J Math Chem (2016) 54:698–706
DOI 10.1007/s10910-015-0581-8

ORIGINAL PAPER

# Introduction of simplex-informational descriptors for QSPR analysis of fullerene derivatives

**Natalia Sizochenko**[1,2] · **Victor Kuz'min**[2,3] · **Liudmila Ognichenko**[3] · **Jerzy Leszczynski**[1]

Received: 10 November 2015 / Accepted: 7 December 2015 / Published online: 15 December 2015


**Abstract** Rational approach towards the QSAR/QSPR modeling requires the selection of descriptors to be computationally efficient and physically and chemically meaningful. However, fullerenes and their derivatives represent challenging compounds in terms of QSPR modeling and there is a lack of efficient and comprehensible descriptors for them. Based on existing informational field model and simplex representation of molecular structure approach, an outline of descriptoral representation for fullerenes was developed. Solubility of fullerene derivatives was chosen as target property for the estimation of descriptors' efficacy. Developed model provides well-defined physical meanings and obtained results are interpreted in terms of basic molecular properties.

**Keywords** QSAR · Molecular descriptors · Simplex · Informational field · Fullerenes · Solubility

## 1 Introduction

In last years, one of the fastest growing areas of modern chemistry is the physical chemistry of nanostructures. For example, fullerenes that have been discovered 30

✉ Natalia Sizochenko
sizochenko@icnanotox.org

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

2 Department of Chemistry, Odessa I. I. Mechnikov National University, Dvoryanskaya Str., 2, 65082 Odessa, Ukraine

3 A.V. Bogatsky Phys.-Chem. Institute NAS of Ukraine, Lustdorfskaja doroga 86, 65080 Odessa, Ukraine

years ago [1] and their derivatives, developed over the last three decades are still in focus of scientists because of their unique electronic structure, physical and chemical properties [2]. Computational approaches are fast, low-cost and play an essential role in evaluation structures and properties of many chemicals. One of the most popular methods is Quantitative Structure-Activity/Property Relationship (QSAR/QSPR) analysis, which statistically establishes a mathematical relationship between target experimental properties and features of chemical structure (descriptors) [3].

Recently, several attempts have been made to reflect using mathematical models the features of fullerene's structure. For instance, topological shape factors have been introduced to preselect the most stable fullerene isomers [4]. However, in QSAR/QSPR studies only quantum-chemical descriptors [5] and descriptors derived from simplified molecular input line entry system (SMILES) were tested [6,7]. At the same time, there are many different commercial and open source generators of descriptors: PaDEL [8], Dragon [9], ISIDA/QSPR [10], Open3DQSAR [11], etc., but almost all of them are applicable only to "classical" chemicals (namely, organics) and are not able to generate descriptors for "big" molecules (namely, fullerenes and small proteins). Most of these tools are based on the idea of a molecular graph to generate fragments and fingerprints and it seems that implemented algorithms are not able to compute molecular graphs for closed spherical or ellipsoid structures (fullerenes). Therefore, specific descriptors scalable to "big" molecules are required.

Another problem with fullerenes is that fullerenes contain many similar atoms (carbons of fullerene's core), and less different atoms (substitutes or functional groups). Regular descriptors are not able to differentiate them. Development of descriptors that allows clearly differentiate atoms could be a great challenge for computational experts.

In the current study, in order to improve existing methods of molecular representation of fullerenes, the topological informational field model combined with Simplex Representation of Molecular Structure (SiRMS) [12,13] was applied. On this basis, series of so-called simplex-informational descriptors [14] for fullerene derivatives were computed. To evaluate quality of developed descriptors, QSPR model for solubility of 27 fullerene derivatives in chlorobenzene was developed.

## 2 Computational details

### 2.1 Simplex-informational descriptors

The information field of molecule can be modeled as a superposition of the appropriate fields of its independent parts, namely—atoms). In fact, such field reflects information about the distribution of the considered property in space [12]. In accordance with information field theory, each vertex of molecular graph is a source of information. A fullerene graph represents a planar, 3-regular and 3-connected graph. Twelve of such graphs are pentagons, and any remaining faces are hexagons. Inclusion of substitution in fullerene structure can dramatically change informational field of such complex molecular system.

Using different physical or chemical labels, one can observe changes in informational field [15]. However, full description of informational theory is out of scope of

this paper. Thus, excellent paper on informational theory is recommended for readers [16].

Let us set-up as starting point assumption that it is possible to obtain different informational descriptors by encoding chemical graph by using of different parameters or labels. For this purpose, 2D Simplex Representation of Molecular Structure (SiRMS) approach was applied [13,17,18].

Following this approach, at the first step, fullerene was represented as molecular graph. At the next step, the vertices of graph were labelled by various physicochemical properties: values of partial charges [19], parameters of Lennard–Jones 6–12 potential [20], polarization parameter, H-bond donor/acceptor labels, and lipophilicity. These properties are typically used for SiRMS-based calculations, so we decided to apply them from informational theory calculations. Using each of these properties, various topological information potentials (IPs) [15] were expressed by equation 1:

$$IP_i = w_i \sum_{j=1}^{n} \left( \frac{\sum_m lb\left(\frac{r}{2R_{ij}+1}\right)}{m} \right), \tag{1}$$

where $m$ is the number of all possible paths between every atom pair, $n$ is the number of atoms in the given molecule, $R_{ij}$ is the path length (number of bonds between the $i$th and $j$th atoms), $w_i$ is physicochemical property on which basis IP is calculated, $r$ is the maximal path length between atoms for investigated set of molecules.

IP calculations were followed by re-labelling procedure. Atoms were labeled as A, B, C, etc on the basis of type of IP near certain atom. At the last step, all molecules in dataset were fragmentized. Typically, SiRMS-based descriptors do label fragments of the size 4, but in current study we calculated fragments of sizes from 3 to 4 [13]. Number of simplexes of certain type (for example, A–B–D–G) for each molecule was set as descriptor. Simplex-informational descriptors were generated by using HiT QSAR software [21].

### 2.2 Statistical analysis

After calculating sets of descriptors, the descriptors that highly correlated with each other were eliminated. When the squared correlation coefficient between descriptors in a pair was higher than a given limit (set here as 0.85), one of variables was deleted. The descriptors having higher sum of squared correlation coefficients calculated in relation to all other descriptors were excluded. In addition, descriptors with no or with very little variance were also eliminated [22].

For assessing the ability of the model to make robust predictions, the initial dataset was splitted into training and a test sets based on random selection considering two rules: (a) the range of the response values of both the training set and the test set should be covered from the lowest to the highest; (b) the highest and lowest response values were included in the training set. Thus, initial dataset was splitted into 22 compounds for training set and 5 compounds for test set. QSPR tasks have been solved using the PLS regression on three latent variables [22]. The statistical fit of

a QSPR equation was assessed by correlation coefficient $R^2$ (both for training and for test sets), cross-validation correlation coefficient $Q^2$, and root-mean-square error RMSE (for training, validation and test sets). Chance correlations between solubility and selected descriptors were additionally tested by using scrambling procedure [23]. Domain of applicability (AD) for developed model was defined by means of Williams plot [24].

## 3 Case study: solubility of $C_{60}$ and $C_{70}$ derivatives in chlorobenzene

Solubility of fullerenes is very important characteristic for the development of different crystallization, extraction, and chromatographic separation techniques of fullerenes [2]. In current study, information on $C_{60}$ and $C_{70}$ fullerene derivatives solubility in chlorobenzene was extracted from literature [25]. This dataset was already treated by means of QSPR analysis several years ago [6]. However, developed in the previous study the SMILES-based model does not allow performing mechanistic interpretation, since SMILES codes are able to describe only presence of chemical elements and types of covalent bonds in molecule. In addition, the previous study provides a model without any transformations of initial values of solubility.

The initial experimental data contains all but one soluble fullerenes. Inclusion into dataset endpoint with such abnormal activity's value (almost insoluble) can implicitly destroy its descriptive and predictive abilities. At the same time, the wide variability of small dataset allows developing models with high statistical characteristics and high error values [26]. Thus, one insoluble compound was replaced by soluble fullerene $C_{60}$ [27]. Original solubility (mg/ml) [25] was converted to molar (mmol/ml) and expressed as log(S) values. Molecular structures, solubility and data transformation are summarized in Table 1.

Chemical structures were first pre-optimized with the Molecular Mechanics Force Field (MM+), and the resulting geometries were further refined by means of the semi-empirical PM7 method [28]. More than a thousand simplex-informational descriptors were generated.

As a result of PLS modeling we obtained six significant descriptors combined into three latent variables. Developed model is characterized by quite good statistical characteristics—training set: $R^2 = 0.939$, RMSE = 0.120; validation set: $Q^2 = 0.904$, RMSE = 0.141; test set: $R^2 = 0.873$, RMSE = 0.146; scrambling: $R^2 = 0.026\, Q^2 = 0.031$. A plot of observed experimentally determined versus predicted values of solubility is presented in Fig. 1. The straight line represents perfect agreement between experimental and calculated values.

Relative influence (%) and influence trend for each descriptor are presented in Table 2. To summarize our results, relative influences of descriptors were grouped by initial physicochemical properties (Fig. 2). Values of each descriptor are presented in Online Resource.

As one can see from Table 2 and Fig. 2, sum of informational descriptors based on partial charges ($S_2$ and $S_3$) have the highest influence on the solubility. Atomic weight-based descriptor ($S_1$) and polarization descriptors ($S_5$ and $S_6$) contributed

**Table 1** Solubility of fullerene derivatives in chlorobenzene

| # | Fullerene core | R1 | R2 | Solubility (mg/ml) | log(S) (mmol/ml) |
|---|---|---|---|---|---|
| 1 | $C_{60}$ | –Ph | $-(CH_2)_3COOMe$ | 50 | −1.26 |
| 2 | $C_{60}$ | –Ph | $-(CH_2)_3COOEt$ | 19 | −1.69 |
| 3 | $C_{60}$ | $-C_4H_3S$ | $-(CH_2)_3COOMe$ | 36 | −1.41 |
| 4 | $C_{60}$ | $-C_4H_3O$ | $-(CH_2)_3COOMe$ | 48 | −1.19 |
| 5 | $C_{70}$ | –Ph | $-(CH_2)_3COOMe$ | 80 | −1.11 |
| 6 | $C_{60}$ | –Ph | $-(CH_2)_2COOMe$ | 10 | −1.95 |
| 7 | $C_{60}$ | –Ph | $-(CH_2)_2COOEt$ | 5 | −2.26 |
| 8 | $C_{60}$ | –Ph | $-(CH_2)_2COOPr$ | 43 | −1.33 |
| 9 | $C_{60}$ | –Ph | $-(CH_2)_2COOPr\text{-}i$ | 22 | −1.62 |
| 10 | $C_{60}$ | –Ph | $-(CH_2)_2COOBu$ | 30 | −1.50 |
| 11 | $C_{60}$ | –Ph | $-(CH_2)_2COOBn$ | 106 | −0.96 |
| 12 | $C_{60}$ | –Ph–OMe | $-(CH_2)_2COOMe$ | 5 | −2.27 |
| 13 | $C_{70}$ | –Ph | $-(CH_2)_2COOMe$ | 12 | −1.93 |
| 14 | $C_{70}$ | –Ph | $-(CH_2)_2COOEt$ | 10 | −2.01 |
| 15 | $C_{70}$ | –Ph | $-(CH_2)_2COOPr$ | 35 | −1.47 |
| 16 | $C_{70}$ | –Ph | $-(CH_2)_2COOBu$ | 30 | −1.55 |
| 17 | $C_{60}$ | $-C_4H_3S$ | $-(CH_2)_2COOPr$ | 45 | −1.32 |
| 18 | $C_{60}$ | $-C_4H_3S$ | $-(CH_2)_2COOBu$ | 70 | −1.13 |
| 19 | $C_{70}$ | $-C_4H_3S$ | $-(CH_2)_2COOPr$ | 130 | −0.91 |
| 20 | $C_{70}$ | $-C_4H_3S$ | $-(CH_2)_2COOBu$ | 124 | −0.93 |
| 21 | $C_{60}$ | –Ph | $-CH_5CH3$ | 31 | −1.46 |
| 22 | $C_{60}$ | –Ph | –Bn | 23 | −1.58 |
| 23 | $C_{60}$ | $-C_4H_3S$ | $-C_6H_{13}$ | 25 | −1.56 |
| 24 | $C_{60}$ | $-CH_2COOMe$ | $-CH_2COOMe$ | 4 | −2.36 |
| 25 | $C_{60}$ | –COOEt | $-COO(CH_2)_2OMe$ | 11 | −1.94 |
| 26 | $C_{60}$ | –H | $-COOC_8H_{17}$ | 9 | −2.00 |
| 27 | $C_{60}$ | – | – | 7 | −2.01 |

almost equally. However, atomic weight-based descriptor has positive contribution in PLS model (increasing predicted property), while polarization-based descriptors contributed negatively (Table 2). Lipophilicity-based informational descriptor ($S_4$) also has negative impact.

Let us take a closely look at developed model. Descriptor $S_1$ (27 %) describes positive influence of informational atomic weights of vertices in molecular graph. Descriptors $S_2$ (21 %) and $S_3$ (19 %) have positive impacts in the modeled PLS equation. These descriptors are based on information in molecular system, induced by charges. Charge distributions occurred by presence of heteroatoms (S, O, N) in the aromatic rings or chains. For instance, same derivatives of $C_{60}$ with different aromatic substituent demonstrate differences in solubility. Figure 3 presents common functional groups for these derivatives.

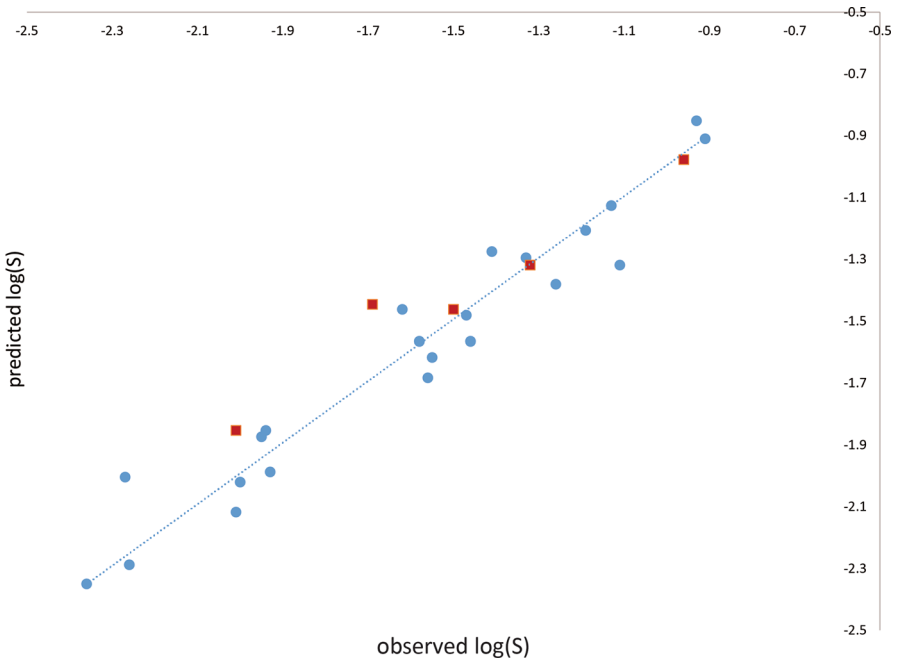


**Fig. 1** Plot of experimentally determined (observed) versus predicted values of (S). *Blue dots* represent compounds from training set, *red squares*—test set (Color figure online)

**Table 2** Relative influence of informational simplex descriptors

| Physicochemical property | Descriptor | Influence trend | Percentage |
|---|---|---|---|
| Atomic weight | $S_1$ | + | 27 |
| Partial charges | $S_2$ | + | 21 |
| | $S_3$ | + | 19 |
| Lipophilicity | $S_4$ | − | 10 |
| Polarization | $S_5$ | − | 15 |
| | $S_6$ | − | 8 |

As one can see from Table 1 for derivatives **1**, **3,** and **4** solubility decreases as follows: furan (**4**) > benzene (**1**) > thiophene (**3**). Similar dependency is observed for derivatives **8** and **17:** benzene (**8**) > thiophene (**17**). Thus, the charge-weighted informational descriptors point to the significant importance of electrostatic interactions of solvating species, related to their high polarity.

Descriptor $S_4$ reflects informational lipophilicity of fragments [–C–C=] in fullerenes $C_{60}$ and $C_{70}$. Fullerene $C_{70}$ includes more [–C–C=] fragments than $C_{60}$. Similarly, $C_{70}$ derivatives have higher solubility than same $C_{60}$ derivatives (Table 1). Solubility and lipophilicity are co-dependent properties, so one can conclude that descriptor $S_4$ directly reflects dissolution properties of studied fullerenes. Comparison of solubility for similar $C_{70}$ and $C_{60}$ derivatives is presented in Table 3.

Descriptors $S_5$ and $S_6$ reflect influence of some polar groups and aromatic fragments. Since chlorobenzene is a polar aprotic solvent it can enforce inflict inter-

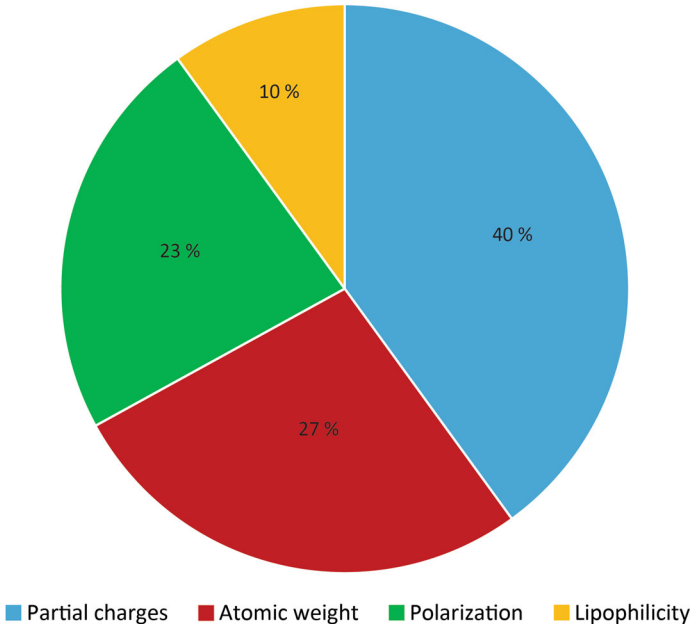


**Fig. 2** Diagram of relative contributions towards fullerenes' solubility in chlorobenzene (in %) of different physicochemical properties

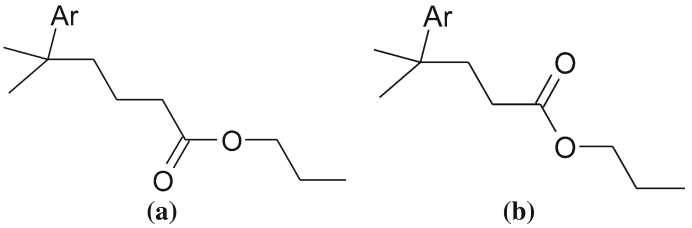


**Fig. 3** Common functional groups for fullerene $C_{60}$ derivatives: **a** common part for compounds **1**, **3**, **4**; **b** common part for compounds **8** and **17**.

**Table 3** Comparison between $C_{60}$ and $C_{70}$ derivatives solubility

| R1 | R2 | $C_{60}$ | | $C_{70}$ | | Difference for $C_{60}$ and $C_{70}$ (observed/predicted) |
|---|---|---|---|---|---|---|
| | | No | log(S) observed/predicted | No | log(S) observed/predicted | |
| –Ph | $-(CH_2)_3COOMe$ | 1 | −1.26/−1.38 | 5 | −1.11/−1.32 | 0.15/0.06 |
| –Ph | $-(CH_2)_2COOMe$ | 6 | −1.95/−1.87 | 13 | −1.93/−1.98 | 0.02/0.11 |
| –Ph | $-(CH_2)_2COOEt$ | 7 | −2.26/−2.29 | 14 | −2.01/−2.11 | 0.25/0.18 |
| –Ph | $-(CH_2)_2COOPr$ | 8 | −1.33/−1.30 | 15 | −1.47/−1.48 | 0.14/0.18 |
| –Ph | $(CH_2)_2COOBu$ | 10 | −1.50/−1.46 | 16 | −1.55/−1.60 | 0.05/0.14 |
| $-C_4H_3S$ | $-(CH_2)_2COOPr$ | 17 | −1.32/−1.32 | 19 | −0.91/−0.91 | 0.41/0.41 |
| $-C_4H_3S$ | $-(CH_2)_2COOPr$ | 18 | −1.13/−1.13 | 20 | −0.93/−0.85 | 0.20/0.28 |

actions, which are relevant to solvation process. Interactions between polar groups of fullerene's derivatives and solvent are among the factors responsible for solvation. These descriptors are related to interaction between solvent's aromatic ring and similar hexagonal structural motifs in fullerene [2].

Thus, developed here model has better statistical representation than mentioned above SMILES-based model [6], and in addition, it has strong mechanistic interpretation.

One can conclude that the chemical information encoded by six simplex-informational descriptors reflects the variation of the experimental solubility of fullerene derivatives in a satisfactory manner, and allowed a proper characterization of structurally heterogeneous compounds from both the training and test sets. It involved theoretical descriptors that have a direct interpretation. The statistical parameters of the proposed model compare fairly well with developed previously models, based on the considered here dataset. In addition, informational descriptors are potentially useful for ADME predictions (ADME stands for "absorption, distribution, metabolism, and excretion" abbreviation in pharmacokinetics and pharmacology).

## 4 Conclusions

Two different methods: simplex approach and the informational field [14] theory were simultaneously applied describing structure features of fullerene derivatives. Fullerene's molecular graphs were differentiated using informational potentials of the influence of near and far surroundings. Due to this fact the set of descriptors becomes more diverse. Proposed here simplex informational descriptors were evaluated in terms of mechanistic interpretation and were recognized as reliable descriptors for chemoinformatics studies.

Based on introduced descriptors we have analyzed quantitative structure-solubility relationships for set of fullerene derivatives and compared this model with the other, reported in the literature. The developed methodology of structural representation is potentially useful for QSAR/QSPR studies of fullerenes ADME evaluations. The QSPR solvation models, based on simplex informational descriptors are fast and have reasonable predictive power.

**Acknowledgments** This work was financially supported by National Science Foundation: NSF-CREST Grant #HRD-0833178 and EPSCoR Grant #362492-190200-01\NSFEPS-0903787. Authors also thank Office of Naval Research: Grant # N00014-13-1-0501.

## References

1. H.W. Kroto, J.R. Heath, S.C. O'Brien, R.F. Curl, R.E. Smalley, Nature **318**, 162 (1985)
2. K.N. Semenov, N.A. Charykov, V.A. Keskinov, A.K. Piartman, A.A. Blokhin, A.A. Kopyrin, J. Chem. Eng. Data **55**, 13 (2010)
3. T. Le, V.C. Epa, F.R. Burden, D.A. Winkler, Chem. Rev. **112**, 2889 (2012)
4. T. Reti, E. Bitay, Mater. Sci. Forum **537**, 439 (2007)
5. K. Choho, W. Langenaeker, G. Van De Woude, P. Geerlings, J. Mol. Struct. **338**, 293 (1995)
6. A. Toropova, A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, Mol. Divers. **15**, 249 (2011)

7. A.A. Toropov, A.P. Toropova, E. Benfenati, D. Leszczynska, J. Leszczynski, J. Comput. Chem. **31**, 381 (2010)
8. C.W. Yap, J. Comput. Chem. **32**, 1466 (2011)
9. A. Mauri, V. Consonni, M. Pavan, R. Todeschini, Match-Commun. Math. Co. **56**, 237 (2006)
10. F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, Mol. Inf. **29**, 855 (2010)
11. P. Tosco, T. Balle, J. Mol. Model. **17**, 201 (2011)
12. V.E. Kuz'min, L.N. Ognichenko, A.G. Artemenko, Mol. Model. **7**, 278 (2001)
13. V.E. Kuz'min, A.G. Artemenko, E.N. Muratov, J. Comput.-Aided Des. **22**, 403 (2008)
14. D. Aleksandrova, A. Yegorova, L.N. Ognichenko, Y.V. Scripinets, I.V. Ukrainets, V.E. Kuz'min, V.P. Antonovich, Metody i Ob'ekty Khimicheskogo Analiza **3**, 50 (2008)
15. L.N. Ognichenko, V.E. Kuz'min, A.G. Artemenko, QSAR Comb. Sci. **28**, 939 (2009)
16. T.D. Schneider, Nano Commun. Netw. **1**, 173 (2010)
17. V.E. Kuz'min, A.G. Artemenko, R.N. Lozytska, A.S. Fedtchouk, V.P. Lozitsky, E.N. Muratov, A.K. Mescheriakov, SAR QSAR Environ. Res. **16**, 219 (2005)
18. A.G. Artemenko, E.N. Muratov, V.E. Kuz'min, N.N. Muratov, E.V. Varlamova, A.V. Kuz'mina, L.G. Gorb, A. Golius, F.C. Hill, J. Leszczynski, A. Tropsha, SAR QSAR Environ. Res. **22**, 575 (2011)
19. A.K. Rappe, C.J. Casewit, K.S. Colwell, W.A. Goddard, W.M. Skiff, J. Am. Chem. Soc. **114**, 10024 (1992)
20. W.L. Jolly, W.B. Perry, J. Am. Chem. Soc. **95**, 5442 (1973)
21. A. Artemenko, HiT QSAR Software. http://www.qsar4u.com/)
22. S. Wold, J. Trygg, A. Berglund, H. Antti, Chemom. Intell. Lab. **58**, 131 (2001)
23. K.E. Hevener, D.M. Ball, J.K. Buolamwini, R.E. Lee, Bioorgan Med. Chem. **16**, 8042 (2008)
24. P. Gramatica, E. Papa, QSAR Comb. Sci. **24**, 953 (2005)
25. P.A. Troshin, H. Hoppe, J. Renz, M. Egginger, J.Y. Mayorova, A.E. Goryachev, A.S. Peregudov, R.N. Lyubovskaya, G. Gobsch, N.S. Sariciftci, V.F. Razumov, Adv. Func. Mater. **19**, 779 (2009)
26. A.H. Asikainen, J. Ruuskanen, K.A. Tuppurainen, SAR QSAR Environ. Res. **15**, 19 (2004)
27. M.T. Beck, G. Mándi, Fuller. Sci. Tech. **5**, 291 (1997)
28. J.J.P. Stewart, J. Mol. Model. **19**, 1 (2013)