Available online at www.sciencedirect.com
ScienceDirect
Journal of Fluorine Chemistry 128 (2007) 484–492
JOURNAL OF FLUORINE CHEMISTRY
www.elsevier.com/locate/fluor

QSPR analysis of fluorophilicity for organic compounds

Andrew G. Mercader [*], Pablo R. Duchowicz, Miguel A. Sanservino, Francisco M. Fernández, Eduardo A. Castro

INIFTA, División Química Teórica, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Diag. 113 y 64, Suc. 4, C.C. 16, 1900 La Plata, Argentina

Received 20 November 2006; received in revised form 20 December 2006; accepted 21 December 2006
Available online 3 January 2007

Abstract

We constructed a QSPR model from 116 organic compounds for the prediction of fluorophilicity. The 1268 theoretical descriptors explored by means of linear regressions, encoding different aspects of the topological, geometrical, and electronic molecular structure, lead to an optimal seven-parameter equation with a correlation coefficient $R = 0.9807$ and cross-validation parameter $R_{l\text{-}15\%\text{-}o} = 0.9677$. As a more realistic and practical application of present optimal QSPR model, it is applied to the estimation of the fluorophilicity of 69 non-yet synthesized molecular structures.

Keywords: QSPR theory; Molecular descriptors; Replacement method; Fluorophilicity

1. Introduction

Nowadays, the fluorous chemistry of compounds exhibit many interesting applications in synthesis and catalysis. The tendency of an organic substance to dissolve in fluorous media has continuously gained importance after the disclosure of the fluorous biphase catalysis in 1994 [1], as biphasic reactions take advantage of the fact that organic and fluorous phases are typically immiscible at room temperature, but may homogenize at elevated temperatures. One can therefore expose reactants in an organic phase to a catalyst in a fluorous phase simply by heating, and separate products (in the organic phase) from the catalyst (in the fluorous phase) on cooling. Other useful advantages of the techniques based on the fluor content rely on the unique physical and chemical properties associated with perfluorinated type of solvents, such as inertness, non-toxicity, and easy separation [2].

The fluorophilicity of a compound can be quantified through the associated partition coefficient (P) between f1uorous ($CF_3C_6F_{11}$) and organic ($CH_3C_6H_5$) layers [3].

$$\ln P = \ln\left[\frac{c(CF_3C_6F_{11})}{c(CH_3C_6H_5)}\right], \quad T = 298\,\mathrm{K} \qquad (1)$$

It is well-known that the experimental design of fluorophilic molecules demands a minimum fluorine content of 60%, the presence of one or more perfluorinated-alkyl chain called ‘ponytail’, and the absence of hydrogen bonding or polar groups which may interact with the organic phase. Furthermore, the fluorination of molecules often takes the form of adding long ponytails [4].

Clearly, the design of fluorous biphase catalysis experiments would be substantially improved by a reliable prediction of the fluorophilicity for a given molecule. A generally accepted remedy for overcoming the lack of experimental data in complex chemical phenomena is the analysis based on quantitative structure-property relationships (QSPR) [5], which in the present case may provide adequate predictions of fluorophilicity. The ultimate role of the different formulations of the QSPR theory is to suggest mathematical models for estimating relevant properties of interest, especially when they cannot be experimentally determined for some reason. These studies simply rely on the assumption that the structure of a compound determines its physicochemical properties. The molecular structure is therefore translated into the so-called molecular descriptors through mathematical formulae obtained from several theories, such as Chemical Graph Theory, Information Theory, Quantum Mechanics, etc. [6,7]. There exist more than a thousand theoretical descriptors available in the literature, and one usually faces the problem of selecting

[*] Corresponding author. Fax: +54 221 425 4642.
E-mail address: amercader@inifta.unlp.edu.ar (A.G. Mercader).

0022-1139/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jfluchem.2006.12.011

those which are the most representative of the property under consideration.

Several QSPR studies on fluorophilicity were published in the last 5 years. In 2001, Kiss et al. [8] estimated this property for 59 fluorinated organic molecules using a neural network (NN) combination of eight molecular descriptors chosen from a pool of almost a hundred variables. In 2002, Huque et al. [9] employed a modified version of the linear free-energy relationships (LFER) on 91 chemicals to derive a five-descriptor model with structural interpretation, characterized with statistical parameters $R = 0.9742$ and $S = 0.566$. In 2004, Duchowicz et al. [10] employed the same data set to propose a different QSPR model based on linear regression and the atoms and chemical bonds as molecular descriptors.

In 2004, de Wolf et al. [11] applied a universal lipophilicity model based on the mobile order/disorder (MOD) solution theory to predict partition coefficients for 88 molecules in either PFMCH/toluene or FC-72/benzene. However, those predictions required the knowledge of molecular volumes and modified non-specific cohesion parameters for the solute; data which are commonly unavailable. The same year, Daniels et al. [12] proposed a modified LFER for 93 organic compounds by means of five molecular surface area descriptors, achieving $R = 0.9716$ and $S = 0.638$ and showing an almost equal accuracy to the previous reported models.

The present study reports the predictions of fluorophilicity for 116 organic compounds whose experimental data were collected from two previous publications [9,13]. For this purpose, two widely applied modeling strategies based on linear regressions are employed, namely the forward stepwise regression and the replacement method [14–17]. In Section 2 we show and discuss our results. In Section 3 we summarize the main conclusions of this study. Finally, in Section 4 we outline the employed methods.

2. Results and discussion

We first apply the RM to the search for the best fluorophilicity–structure relationships. The equation that leads to the best cross-validation parameters $R_{\text{l-15\%-o}}$ and $S_{\text{l-15\%-o}}$ contains seven molecular descriptors of different type:

$$\begin{aligned}\ln P = {} & -5.688(\pm 0.5) - 0.348(\pm 0.01)\text{SEigp} \\ & +0.164(\pm 0.02)\text{RDF055p} + 0.0531(\pm 0.05)\text{MAXDP} \\ & -0.197(\pm 0.01)\text{Har} + 1.178(\pm 0.1)\text{CIC1} \\ & -6.488(\pm 0.9)\text{MATS1v} + 1.606(\pm 0.1)\text{HOMA} \\ N = 116, \quad & R = 0.9806, \quad S = 0.494, \quad F = 387.7, \\ p < 10^{-4}, \quad & \text{AIC} = 0.280, \quad \text{FIT} = 16.450 \\ R_{\text{loo}} = 0.9778, \quad & S_{\text{loo}} = 0.511 R_{\text{l-15\%-o}} = 0.9677, \\ S_{\text{l-15\%-o}} = 0.620 & \end{aligned} \qquad (2)$$

where the absolute errors of the regression coefficients are given in parentheses; R the correlation coefficient of the model, F the Fisher ratio, p the significance of the model, AIC the Akaike's information criterion and FIT is the Kubinyi function. A brief description for each variable appearing in all the proposed models is presented in Table 1. The application of the FSR procedure does not improve the quality of this relationship, as it leads to the following equation:

$$\begin{aligned}\ln P = {} & -2.238(\pm 0.6) + 0.151(\pm 0.005)\text{RTm} \\ & -1.902(\pm 0.2)\text{HATSp} - 0.00611(\pm 0.0007)\text{PCD} \\ & +0.526(\pm 0.07)\text{H2e} - 0.362(\pm 0.06)\text{C} - 006 \\ & +0.760(\pm 0.2)\text{N} - 068 + 0.807(\pm 0.3)\text{GATS1p} \\ N = 116, \quad & R = 0.9740, \quad S = 0.572, \quad F = 285.5, \\ p < 10^{-4}, \quad & \text{AIC} = 0.375, \quad \text{FIT} = 12.110 \\ R_{\text{loo}} = 0.9700, \quad & S_{\text{loo}} = 0.5932 R_{\text{l}-15\%-\text{o}} = 0.9281, \\ S_{\text{l}-15\%-\text{o}} = 0.9153 & \end{aligned} \qquad (3)$$

Once again, we conclude that the RM is preferable to the FSR for exploring large sets of descriptors.

Table 2 shows the predicted values of fluorophilicity and the corresponding residuals between parentheses. We may consider the closely related aromatic esters **64** ($R_{f7}C(O)OCH_2Ph$) and **65**

Table 1
Classification of molecular descriptors involved in the QSPR models

Symbol	Description	Type
MATS1v	Moran autocorrelation-lag 1/weighted by atomic van der Waals volumes	2D-autocorrelations
RDF055p	RDF-5.5 weighted by atomic polarizabilities	RDF[a]
MAXDP	Maximal electrotopological positive variation	Topological
Har	Harary H index	Topological
CIC1	Complementary information content (neighborhood symmetry of first-order)	Topological
HOMA	Armonic oscillator model of aromaticity index	Aromaticity indices
SEigp	Eigenvalue sum from polarizability weighted distance matrix	Topological
RTm	*R* total index/weighted by atomic masses	GETAWAY[b]
HATSp	Leverage-weighted total index/weighted by atomic polarizabilities	GETAWAY
PCD	Difference of multiple path counts to path counts	Topological
H2e	H autocorrelation of lag 2/weighted by atomic Sanderson electronegativities	GETAWAY
C-006	Number of CH_2RX groups	Atom-centred fragments
N-068	Number of Al_3–N groups[c]	Atom-centred fragments
GATS1p	Geary autocorrelation-lag 1/weighted by atomic polarizabilities	2D-autocorrelations

[a] RDF, radial distribution function.

[b] GETAWAY, GEometry, Topology and Atoms-Weighted AssemblY.

[c] Al, aliphatic groups.

Table 2
Experimental values of fluorophilicity and predictions achieved by Eq. (2) and Huque et al. [9]

N	Compound name	Exp.	Eq. (2)	Huke et al. [9]
1	Decane	−2.86	−3.09 (0.23)	−3.07 (0.21)
2	Undecane	−3.13	−3.28 (0.15)	−3.13 (0.00)
3	Dodecane	−3.35	−3.47 (0.12)	−3.19 (−0.16)
4	Tridecane	−3.71	−3.67 (−0.04)	−3.24 (−0.47)
5	Tetradecane	−3.94	−3.87 (−0.07)	−3.30 (−0.64)
6	Hexadecane	−4.50	−4.30 (−0.20)	−3.41 (−1.09)
7	Dec-1-ene	−2.99	−3.50 (0.51)	−3.29 (0.30)
8	Undec-1-ene	−3.26	−3.65 (0.39)	−3.34 (0.08)
9	Dodec-1-ene	−3.66	−3.81 (0.15)	−3.40 (−0.26)
10	Tridec-1-ene	−3.94	−3.98 (0.04)	−3.46 (−0.48)
11	Tetradec-1-ene	−4.12	−4.17 (0.05)	−3.51 (−0.61)
12	Hexadec-1-ene	−4.70	−4.56 (−0.14)	−3.62 (−1.08)
13	$R_{f8}CH=CH_2$	2.67	1.65 (1.02)	2.82 (−0.15)
14	Cyclohexanone	−3.79	−3.87 (0.08)	−3.96 (0.17)
15	Cyclohexenone	−4.06	−4.57 (0.51)	−4.25 (0.19)
16	Cyclohexanol	−4.12	−4.31 (0.19)	−4.74 (0.62)
17	Trifluoroethanol	−1.77	−1.92 (0.15)	−1.37 (−0.40)
18	$(CF_3)_2CHOH$	−1.02	−1.07 (0.05)	−0.70 (−0.32)
19	$R_{f6}(CH_2)_2OH$	0.10	0.05 (0.05)	0.47 (−0.37)
20	$R_{f6}(CH_2)_3OH$	−0.24	−0.14 (−0.10)	0.50 (−0.74)
21	$R_{f8}(CH_2)_2OH$	1.02	1.47 (−0.45)	0.72 (0.30)
22	$R_{f8}(CH_2)_3OH$	0.59	1.16 (−0.57)	0.80 (−0.21)
23	$R_{f10}(CH_2)_3OH$	1.42	2.10 (−0.68)	1.25 (0.17)
24	Pentafluorobenzene	−1.24	−1.40 (0.16)	−0.58 (−0.66)
25	Hexafluorobenzene	−0.94	−0.34 (−0.60)	−0.12 (−0.82)
26	Ethylbenzene	−4.41	−3.31 (−1.10)	−4.23 (−0.18)
27	Dodecylbenzene	−4.70	−4.53 (−0.17)	−4.79 (0.09)
28	$R_{f8}(CH_2)_3C_6H_5$	−0.02	0.48 (−0.50)	0.38 (−0.40)
29	o-$R_{f6}(CH_2)_3C_6H_4(CH_2)_3R_{f6}$	1.03	1.20 (−0.17)	1.37 (−0.34)
30	o-$R_{f8}(CH_2)_3C_6H_4(CH_2)_3R_{f8}$	2.34	2.69 (−0.35)	2.32 (0.02)
31	o-$R_{f10}(CH_2)_3C_6H_4(CH_2)_3R_{f10}$	3.62	3.40 (0.22)	3.23 (0.39)
32	m-$R_{f8}(CH_2)_3C_6H_4(CH_2)_3R_{f8}$	2.28	2.96 (−0.68)	2.32 (−0.04)
33	p-$R_{f8}(CH_2)_3C_6H_4(CH_2)_3R_{f8}$	2.33	2.97 (−0.64)	2.32 (0.01)
34	$R_{f8}(CH_2)_3Cl$	0.03	0.74 (−0.71)	0.37 (−0.34)
35	$R_{f8}(CH_2)_3NH_2$	0.85	0.29 (0.56)	1.29 (−0.44)
36	$R_{f8}(CH_2)_3NH(CH_2)_3R_{f8}$	3.32	2.75 (0.57)	3.34 (−0.02)
37	$(R_{f6}(CH_2)_2)_3P$	4.41	4.46 (−0.05)	3.75 (0.66)
38	$(R_{f8}(CH_2)_3)_3P$	4.41	–	4.79 (−0.38)
39	$(R_{f8}(CH_2)_4)_3P$	4.50	–	4.53 (−0.03)
40	$(R_{f8}(CH_2)_5)_3P$	4.50	–	4.27 (0.23)
41	$(R_{f6}(CH_2)_2)_2PC_{10}H_{19}$ (menthyl)	1.29	0.92 (0.37)	1.11 (0.18)
42	$(R_{f8}(CH_2)_2)_2PC_{10}H_{19}$ (menthyl)	2.70	1.92 (0.78)	2.10 (0.60)
43	$(p$-$R_{f6}C_6H_4)_3P$	−1.32	−1.31 (−0.01)	−0.57 (−0.75)
44	$(p$-$R_{f8}C_6H_4)_3P$	0.76	–	0.78 (−0.02)
45	$Ph(CH_2)_2SiH_3$	−3.29	−3.22 (−0.07)	−4.53 (1.24)
46	$Ph(CH_2)_2SiOC_8H_{15}$	−5.11	−5.15 (0.04)	−5.56 (0.45)
47	$Ph(CH_2)_2SiOC_6H_{11}$ (cyclohexyl)	−4.82	−4.89 (0.07)	−5.56 (0.74)
48	$R_{f6}I$	1.31	1.53 (−0.22)	0.34 (0.97)
49	$R_{f8}I$	2.04	2.69 (−0.65)	0.93 (1.11)
50	$R_{f10}I$	2.84	2.43 (0.41)	1.48 (1.36)
51	$R_{f8}CH=CH_2$	2.67	1.92 (0.75)	2.82 (−0.15)
52	$R_{f8}(CH_2)_3SH$	0.24	1.04 (−0.80)	1.23 (−0.99)
53	$R_{f8}N(CH_2CH_2)_2$	0.86	0.91 (−0.05)	1.48 (−0.62)
54	$R_{f6}S(CH_2)_2CO_2Et$	−0.67	−0.17 (−0.50)	−0.05 (−0.62)
55	$R_{f8}S(CH_2)_2CO_2Et$	0.04	0.76 (−0.72)	0.49 (−0.45)
56	CF_3SPh	−2.45	−2.87 (0.42)	−2.01 (−0.44)
57	m-$CF_3SC_6H_4CF_3$	−1.58	−2.17 (0.59)	−0.85 (−0.73)
58	$R_{f8}SPh$	0.59	1.03 (−0.44)	−0.15 (0.74)
59	$R_{f7}CH_2NHMe$	1.07	0.79 (0.28)	1.49 (−0.42)
60	$R_{f7}CH_2NMe_2$	1.53	1.10 (0.43)	1.63 (−0.10)
61	$R_{f7}CH_2N(CH_2CH_2)_2O$	0.14	0.43 (−0.29)	0.60 (−0.46)
62	$R_{f7}CH_2NHCH(Me)Ph$	−0.87	−0.73 (−0.14)	−0.65 (−0.22)
63	$R_{f7}C(O)Ph$	0.48	0.18 (0.30)	–
64	$R_{f7}C(O)OCH_2Ph$	2.14	0.54 (1.60)	–

Table 2 (*Continued*)

N	Compound name	Exp.	Eq. (2)	Huke et al. [9]
65	*p*-$R_{f7}C(O)OCH_2C_6H_4OCF_3$	3.15	1.55 (1.60)	–
66	$R_{f7}C(O)SMe$	1.16	0.92 (0.24)	0.57 (0.59)
67	$R_{f7}C(O)NHMe$	0.15	0.82 (−0.67)	−0.23 (0.38)
68	$R_{f7}C(O)NMe_2$	0.34	0.72 (−0.38)	0.66 (−0.32)
69	$R_{f7}C(O)N(CH_2CH_2)_2O$	−0.62	−0.32 (−0.30)	−0.38 (−0.24)
70	$R_{f7}C(S)Me$	1.08	1.46 (−0.38)	0.19 (0.89)
71	$R_{f7}C(S)NMe_2$	−0.66	0.22 (−0.88)	−0.20 (−0.46)
72	$R_{f7}C(S)N(CH_2CH_2)_2O$	−1.56	−1.06 (−0.50)	−1.18 (−0.38)
73	$R_{f7}C(S)NHCH(Me)Ph$	−1.84	−1.03 (−0.81)	−3.18 (1.34)
74	C_6H_6	−2.77	−2.58 (−0.19)	−4.12 (1.35)
75	CF_3Ph	−1.96	−2.39 (0.43)	−1.82 (−0.14)
76	$R_{f6}Ph$	0.54	0.33 (0.21)	0.24 (0.30)
77	$R_{f8}Ph$	1.24	1.38 (−0.14)	0.78 (0.46)
78	$R_{f10}Ph$	1.77	2.29 (−0.52)	1.28 (0.49)
79	*o*-$R_{f8}C_6H_4CF_3$	1.50	1.52 (−0.02)	1.37 (0.13)
80	*m*-$R_{f8}C_6H_4CF_3$	2.37	1.99 (0.38)	1.37 (1.00)
81	*p*-$R_{f8}C_6H_4CF_3$	2.13	2.01 (0.12)	1.37 (0.76)
82	*p*-$R_{f8}C_6H_4R_{f8}$	4.98	4.63 (0.35)	–
83	[*p*-$CF_3C_6H_4(CF_2)_4]_2$	−0.56	−0.10 (−0.46)	−0.18 (−0.38)
84	*o*-$R_{f6}(CH_2)_2C_6H_4Cl$	−0.64	−0.99 (0.35)	−0.63 (−0.01)
85	*p*-$R_{f6}(CH_2)_2C_6H_4Cl$	−1.02	−1.02 (0.00)	−0.63 (−0.39)
86	*p*-$R_{f8}(CH_2)_2C_6H_4Cl$	−0.37	−0.05 (−0.32)	−0.04 (−0.33)
87	*o*-$R_{f6}(CH_2)_2C_6H_4Br$	−1.05	−1.12 (0.07)	−1.22 (0.17)
88	*m*-$R_{f6}(CH_2)_2C_6H_4Br$	−1.44	−1.09 (−0.35)	−1.22 (−0.22)
89	*p*-$R_{f6}(CH_2)_2C_6H_4Br$	−1.49	−1.13 (−0.36)	−1.22 (−0.27)
90	*o*-$R_{f8}C_6H_4CO_2Me$	−0.39	0.34 (−0.73)	−0.18 (−0.21)
91	*m*-$R_{f8}C_6H_4CO_2Me$	0.12	−0.11 (0.23)	−0.18 (0.30)
92	*p*-$R_{f8}C_6H_4CO_2Me$	−0.01	0.00 (−0.01)	−0.18 (0.17)
93	1,3,5-$R_{f8}C_6H_3(CF_3)_2$	4.05	2.70 (1.35)	–
94	1,3,5-$(R_{f8})_2C_6H_3CO_2Me$	4.41	3.60 (0.81)	–
95	1,3,5$(R_{f8})_2C_6H_3CH_2OH$	3.62	3.19 (0.43)	–
96	1,3,5-$(R_{f8})_2C_6H_3CHO$	4.25	4.10 (0.15)	–
97	2-$R_{f8}C_5H_4N$ (pyridine)	0.54	1.12 (−0.58)	0.64 (−0.10)
98	3-$R_{f8}C_5H_4N$ (pyridine)	0.88	1.02 (−0.14)	0.64 (0.24)
99	4-$R_{f8}C_5H_4N$ (pyridine)	0.80	1.28 (−0.48)	0.64 (0.16)
100	$(CF_3)_3CO(CH_2)_2NH_2$	−0.14	−0.46 (0.32)	–
101	$(CF_3)_3CO(CH_2)_2NH(CH_3)$	−0.08	−0.19 (0.10)	–
102	$[(CF_3)_3CO(CH_2)_2]_2NH$	1.82	2.02 (−0.20)	–
103	$(CF_3)_3CO(CH_2)_2NH(CH_2)_3R_{f8}$	2.69	2.32 (0.37)	–
104	$(CF_3)_3CO(CH_2)_2N(CH_3)_2$	0.34	0.25 (0.08)	–
105	$[(CF_3)_3CO(CH_2)_2]_2NCH_3$	2.03	2.20 (−0.17)	–
106	$[(CF_3)_3CO(CH_2)_2]_3N$	3.62	3.63 (−0.01)	–
107	$R_{f8}(CH_2)_3NH_2$	0.85	0.29 (0.56)	–
108	$R_{f8}(CH_2)_4NH_2$	0.54	0.40 (0.14)	–
109	$R_{f8}(CH_2)_5NH_2$	0.28	0.24 (0.04)	–
110	$R_{f7}CH_2NH(CH_3)$	1.07	0.79 (0.28)	–
111	$R_{f8}(CH_2)_3NH(CH_3)$	0.88	0.76 (0.12)	–
112	$[R_{f4}(CH_2)_3]_2NH$	0.71	0.91 (−0.20)	–
113	$[R_{f6}(CH_2)_3]_2NH$	1.98	2.12 (−0.14)	–
114	$[R_{f10}(CH_2)_3]_2NH$	4.08	4.28 (−0.20)	–
115	$[R_{f8}(CH_2)_3]_2NH$	3.32	3.53 (−0.21)	–
116	$[R_{f8}(CH_2)4]_2NH$	2.97	3.43 (−0.46)	–
117	$[R_{f8}(CH_2)_5]_2NH$	2.59	2.88 (−0.29)	–
118	$R_{f7}CH_2N(CH_3)_2$	1.53	1.10 (0.43)	–
119	$R_{f8}(CH_2)_3N(CH_3)_2$	1.37	0.92 (0.45)	–
120	$[R_{f8}(CH_2)_3]_2NCH_3$	3.63	3.28 (0.35)	–

Residuals are presented in parentheses. R_{fn} refers to $(CF_2)_{n-1}CF_3$.

(*p*-$R_{f7}C(O)OCH_2C_6H_4OCF_3$) as outliers, with a residual value exceeding 3*S*. It is not possible to determine if such deviation is either a statistical consequence of present selection of descriptors in Eq. (2) or a physical (meaningful) result. It is quite possible that these compounds are somewhat structurally different to the others in the training set. In many cases present residuals are smaller than those of Huque et al. [9] also shown in Table 2. We mention that we were unable to include four molecules (**38**, **39**, **40** and **44**) from that previous study because the freely downloadable version of Dragon 3.0 is limited up to 100 atoms. On the other hand, those molecules identified by Huque et al. as outliers and omitted from their

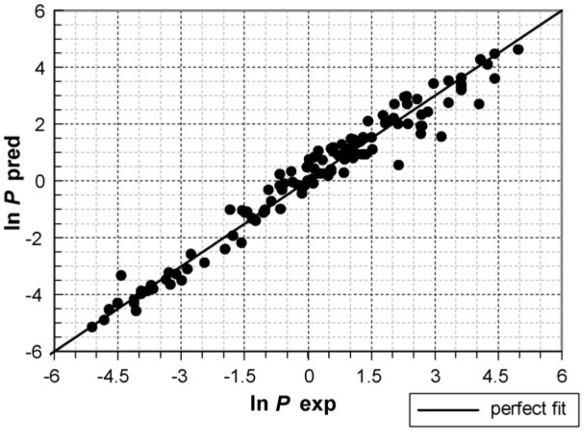

Fig. 1. Predicted vs. experimental fluorophilicity.

final model (**63**, **64**, **65**, **82**, **93**, **94**, **95**, **96**) were included in present calculation.

The plot of predicted versus experimental values of fluorophilicity shown in Fig. 1 suggests that the 116 compounds roughly follow a straight line. Fig. 2 shows the residuals in terms of the experimental fluorophilicities, and demonstrates that the best molecular descriptors given in Eq. (2) lead to a model that follows a normal distribution and that does not obey any kind of undesired pattern that would probably suggest the presence of non-modelled factors contributing to the fluorophilicity. The correlation matrix for Eq. (2) (indicated in Table 3) reveals that there exists some degree of intercorrelation between SEigp and Har ($R_{ij} = 0.9738$), although these descriptors carry some non-overlapping structural information that makes the model to exhibits adequate predictive l-10%-o cross-validation parameters measured on 100,000 randomly generated cases of compounds exclusion. Figs. 3–10 include the histograms of the experimental property and of each molecular descriptor selected, revealing the distribution of the chemical compounds in the different numerical intervals of the descriptor variation.

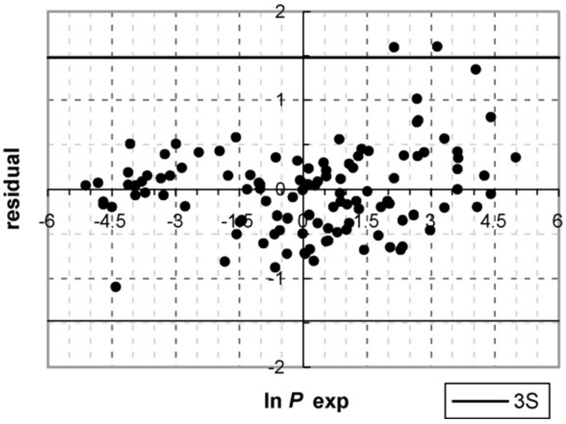

Fig. 2. Dispersion plot of the residuals for Eq. (2).

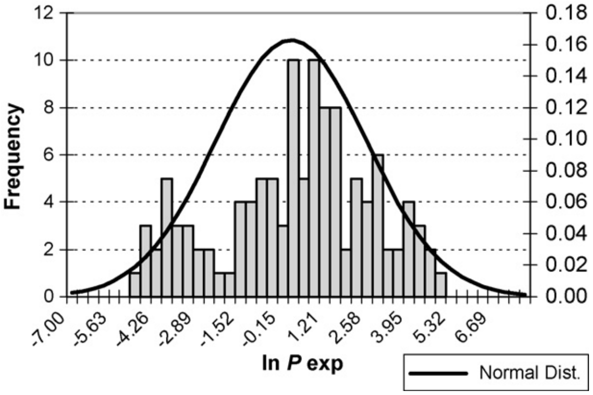

Fig. 3. Histogram of experimental data.

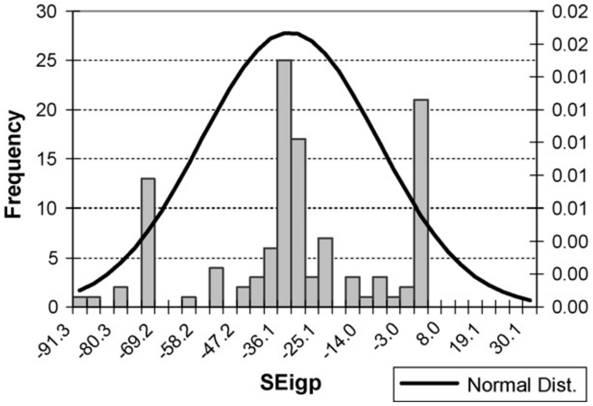

Fig. 4. Histogram of SEigp descriptor.

Present QSPR equation includes different theoretical descriptors derived from the molecular graph (G) and the three-dimensional geometry, as Table 2 shows. The numerical characterization of the molecular structure is given by: (a) four topologicals: SEigp, the eigenvalue sum from polarizability weighted distance matrix; MAXDP, the maximal electrotopological positive variation; Har, the Harary index; CIC1, the complementary information content (neighborhood symmetry of first-order); (b) a 2D-autocorrelation: MATS1v, the Moran autocorrelation-lag 1/weighted by atomic van der Waals volumes; (c) a radial distribution function: RDF055p, −5.5 weighted by atomic polarizabilities; (d) an aromaticity index: HOMA, the armonic oscillator model of aromaticity index.

It is possible to argue some structural interpretation for the numerical variables appearing in Eq. (2). In Chemical Graph Theory, the distance matrix introduced by Harary in the 1960s [18] accounts for the ‘‘through bond’’ interactions of atoms in

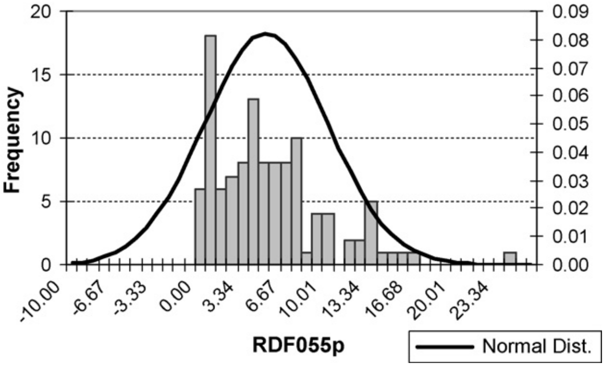

Fig. 5. Histogram of RDF055p descriptor.

Table 3
Correlation matrix for the descriptors of Eq. (2)

	SEigp	RDF055p	MAXDP	Har	CIC1	MATS1v	HOMA
SEigp	1	0.8059	0.7854	0.9738	0.4121	0.09542	0.09487
RDF055p		1	0.6357	0.8496	0.4684	0.2885	0.1510
MAXDP			1	0.7283	0.0242	0.1353	0.2478
Har				1	0.5284	0.1988	0.1868
CIC1					1	0.2766	0.09254
MATS1v						1	0.5124
HOMA							1

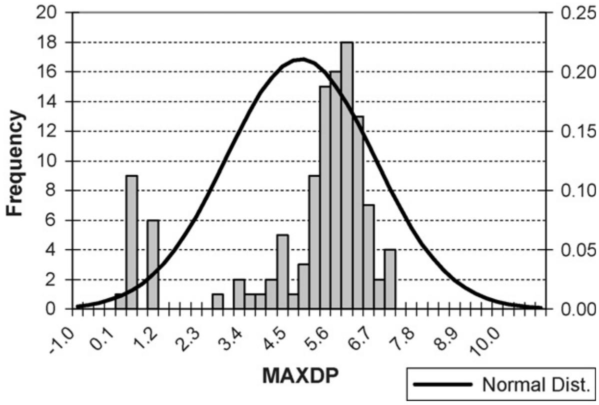

Fig. 6. Histogram of MAXDP descriptor.

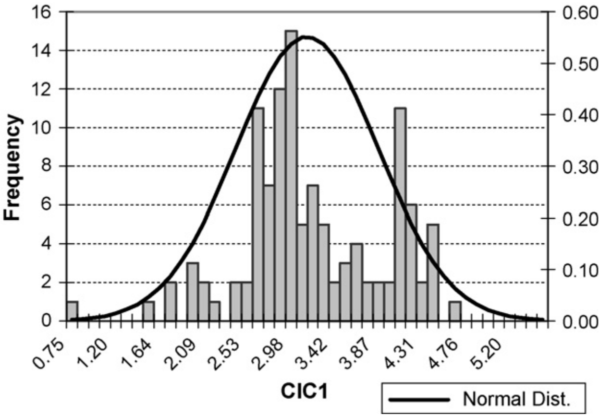

Fig. 8. Histogram of ClC1 descriptor.

molecules. The molecular descriptor SEigp characterizes the distribution of the topological distances in G and differentiates the nature of atoms through the atomic polarizability values. Another descriptor from this equation is the index Har, which contemplates in its calculation the reciprocal entries of the distance matrix. This results from the fact that the interactions among atoms decrease as the distance between them increases. The topological descriptor CIC1 is obtained from the Information Theory [19], and this sort of theoretical descriptor measures the complexity of the molecule in terms of the diversity of elements that includes in its chemical structure, such as the type of atoms, bonds, cycles, etc. It expresses the molecular symmetry by measuring the neighborhood of the atoms (through the value of the vertex degrees) located at a first-order distance (one single bond) of a considered atom, for each vertex in G. The variable MAXDP is derived from the hydrogen-depleted molecular graph and obtained from the Kier–Hall intrinsic states of atoms as [20]:

$$\mathrm{MAXDP} = \max_i |\Delta I_i| \tag{4}$$

where ΔI_i is the field effect on the ith atom due to the perturbation of all other atoms as defined by Kier and Hall:

$$\Delta I_i = \sum\nolimits_i \frac{(I_i - I_j)}{(d_{ij} + 1)} \tag{5}$$

The sum runs over all the other atoms in the molecular graph, I is the atomic intrinsic state and d_{ij} is the topological

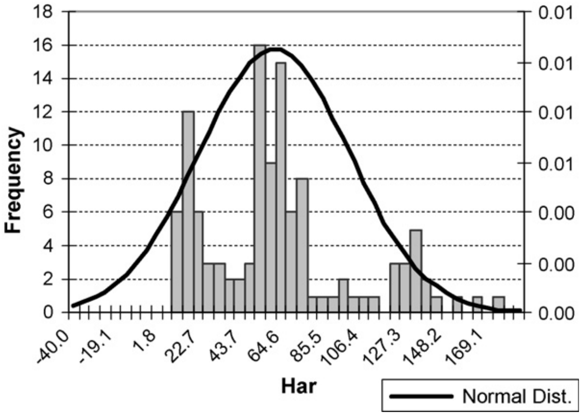

Fig. 7. Histogram of Har descriptor.

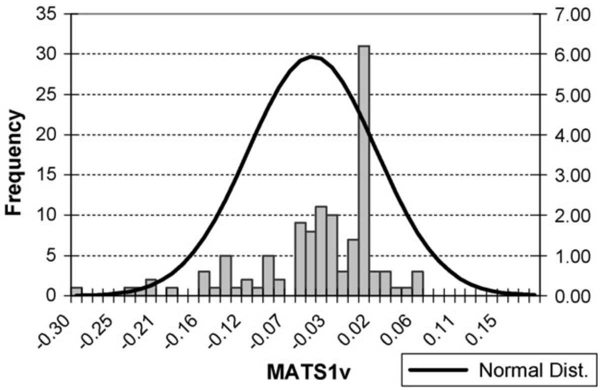

Fig. 9. Histogram of MATS1v descriptor.

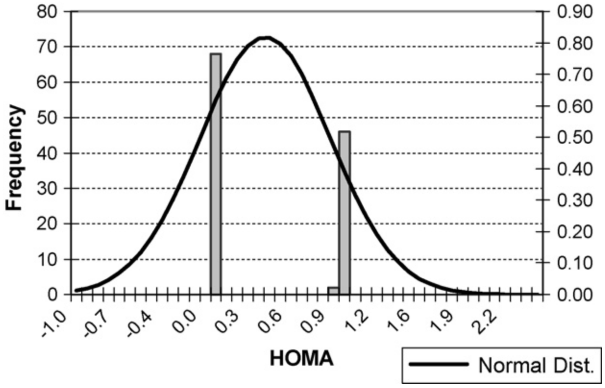

Fig. 10. Histogram of HOMA descriptor.

distance between the two considered atoms. The intrinsic state of an atom is calculated as the ratio between the Kier–Hall atomic electronegativity and the vertex degree, i.e. the number of bonds of an atom, encoding information related to both atomic partial charges and their topological position relative to the whole molecule. Therefore, MAXDP represents the maximum positive intrinsic state difference and can be related to the electrophilicity of the molecule. This fact reveals the importance of describing the fluorous electronegative contributions to fluorophilicity, in complete agreement with the empirical criteria for the design of fluorophilic compounds.

The structural variables introduced by Moran [21] correspond to bi-dimensional autocorrelations between pairs of atoms in the molecule, and are defined in order to reflect the contribution of a considered atomic property to the property being investigated. These can be readily calculated, i.e.: by summing products of atomic weights of the terminal atoms of all the paths of a prescribed length. For the case of MATS1v, the path connecting a pair of atoms has a bond length and involves the van der Waals volumes as weighting scheme to distinguish their nature. Another optimal molecular descriptor selected by the RM is RDF055p. The radial distribution function [22] of an ensemble of atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of certain radius. For RDF055p, the sphere radius is 5.5 Å and atomic polarizabilities are employed as weighting scheme. Finally, the aromaticity index HOMA [23,24] considers the presence/absence of aromatic rings in the set of compounds, measuring thus the tendency of a bond length to be comprised between a single and a double bond length. HOMA takes the value 0 for an hypotetic non-aromatic system, while it equals 1 in case of aromaticity.

The standardization of the regression coefficients [14] in Eq. (2) enables to assign more importance to the variables of the model exhibiting larger absolute standardized coefficients (shown in parentheses), therefore achieving the following ranking of contributions to ln *P*:

$$\underset{(3.398)}{\text{Seigp}} > \underset{(3.053)}{\text{Har}} > \underset{(0.348)}{\text{CIC1}} > \underset{(0.327)}{\text{RDF055p}} > \underset{(0.320)}{\text{HOMA}} > \underset{(0.178)}{\text{MATS1v}} > \underset{(0.041)}{\text{MAXDP}} \qquad (6)$$

From Eq. (2) it can be concluded that increased numerical values of descriptors such as CIC1, RDF055p, HOMA and

Table 4
Prediction of unknown fluorophilicity

N	Name	Eq. (2)
1	$(R_{f8})_3P$	5.84
2	1,3,4-$(R_{f8})_3C_6H_3$	5.21
3	*p*-$R_{f8}C_6F_4R_{f8}$	4.85
4	$R_{f11}CF{=}CF_2$	4.61
5	$R_{f9}CF_3$	4.47
6	$R_{f9}(CH_2)_2NH(CH_2)_2R_{f9}$	4.45
7	*p*-$R_{f8}C_6H_3FR_{f8}$	4.19
8	1,3,5-$(R_{f4})_2C_6H_3R_{f8}$	4.16
9	$(R_{f7}CH_2)_3P$	4.10
10	*m*-$R_{f16}C_6H_3CHO$	3.97
11	*m*-$R_{f16}C_6H_3CH_2OH$	3.69
12	*m*-$R_{f16}C_6H_3CO_2Me$	3.64
13	1,3-$(R_{f8})_2$-5-$CO_2CF_3C_6H_3$	3.05
14	$CH_3CH_2NH(CH_2)_3R_{f20}$	3.01
15	1-R_{f16}-2,3-F_2-5-$CH_2OHC_6H_2$	2.94
16	*m*-$R_{f17}C_6H_3CO_2Me$	2.55
17	$R_{f9}(CH_2)_2NH(CH_2)_2R_{f2}$	2.31
18	$(CF_3)_3CO(CF_2)_2NF_2$	2.18
19	*o*-$R_{f10}(CH_2)_3C_6H_4(CH_2)_3R_{f3}$	1.68
20	$CH_3(CH_2)_2NCH_3(CH_2)_3R_{f16}$	1.55
21	$R_{f8}(CH_2)_3NH(CH_2)_3CF_3$	1.07
22	$F_5C_6(CF_2)_2SiOC_8F_{15}$	0.95
23	$CF_3SC_6(CF_3)_5$	0.68
24	*o*-$R_{f3}(CH_2)_3C_6F_4(CH_2)_3R_{f2}$	0.55
25	*o*-$R_{f4}(CH_2)_3C_6H_4(CH_2)_3R_{f3}$	0.32
26	*o*-$R_{f3}(CH_2)_3C_6H_4(CH_2)_3R_{f3}$	−0.04
27	1,3-$(R_{f5})_2$-5-$(CH_2)_2SiOC_8H_{15}C_6H_3$	−0.12
28	*o*-$R_{f3}(CH_2)_3C_6H_4(CH_2)_3R_{f2}$	−0.47
29	1,3-$(R_{f5})_2$-2,4-F_2-5-$(CH_2)_2SiOC_8H_{15}C_6H$	−0.73
30	$R_{f8}C(S)NHCH(Me)Ph$	−0.76
31	Pentafluoroethanol	−0.81
32	7,10-R_{f4}-hexadec-1-ene	−1.63
33	1-R_{f7}-4-$(CH_2)_2SiOC_8H_{15}C_6H_4$	−2.06
34	$Ph(CH_2)_2SiF_3$	−2.08
35	4-F-1-$CF_3C_6H_4$	−2.35
36	1,3,5-Trifluorobenzene	−2.44
37	1,13-Difluorotridecane	−2.90
38	1-$(CF_2)_4CF_2H$-4-$(CH_2)_2SiOC_8H_{15}C_6H_4$	−2.92
39	6-Fluoroundecane	−2.97
40	1,14-Difluorotetradecane	−3.09
41	1-Fluorododecane	−3.10
42	*Cis*-1,2-difluorododec-1-ene	−3.13
43	6-Fluorodec-1-ene	−3.24
44	2-Fluoroundec-1-ene	−3.28
45	1,1-Difluorotridec-1-ene	−3.30
46	10-R_{f4}-hexadec-1-ene	−3.34
47	Propylbenzene	−3.41
48	$F_5C_6(CH_2)_2SiOC_8H_{15}$	−3.72
49	$F_5C_6(CF_2)_2SiOC_8H_{15}$	−3.82
50	$F_5C_6(CFH)_2SiOC_8H_{15}$	−3.88
51	Tetradec-1,13-ene	−4.33
52	PhSiOPh	−4.34
53	Hexadec-1,5,9,13-ene	−4.46
54	$CH_2{=}CH(CH_2)_{10}C_6H_5$	−4.60
55	1,2,3,4-F_4-6-$(CH_2)_2SiOC_8H_{15}C_6H$	−4.67
56	$F_5C_6CFHCH_2SiOC_8H_{15}$	−4.67
57	$C_8H_{15}SiOC_8H_{15}$	−4.70
58	$Ph(CH_2)_2SiOPh$	−4.82
59	1,2,3-F_3-5-$(CH_2)_2SiOC_8H_{15}C_6H_2$	−4.83
60	2,5-Cyclohexadienone	−4.86
61	1,2,3-F_3-4-$(CH_2)_2SiOC_8H_{15}C_6H_2$	−4.88
62	3-Cyclohexenol	−5.01
63	1,2-F_2-4-$(CH_2)_2SiOC_8H_{15}C_6H_3$	−5.03
64	1,2-F_2-3-$(CH_2)_2SiOC_8H_{15}C_6H_3$	−5.04

Table 4 (*Continued*)

N	Name	Eq. (2)
65	Hexadec-1,3,5,7,9,11,13,15-ene	−5.11
66	*m*-$FC_6H_4(CH_2)_2SiOC_8H_{15}$	−5.18
67	Icosane	−5.21
68	*o*-$FC_6H_4(CH_2)_2SiOC_8H_{15}$	−5.24
69	1,3,5-Trihydroxybenzene	−5.43

R_{fn} refers to $(CF_2)_{n-1}CF_3$.

MAXDP (with positive regression coefficients) and decreasing values for the descriptors SEigp, Har and MATS1v (with negative coefficients) would tend to predict higher fluorophilicities. The ranking of contributions given by Eq. (6) suggest that the distribution of topological distances in the molecules under investigation, expressed by descriptors such as SEigp and Har, plays an essential role that influences the values of ln *P*.

As a practical application of the model obtained above we predict the fluorophilicity of some non-yet synthesized structures. Table 4 shows 69 compounds sorted according to their fluorofilicity values. Present theoretical analysis reveals greatly fluorophilic organic chemicals: $(R_{f8})_3P$(ln *P* = 5.84), 1,3,4-$(R_{f8})_3C_6H_3$ (ln *P* = 5.21), *p*-$R_{f8}C_6F_4R_{f8}$ (ln *P* = 5.85), that in principle could be synthesized and employed as new candidates for synthesis and catalysis.

3. Conclusions

We have derived an alternative structure–fluorophilicity relationship presenting good predictive performance in a training set composed of 116 organic compounds. Some of the molecules included in the analysis were synthesized in a recent study [9,13]. The statistical parameters of present QSPR model compare fairly well with other ones reported previously in the literature based on LFER and MOD [9,11]. The best theoretical descriptors appearing in the final equation are able to reflect the molecular size, symmetry, aromaticity, as well as the importance of the fluorous-content in the organic compounds under investigation.

4. Experimental

4.1. General procedures

The structures of the compounds are firstly pre-optimized with the molecular mechanics force field (MM+) procedure included in Hyperchem version 6.03 [25], and the resulting geometries are further refined by means of the semiempirical method Parametric Method-3 (PM3). We chose a gradient norm limit of 0.01 kcal/Å for the geometry optimization.

We derive a set of 1268 molecular descriptors including several types of variables: constitutional, topological, geometrical, charge, GEometry, Topology and Atoms-Weighted AssemblY (GETAWAY), Weighted Holistic Invariant Molecular descriptors (WHIM), 3D-Molecular Representation of Structure based on Electron diffraction (3D-MoRSE), molecular walk counts, BCUT descriptors, 2D-autocorrelations, aromaticity indices, Randic molecular profiles, radial distribution functions, functional groups and atom-centred fragments [26]. To this end we resort to the free-software Dragon version 3.0 available in the Web [27]. We excluded from our calculations the empirical and property-based descriptors provided by the software. Furthermore ten constitutional descriptors and four quantum-chemical descriptors (molecular dipole moments, total energies, and homo–lumo energies), not provided by the program, were added to the pool.

As it is custommary in this sort of studies, we validate the predictive power of the model. The theoretical validation practiced on our models is based on the leave-more-out cross-validation procedure (l-*n*%-o) [28], with *n*% representing the number of molecules removed from the training set. The number of cases for random removal in l-*n*%-o are 100,000. We proceed now to describe briefly the two different techniques employed for searching the best descriptors via linear regression algorithms.

4.2. The forward stepwise regression

It is our purpose to search between a large number of *D* descriptors for an optimal subset of *d* descriptors that minimize the standard deviation *S*. Therefore, *d* will be the number of descriptors which are used for constructing a model. In other words, we want to obtain the global minimum of $S(\mathbf{d})$ where **d** is a point in a space of $D![d!(D-d)!]$ ones. A full search (FS) of optimal variables requires $D![d!(D-d)!]$ linear regressions. The forward stepwise regression (FSR) [14] consists of a step by step addition of descriptors to the model, initially without any independent variable present in the regression, until there is no variable left outside the equation that minimizes its *S*.

Here we define the standard deviation as follows:

$$S=\frac{1}{N-d-1}\sum_{i=1}^{N}\mathrm{res}_i^2 \quad (7)$$

where *N* is the number of molecules in the training set, and res_i is the residual for molecule *i* (difference between the experimental and predicted property). The FSR sacrifices accuracy for a much smaller number of linear regressions than a FS.

4.3. The replacement method

Some time ago we proposed the replacement method (RM) [15–17] that produces QSPR models that are quite close the FS ones with much less computational work. The RM gives better statistical parameters than the FSR and similar ones to the more elaborated Genetics Algorithms [29]. The RM approaches the minimum of *S* by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of *d* descriptors $d=\{X_1, X_2, \ldots, X_d\}$.

The quality of the final optimized equations obtained via the two approaches FSR and RM was compared by means of two different criteria: the Akaike criterion and the Kubinyi function. Akaike's information criterion (AIC) [30,31] considers the statistical goodness of fit and the number of parameters that have to be estimated to achieve that degree of fit:

$$\mathrm{AIC} = \left(\sum_{i=1}^{N} \mathrm{res}_i^2\right) \frac{N+d+1}{(N-d-1)^2} \tag{8}$$

Therefore, the model that produces the minimum AIC value should be considered potentially the most useful. The Kubinyi function (FIT) [32,33] closely relates to the Fisher ratio (F), although the main disadvantage of F is its sensitivity to changes in d, if d is small, and its lower sensitivity if d is large. The FIT criterion has a low sensitivity toward changes in d values, as long as they are small numbers, and a substantially increasing sensitivity for large d values. The following equation is employed:

$$FIT = \frac{R^2(N-d-1)}{(N+d^2)(1-R^2)} \tag{9}$$

where R is the correlation coefficient. The best model will present the highest value of the FIT function.

Acknowledgement

We would like to thank the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

References

[1] I.T. Horvath, J. Rabai, Science 72 (1994) 266.

[2] D.S. Jozsef Rabai, E.K. Borbas, I. Kovesi, I. Kovesdi, A.G. Antal Csampai, V.E. Pashinnik, Y.G. Shermolovich, J. Fluorine Chem. 114 (2002) 199–207.

[3] C.D.R. Rocaboy, B.L. Bennett, J.A. Gladysz, J. Phys. Org. Chem. 13 (2000) 596–603.

[4] P. Bhattacharyya, B. Croxtall, J. Fawcett, D. Gudmunsen, E.G. Hope, R.D.W. Kemmitt, D.R. Paige, D.R. Russell, A.M. Stuart, D.R.W. Wood, J. Fluorine Chem. 101 (2000) 247–255.

[5] C. Hansch, A. Leo, Exploring QSAR Fundamentals. Applications in Chemistry and Biology, American Chemical Society, Washington, DC, 1995.

[6] A.R. Katritzky, V.S. Lobanov, M. Karelson, Chem. Rev. Soc. 24 (1995) 279–287.

[7] N. Trinajstic, Chemical Graph Theory, CRC Press, Boca Raton, FL, 1992

[8] L.E. Kiss, I. Kövesdi, J. Rábai, J. Fluorine Chem. 108 (2001) 95.

[9] F.T.T. Huque, K. Jones, R.A. Saunders, J.A. Platts, J. Fluorine Chem. 115 (2002) 119–128.

[10] P.R. Duchowicz, F.M. Fernández, E.A. Castro, J. Fluorine Chem. 125 (2004) 43–48.

[11] E. de Wolf, P. Ruelle, J. van den Brocke, B. Deelman, G. van Koten, J. Phys. Chem. 108 (2004) 1458.

[12] M.S. Daniels, R.A. Saunders, J.A. Platts, J. Fluorine Chem. 125 (2004) 1291–1298.

[13] D. Szabó, J. Mohl, A.M. Bálint, A. Bodor, J. Rábai, J. Fluorine Chem. 127 (2006) 1496–1504.

[14] N.R. Draper, H. Smith, Applied Regression Analysis, John Wiley & Sons, New York, 1981.

[15] P.R. Duchowicz, E.A. Castro, F.M. Fernández, M.P. González, Chem. Phys. Lett. 412 (2005) 376–380.

[16] P.R. Duchowicz, E.A. Castro, F.M. Fernández, Commun. Math. Comput. Chem. (MATCH) 55 (2006) 179–192.

[17] A.M. Helguera, P.R. Duchowicz, M.A.C. Pérez, E.A. Castro, M.N.D.S. Cordeiro, M.P. González, Chemometr. Intell. Lab. 81 (2006) 180–187.

[18] F. Harary, Graph Theory, Addison-Wesley, Miami, 1969.

[19] D. Bonchev, Information Theoretic Indices for Characterization of Chemical Structures, RSP-Wiley, Chichester, UK, 1983.

[20] L.B. Kier, L.H. Hall, J.W. Frazer, J. Math. Chem. 7 (1991) 229.

[21] P.A.P. Moran, Biometrika 37 (1950) 17–23.

[22] M.C. Hemmer, V. Steinhauer, J. Gasteiger, Vib. Spectrosc. 19 (1999) 151.

[23] T.M. Krygowski, J. Chem. Inf. Model. 33 (1993).

[24] T.M. Krygowski, M. Cyrañski, Chem. Rev. 101 (2001).

[25] HYPERCHEM (Hypercube) available from <http://www.hyper.com>.

[26] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley VCH, 2002.

[27] DRAGON, Web 3.0 available from <http://www.disat.unimib.it/chm>.

[28] D.M. Hawkins, S.C. Basak, D. Mills, J. Chem. Inf. Model. 43 (2003) 579.

[29] S.S. So, M. Karplus, J. Med. Chem. 39 (1996) 1521.

[30] H. Akaike, in: B.N. Petrov, F. Csáki (Eds.), Second International Symposium on Information Theory, Akademiai Kiado, Budapest, (1973), pp. 267–281.

[31] H. Akaike, IEEE Trans. Automat. Control AC-19 (1974) 716.

[32] H. Kubinyi, Quant. Struct. Act. Relat. 13 (1994) 393–401.

[33] H. Kubinyi, Quant. Struct. Act. Relat. 13 (1994) 285–294.