the Eq. 2 should be tested using compounds of an external validation set (i.e., compounds which have not been used for calculation of the model calculated with Eq. 2). The CORAL is a provider of these data which are calculated by special algorithm (CHEMPREDICT at: http://www.insilico.eu/coral).

The algorithm can be represented by two phases. The first phase is the preparation of the list of all SMILES attributes which take place in training and validation sets. The second phase is the calculation (by the Monte Carlo method) of values for correlation weights for these attributes which give maximum of the correlation coefficient between the SMILES-based descriptor and the endpoint for the training set. SMILES attributes, which are absent in the training set, have no influence on the model. Still, SMILES attributes, which are rare in the training set, can lead to overtraining (i.e., the model will be ideal for the training set, but poor for the external validation set). The influence of rare attributes can be reduced if the rare attributes will not involved in the modeling process. For this, one can define threshold: if the number of SMILES (in the training set) which contain the given attribute (SA) is smaller than the threshold the correlation weight for the SA should be equal to zero. Consequently, the influence of the SA upon the model will be blocked. Apparently, different thresholds can give models with different predictive ability. The predictive ability can be estimated in sequence of the runs of the modeling with different threshold. The threshold that gives the best statistical quality for external validation set should be defined as preferable for practical use. SMILES attributes which are absent in the training set (i.e., attributes which take place only in the validation set) are not involved in the modeling process.

Canonical SMILES notations have been built up with ACD/ChemSketch Freeware [5].

Results

Table 1 shows molecular structure of fullerene derivatives. Three splits of compounds were selected for the validation sets. Table 2 contains numbers of compounds used in the external validation set for the Split A, B, and C, respectively. Table 3 contains statistical characteristics of models for solubility.

The first run of the Monte Carlo optimization (with Threshold = 1) for split A yields the following model for the solubility:

$$S\ (\mathrm{mg/mL}) = -1585.71(\pm 29.09) + 7.1324\ (\pm 0.1253) * \mathrm{DCW}(1) \tag{3}$$

$n = 18$, $r^2 = 0.758$, $q^2 = 0.7211$, $s = 17.6$ mg/mL, $F = 50$ (training set)
$n = 9$, $r^2 = 0.925$, $R^2_m = 0.9017$, $s = 12.5$ mg/mL, $F = 87$ (validation set)

Table 2 Three random splits used in this study

Split	List of compounds in the external validation set
A	3, 6, 9, 12, 15, 18, 21, 24, 27
B	1, 2, 7, 10, 13, 19, 21, 22, 26
C	2, 6, 8, 12, 15, 19, 20, 22, 25

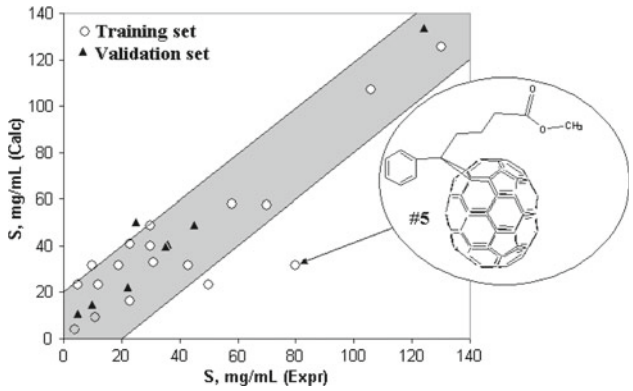


Fig. 1 Experimental and calculated using Eq. 3 values of [C_{60}] and [C_{70}] fullerene derivatives solubility in chlorobenzene. Structure #5 is an outlier

where R^2_m is a measure of the predictability according to [32]:

$$R^2_m = r^2 * \left(1 - \sqrt{r^2 - r^2_o}\right) \quad (4)$$

In the above equation r^2 and r^2_o indicate determination coefficient between observed and predicted values with and without intercept, respectively. These calculations were performed using CORAL freeware (CHEMPREDICT at: http://www.insilico.eu/coral).

Figure 1 shows the model calculated with Eq. 3, graphically. Electronic Supplementary material contains correlation weights, experimental and calculated with Eq. 3 solubility [(S(mg/mL)] values and an example of the DCW(1) calculation.

Discussion

According to Organisation for Economic Co-operation and Development (OECD) principles (OECD at: http://www.oecd.org/dataoecd/33/37/37849783.pdf) QSAR model should be associated with the following information:

(1) a defined endpoint
(2) an unambiguous algorithm
(3) a defined domain of applicability
(4) appropriate measures of goodness-of-fit, robustness, and predictivity
(5) a mechanistic interpretation, if possible.

The same principles can be useful for the QSPR case, i.e., for the modeling of physicochemical parameters. In particular, the solubility has regulatory importance, because ecological effect often is defined by the solubility of a substance. Of course, water solubility is a very important parameter, in this aspect; however, solubility in chlorobenzene can also be an ecologic indicator.

The Algorithm used for examined models (Table 3) is described in the literature [33] and also represented in the CORAL freeware.

The list of SMILES attributes and their correlation weights can be used to define the applicability domain of examined models: firstly, the models can be used for fullerene derivatives, and secondly, SMILES of these substances must contain attributes which take place in SMILES of the training set.

The predictability of SMILES-based models can be estimated by widely used statistical criterions: correlation coefficient, standard error, and Fischer F-ratio. The reproducibility of statistical quality of the QSPR model for three splits into training and test sets is an additional guarantee of the reliability of model.

Each SMILES element is an image of molecular reality. It is not only the presence of chemical elements, but also, the presence of branchings (brackets) in the molecular skeleton, *cis*- and *trans*- isomerism ('/' and '\'), covalent chemical bonds ('=' and '#') and others. The described approach gives a possibility to extract SMILES elements which are promoters of the solubility increase and vice versa promoters of the solubility decrease, as well as one can detect SMILES attributes of undefined role. In fact, it can be basis for heuristic hypotheses about molecular mechanisms of the solubility for fullerene derivatives. It should be noted that a split have influence on the distribution of the attributes in these classes, e.g., the split A has attributes of undefined role, whereas, the split B has not such attributes. These details are represented in the Electronic Supplementary material.

The Monte Carlo optimization is a random process. If the correlation weight for the SMILES attribute SA in sequence of the runs of the optimization has values which all are larger than zero, then the attribute can be estimated as stable promoter of increase of the endpoint, i.e., the presence of the molecular fragment encoded by this SA is indicator of increase of the endpoint. Vice versa, if the correlation weight for the SA has in sequence of the runs of optimization values which all are smaller than zero, the attribute can be estimated as a stable promoter of decrease of the given endpoint. Finally, if a SMILES attribute in three runs of the optimization has both correlation weights: smaller and larger than zero values, one can estimate the attribute as an attribute of undefined role.

By the reasons given in the previous paragraph, the described approach obeys the above-mentioned OECD principles.

Substance #5 is an outlier for the model calculated with Eq. 3. There are seven [C_{70}] fullerene derivatives (Fig. 2), however, only substance #5 has $-CH_2-CH_2-CH_2-$ con-

Table 3 Statistical characteristics of models for the solubility of fullerene [C_{60}] and [C_{70}] in chlorobenzene

Runs	Training set, $n = 18$			Validation set, $n = 9$			
	r^2	S (mg/mL)	F	r^2	S (mg/mL)	F	R^2_m
All substances							
Split A							
1	0.7580	17.585	50	0.9252	12.522	87	0.9017
2	0.7604	17.499	51	0.9136	13.581	74	0.8415
3	0.7604	17.498	51	0.9138	13.365	74	0.8553
Average	0.7596	17.527	51	0.9175	13.156	78	0.8662
Split B							
1	0.8153	15.077	71	0.8055	17.290	29	0.7572
2	0.8096	15.307	68	0.8138	17.056	31	0.7499
3	0.8140	15.130	70	0.8057	17.508	29	0.7437
Average	0.8130	15.171	70	0.8083	17.285	30	0.7503
Split C							
1	0.7498	16.936	48	0.9303	14.079	93	0.7880
2	0.7464	17.051	47	0.9359	13.745	102	0.7978
3	0.7478	17.003	47	0.9361	14.000	103	0.7842
Average	0.7480	16.997	47	0.9341	13.941	99	0.7900
Runs	Training set, $n = 17$			Validation set, $n = 9$			
	r^2	S (mg/mL)	F	r^2	S (mg/mL)	F	R^2_m
Without the outlier #5							
Split A							
1	0.8994	11.243	134	0.9054	14.401	67	0.7752
2	0.8991	11.257	134	0.9112	13.900	72	0.7943
3	0.8979	11.326	132	0.9024	14.920	65	0.7559
Average	0.8988	11.275	133	0.9064	14.407	68	0.7751
Split B							
1	0.9325	8.979	207	0.7868	19.071	26	0.7756
2	0.9339	8.885	212	0.7998	18.480	28	0.7748
3	0.9343	8.859	213	0.7972	18.401	28	0.7854
Average	0.9336	8.907	211	0.7946	18.650	27	0.7786
Split C							
1	0.8969	10.678	131	0.9373	14.020	105	0.8262
2	0.8957	10.741	129	0.9427	13.724	115	0.8338
3	0.8964	10.704	130	0.9400	13.918	110	0.8289
Average	0.8963	10.708	130	0.9400	13.887	110	0.8296

nector of R_1 and $-COOR_2$ fragments, whereas, all others have $-CH_2-CH_2-$ connector. Probably, this feature leads to untypical behavior for #5. We have detected an apparent feature for the #5. This feature does not take place for other substances examined in the present study. Thus, we have suggested the hypothesis that will be confirmed or vice versa rejected in the future researches.

The attempts to build up the model without #5 have shown that there is an improvement of the model for the training set (for all examined splits). However, the statistical characteristics of the model for the validation set remained approximately the same (excepting the split b, where the prediction becomes poorer). Thus, the removing #5 did not improve the predictive ability of the model.

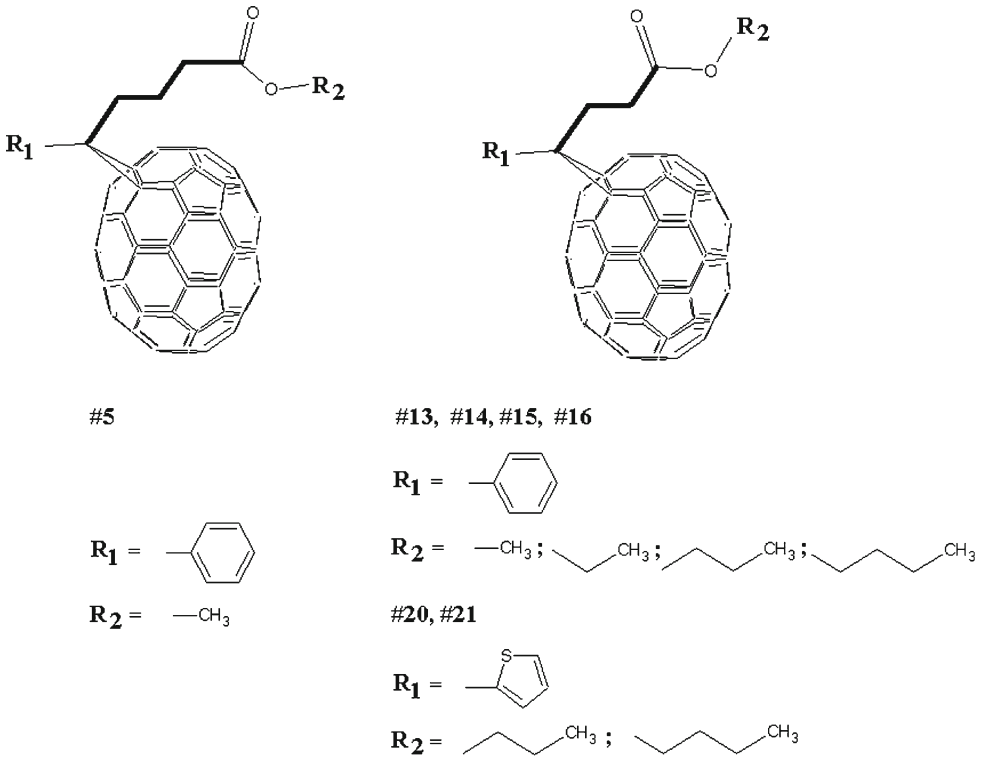


Fig. 2 Structure of seven fullerene [C_{70}] derivatives: substance #5 has unique $-CH_2-CH_2-CH_2-$ fragment that is connector R_1 and $COOR_2$. All other fullerene [C_{70}] derivatives have $-CH_2-CH_2-$ connector in this position

Conclusions

This study shows that the CORAL software can be a tool for modeling of solubility of fullerene derivatives [S, (mg/mL), in chlorobenzene]. The SMILES-based optimal descriptors calculated using CORAL freeware give a reasonable prediction of the solubility of both the [C_{60}] and [C_{70}] fullerene derivatives. The study was performed for three random splits of the data into the training and validation sets, hence the statistical quality of the model is not random one.

Acknowledgments The authors are grateful to the Marie Curie Fellowship for financial support (contract 39036, CHEMPREDICT) and the NSF CREST Interdisciplinary Nanotoxicity Center NSF-CREST Grant No. HRD-0833178.

References

1. Troshin PA, Hoppe H, Renz J et al (2009) Material solubility-photovoltaic performance relationship in the design of novel fullerene derivatives for bulk heterojunction solar cells. Adv Funct Mater 19:779–788. doi:10.1002/adfm.200801189
2. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36. doi:10.1021/ci00057a005
3. Weininger D, Weininger A, Weininger JL (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. J Chem Inf Comput Sci 29:97–101. doi:10.1021/ci00062a008
4. Weininger D (1990) SMILES. 3. DEPICT. Graphical depiction of chemical structures. J Chem Inf Comput Sci 30:237–243. doi:10.1021/ci00067a005
5. ACD/ChemSketch Freeware, version 11.00 (2007) Advanced Chemistry Development, Inc., Toronto, ON, Canada. www.acdlabs.com
6. Vidal D, Thormann M, Pons M (2005) LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. J Chem Inf Model 45:386–393. doi:10.1021/ci0496797
7. Vidal D, Thormann M, Pons M (2006) A novel search engine for virtual screening of very large databases. J Chem Inf Model 46:836–843. doi:10.1021/ci050458q
8. Vidal D, Blobel J, Pérez Y et al (2007) Structure-based discovery of new small molecule inhibitors of low molecular weight protein tyrosine phosphatise. Eur J Med Chem 42:1102–1108. doi:10.1016/j.ejmech.2007.01.017
9. Toropov AA, Leszczynska D, Leszczynski J (2007) QSPR study on solubility of fullerene C60 in organic solvents using optimal descriptors calculated with SMILES. Chem Phys Lett 441:119–122. doi:10.1016/j.cplett.2007.04.094
10. Toropov AA, Toropova AP, Raska I Jr (2008) QSPR modeling of octanol/water partition coefficient for vitamins by optimal descriptors calculated with SMILES. Eur J Med Chem 43:714–740
11. Toropov AA, Benfenati E (2008) Additive SMILES-based optimal descriptors in QSAR modeling bee toxicity: Using rare SMILES attributes to define the applicability domain. Bioorg Med Chem 16:4801–4809. doi:10.1016/j.bmc.2008.03.048
12. Toropov AA, Toropova AP, Benfenati E (2008) QSPR modeling for enthalpies of formation of organometallic compounds by means of SMILES-based optimal descriptors. Chem Phys Lett 461:343–347. doi:10.1016/j.cplett.2008.07.027
13. Rasulev BF, Toropov AA, Hamme AT II et al (2008) Multiple linear regression analysis and optimal descriptors: predicting the cholesteryl ester transfer protein inhibition activity. QSAR Comb Sci 27:595–606. doi:10.1002/qsar.200710006
14. Toropov AA, Rasulev BF, Leszczynski J (2007) QSAR modeling of acute toxicity for nitrobenzene derivatives towards rats: com-

parative analysis by MLRA and optimal descriptors. QSAR Comb Sci 26:686–693. doi:10.1002/qsar.200610135
15. Liu H, Yao X, Zhang R et al (2005) Accurate quantitative structure-property relationship model to predict the solubility of C60 in various solvents based on a novel approach using a least-squares support vector machine. J Phys Chem B 109:20565–20571. doi:10.1021/jp052223n
16. Gharagheizi F, Alamdari RF (2008) A molecular-based model for prediction of solubility of C60 fullerene in various solvents. Fuller Nanotub Car N 16:40–57. doi:10.1080/15363830701779315
17. Gutman I, Toropov AA, Toropova AP (2005) The graph of atomic orbitals and its basic properties. 1. Wiener index. MATCH Commun Math Comput Chem 53:215–224
18. Durdagi S, Mavromoustakos T, Papadopoulos MG (2008) 3D QSAR CoMFA/CoMSIA, molecular docking and molecular dynamics studies of fullerene-based HIV-1 PR inhibitors. Bioorg Med Chem Lett 18:6283–6289. doi:10.1016/j.bmcl.2008.09.107
19. Durdagi S, Mavromoustakos T, Chronakis N et al (2008) Computational design of novel fullerene analogues as potential HIV-1 PR inhibitors: analysis of the binding interactions between fullerene inhibitors and HIV-1 PR residues using 3D QSAR, molecular docking and molecular dynamics simulations. Bioorg Med Chem 16:9957–9974. doi:10.1016/j.bmc.2008.10.039
20. Kuz'min VE, Muratov EN, Artemenko AG et al (2008) The effect of nitroaromatics' composition on their toxicity in vivo: novel, efficient non-additive 1D QSAR analysis. Chemosphere 72:1373–1380. doi:10.1016/j.chemosphere.2008.04.045
21. Afantitis A, Melagraki G, Sarimveis H et al (2006) A novel QSAR model for evaluating and predicting the inhibition activity of dipeptidyl aspartyl fluoromethylketones. QSAR Comb Sci 25:928–935. doi:10.1002/qsar.200530208
22. Afantitis A, Melagraki G, Sarimveis H et al (2006) Prediction of intrinsic viscosity in polymer-solvent combinations using a QSPR model. Polymer 47:3240–3248. doi:10.1016/j.polymer.2006.02.060
23. Puzyn T, Mostrag A, Suzuki N et al (2008) QSPR-based estimation of the atmospheric persistence for chloronaphthalene congeners. Atmos Environ 42:6627–6636. doi:10.1016/j.atmosenv.2008.04.048
24. Puzyn T, Suzuki N, Haranczyk M (2008) How do the partitioning properties of polyhalogenated POPs change when chlorine is replaced with bromine. Environ Sci Tech 42:5189–5195. doi:10.1021/es8002348
25. Puzyn T, Suzuki N, Haranczyk M et al (2008) Calculation of quantum-mechanical descriptors for QSPR at the DFT level: is it necessary?. J Chem Inf Model 48:1174–1180. doi:10.1021/ci800021p
26. Gutman I, Furtula B, Toropov AA et al (2005) The graph of atomic orbitals and its basic properties. 2. Zagreb indices. MATCH Commun Math Comput Chem 53:225–230
27. Castro EA, Toropova AP, Toropov AA et al (2005) QSPR modeling of Gibbs free energy of organic compounds by weighting of nearest neighboring codes. Struct Chem 16:305–324. doi:10.1007/s11224-005-4462-0
28. Roy K, Toropov AA (2005) QSPR modeling of the water solubility of diverse functional aliphatic compounds by optimization of correlation weights of local graph invariants. J Mol Model 11:89–96. doi:10.1007/s00894-004-0224-7
29. Duchowicz PR, Castro EA, Toropov AA et al (2004) QSPR modeling the aqueous solubility of alcohols by optimization of correlation weights of local graph invariants. Mol Divers 8:325–330. doi:10.1023/B:MODI.0000047498.49219.ab
30. Toropov AA, Benfenati E (2004) QSAR modeling of aldehyde toxicity against a protozoan, *Tetrahymena pyriformis* by optimization of correlation weights of nearest neighboring codes. J Mol Struct THEOCHEM 679:225–228. doi:10.1016/j.theochem.2004.04.020
31. Toropov AA, Benfenati E (2004) QSAR modeling of aldehyde toxicity by means of optimisation of correlation weights of nearest neighbouring codes. J Mol Struct THEOCHEM 676:165–169. doi:10.1016/j.theochem.2004.01.023
32. Roy PP, Roy K (2009) QSAR Studies of CYP2D6 Inhibitor Aryloxypropanolamines Using 2D and 3D Descriptors. Chem Biol Drug Des 73:442–455. doi:10.1111/j.1747-0285.2009.00791.x
33. Toropov AA, Toropova AP, Benfenati E (2009) Additive SMILES-based carcinogenicity models: probabilistic principles in the search for robust predictions. Int J Mol Sci 10:3106–3127. doi:10.3390/ijms10073106