A QSPR Study of the *p* Solute Polarity Parameter to Estimate Retention in HPLC

Ramón Bosque and Joaquim Sales*

Departament de Química Inorgànica, Universitat de Barcelona, Martí i Franquès, 1, 08028-Barcelona, Spain

Elisabeth Bosch and Martí Rosés

Departament de Química Analítica, Universitat de Barcelona, Martí i Franquès, 1, 08028-Barcelona, Spain

M. C. García-Alvarez-Coque and J. R. Torres-Lapasió

Departament de Química Analítica, Universitat de València Dr. Moliner, 50, 46100-Burjassot, València, Spain

Received April 10, 2003

A Quantitative Structure−Property Relationship (QSPR) model is developed to calculate the solute polarity parameter *p* of a set of 233 compounds of a very different chemical nature. The proposed model, derived from multiple linear regression, contains four descriptors calculated from the molecular structure and the well-known hydrophobicity parameter log $P_{o/w}$. According to the statistics obtained with the prediction set, the model has a very good prediction capacity ($R^2 = 0.954$, $F = 889$, $n = 45$, and SD $= 0.27$). The study shows that log $P_{o/w}$ and hydrogen bond acidity of the solutes are the most relevant descriptors to predict *p* values. This *p* parameter is embodied in a general equation to predict retention in reversed-phase liquid chromatography (RP-HPLC). It describes analyte retention exclusively on the basis of mobile phase/analyte/stationary phase polar interactions. Equations and procedures to determine polarity of both chromatographic phases had been successfully developed previously. Therefore, the proposed QSPR model for *p* estimation becomes a very useful tool in RP-HPLC optimization of procedures and methods in the everyday analytical work.

INTRODUCTION

Liquid chromatography is a very powerful analytical technique widely used in a variety of application fields, such as environmental, food and drug analysis, and many others. Several empirical and theoretical models have been proposed to optimize the experimental working conditions, mainly devoted to select the best composition of the mobile phase.[1] For practical purposes, a previous selection of the column, organic modifier of the mobile phase, and several measurements in the selected system are necessary to build up an empirical model. Therefore, the derived optimization equations are valid only for the chromatographic arrangement under study. Among the theoretical approaches to predict retention in reversed phase liquid chromatography (RP-HPLC), a general model useful for binary modifier/water eluents, was selected in this study. It explains the retention of the analyte exclusively on the basis of mobile phase/analyte/stationary phase polar interactions. The following general equation, widely tested, summarizes the chosen model

$$\log k = (\log k)_0 + p(P_m^N - P_s^N) \qquad (1)$$

where log *k* refers to the analyte retention factor. Other parameters account for the polarity of mobile phase (P_m^N), stationary phase ((log $k)_0$ and P_s^N), and analyte (*p*).[2−6]

P_m^N was developed for acetonitrile/water and methanol/water mixtures, which are very common mobile phases. It is closely related to the well-known E_T^N polarity parameter and can be easily calculated from the mobile phase composition solely.[3] Polarity parameters of the stationary phase can be determined from the retention of several compounds.[3,5] The polarity of the analyte, *p* quantity,[2,3] mainly gathers the polarity contribution of the solute, but it is actually a relative magnitude, since it depends also on the environment inside the column. Nevertheless, *p* values determined in a particular chromatographic system correlate linearly with those from another one. Therefore, a reference chromatographic system to refer *p* values obtained in different column/mobile phase arrangements was established.[5] This reference data set was selected from the retention measurements of Smith and Burr in a Spherisorb ODS-2 column and acetonitrile/water mobile phases.[7−12] Thus, a wide database of more than 200 compounds of a different nature is now available.[5,6] Equation 1 and the described polarity parameters have demonstrated to be able to transfer successfully retention data between solvent systems and between columns.[5]

For any chromatographic system, mobile phase polarity is easily calculable, and stationary phase parameters can be determined from a few retention measurements.[3,5] Since the calculation of *p* needs knowledge of the retention in a certain number of mobile phases for each analyte, it is convenient to find out a direct way to predict this parameter without developing new experiments. At the moment, the only computational approach proposed for this purpose is a multilinear relationship between *p* values and the Abraham parameters and/or solute hydrophobicity.[6] These models constitute useful tools to predict *p* values for new compounds,

* Corresponding author phone: +34 934021266; fax: +34 934907725; e-mail: joaquim.sales@qi.ub.es.

10.1021/ci0340660 CCC: $25.00
Published on Web 06/19/2003

and, in addition, they allow a physicochemical interpretation of retention. According to it, the most significant contribution to *p* parameter is the solute hydrophobicity expressed as octanol/water partition coefficient, log $P_{o/w}$, and the Abraham effective hydrogen bond acidity, *A*. Values of log $P_{o/w}$ are known for a vast number of compounds or, at least, are easily calculable by commercial software programs. However, *A* values are known only for a limited set of analytes. This fact prevents the evaluation of *p* values for many compounds and, consequently, the prediction of their chromatographic retention. Thus, an approach able to estimate the *p* value of a compound without any other requirement apart from its chemical structure should be very useful in predicting retention by means of eq 1.

Quantitative Structure−Property Relationship approach (QSPR) has become very useful in the prediction and interpretation of several physical and chemical properties. It seems to be an appropriate tool to study the solute polarity since it is reasonable to suppose that its numerical value depends mainly on the molecular structure. The basis of this methodology is the assumption that the behavior of compounds, as expressed by any measured physical or chemical property, can be correlated with molecular features of the compounds termed descriptors. While common approaches often need some intuitive vision to derive the relevant mathematical relationship, QSPR methods are based on statistically determined linear or nonlinear functional forms that relate the property of interest with descriptors. These descriptors are numerical values that account for the molecular structure characteristics related to the property being studied. Recently, Katritzky et al.[13] have reviewed the applications of the QSPR approach to technologically relevant physical properties, including chromatographic retention parameters.

Chromatography can readily yield a great amount of precise and reproducible data when the experimental conditions are kept constant. In this instance, solute structure becomes the single independent variable. This factor has promoted the generation of a specific class into the structure−property methodology, the so-called Quantitative Structure-Retention Relationship (QSRR).[14] Regarding HPLC, Kaliszan et al.[15] have published several papers describing retention through three analyte descriptors: total dipole moment, electron excess charge of the most negative charged atom, and the water-accessible molecular surface area. Ledesma et al.[16] studied retention data of 12 substituted PAH (polycyclic aromatic hydrocarbons) and their six unsubstituted parent compounds, which were successfully correlated with polarizability and subpolarity parameters of the solutes. More recently, Katrizky et al.[17] studied the RP-HPLC lipophilicities of 73 hydantoin derivatives, using the CODESSA program. They obtained good correlations with three descriptors: the rotational entropy (300 K), the charged solvent accessible area of nitrogen atoms, and the minimum Coulombic interaction for a C−N bond.

In this paper, we build a QSPR model for solute polarity parameter, *p*, for a large variety of solutes. The set studied contains 233 compounds of different chemical nature, which makes up a representative selection of common analytes, including molecules with different size, shape, substituents, and chemical properties. Therefore, a robust model to be used for *p* estimation and prediction of RPLC retention is proposed.

DATA AND COMPUTATIONAL METHODS

Data Set. The set of solutes contains 233 compounds of very different chemical characteristics (Table 1). These included mainly aromatic derivatives such as phenols, anilines, phenones, halobenzenes, nitrobenzenes, alkylbenzenes, and esters and also several aliphatic compounds such as alkanes and alcohols. The *p* values for 152 compounds were calculated from retention data published by Smith and Burr, which were determined in a Spherisorb ODS-2 (100 × 5 mm) column using several acetonitrile−water mobile phases buffered at pH 7.[7−12] For other 63 compounds, *p* values were calculated from the data published by Hanai and Hubert obtained in ERC-1000 ODS (150 × 6 mm), Develosil ODS-5 (150 × 4.6 mm), or Unisil Q C_{18} (150 × 4.1 mm) columns and acetonitrile−water mobile phases[18−20] and by Rosés and Bosch in a LiChrospher 100 RP-18 (100 × 5 mm) column, using acetonitrile−water mobile phases too.[2] All these values were transferred to the Smith and Burr chromatographic system by linear regression.[5] Eighteen other *p* values were calculated from the data of Kaibara et al.[21] in a Nucleosil C_{18} (150 × 6 mm) column, using methanol as the organic modifier of the mobile phases. These values were also referred to the Smith and Burr chromatographic column, p_{MeOH}, and finally, were transferred to acetonitrile−water conditions, p_{MeCN}, through eq 2[5]

$$p_{\mathrm{MeCN}} = 0.979\, p_{\mathrm{MeOH}} - 0.077 \qquad (2)$$

The *p* data set ranged from 0.77 to 7.79 with a mean value of 3.80. This data set was split randomly into a 188 member working set and an external prediction set of 45 compounds.

Structural Descriptors. The generation of the QSPR descriptors and the multilinear regression analysis were performed with the CODESSA[22] and SPSS[23] programs, respectively. The structures of the compounds were drawn with HyperChem Lite and exported in a file format suitable for MOPAC. The geometry optimization was performed with the semiempirical quantum method AM1[24] using the MOPAC 6.0 program.[25] All the geometries were fully optimized without symmetry restrictions. In all instances frequency calculations were performed in order to ensure that all the calculated geometries correspond to true minima. The MOPAC output files were used to calculate about 600 descriptors by the CODESSA program. CODESSA computes five classes of descriptors: constitutional, that reflect only the molecular composition of the compound (number of various types of atoms and bonds, number of rings, molecular weight, etc.); topological, that describe the atomic connectivity in the molecule (Wiener index, Randic indices, Kier&Hall shape indices, etc.); geometrical, that represent structural molecular descriptors, derived from the three-dimensional coordinates of the atoms in the given molecule (moments of inertia, molecular volume, molecular surface area, etc.); electrostatic, that reflect characteristics of the charge distribution of the molecule (minimum and maximum partial charges, polarity parameter, charged partial surface area descriptors, etc.); and quantum-chemical, that are obtained from MOPAC calculations (reactivity indices, dipole moment, HOMO and LUMO energies, etc.).

Procedures. The heuristic multilinear regression procedures available in the framework of CODESSA program were used to find the best correlation models. These

Table 1. Experimental and Calculated *p* for the Working and Prediction Sets

no.	compound	exptl	calc	error^b
1	1,2,3,4-tetrachlorobenzene	6.18	6.12	1.1
2	1,2,3,5-tetrachlorobenzene	6.40	6.14	4.2
3	1,2,4,5-tetrachlorobenzene	6.33	6.08	4.0
4	1,2,4-trichlorobenzene	5.69	5.60	1.7
5	1,2-dichlorobenzene	4.91	5.09	3.6
6	1,2-dihydroxybenzene	1.79	1.56	14.9
7	1,2-dimethylbenzene	4.83	4.89	1.3
8	1,3,5-trichlorobenzene	6.01	5.79	3.9
9	1,3-dichlorobenzene	5.13	4.93	4.2
10	1,3-dihydroxybenzene	1.36	1.47	7.5
11	1,3-dimethylbenzene	4.94	4.96	0.5
12	1,4-dichlorobenzene	5.01	4.82	4.0
13	1,4-dihydroxybenzene	1.17	1.32	11.3
14	1,4-dimethylbenzene	4.95	4.90	0.9
15	1-chloro-4-nitrobenzene	3.63	4.10	11.6
16	1-naphthoic acid	1.84	3.31	44.4
17	1-naphthylamine	3.16	3.24	2.6
18	1-phenyl-2-butanone	3.73	3.63	2.8
19	1-phenyl-2-propanol	2.90	2.70	7.5
20	2,3,4,5-tetrachlorophenol	4.68	4.59	2.0
21	2,3,5,6-tetrachlorophenol	4.74	4.41	7.4
22	2,3,5,6-tetramethylphenol	4.22	4.40	4.1
23	2,3,5-trichlorophenol	3.68	4.44	17.1
24	2,3,5-trimethylphenol	3.68	3.72	1.2
25	2,3,6-trichlorophenol	4.01	4.36	8.1
26	2,3,6-trimethylphenol	3.83	3.51	9.0
27	2,3-dichlorophenol	3.38	3.56	4.9
28	2,3-dimethylphenol	3.27	3.22	1.7
29	2,4,5-trichlorophenol	4.13	4.15	0.4
30	2,4,6-trichlorophenol	4.24	4.32	1.8
31	2,4-dibromophenol	3.85	3.94	2.2
32	2,4-dichlorophenol	3.55	3.49	1.9
33	2,4-dinitrophenol	2.84	2.75	3.4
34	2,5-dichlorophenol	3.51	3.46	1.5
35	2,5-dimethylphenol	3.34	3.23	3.3
36	2,5-dinitrophenol	2.97	2.87	3.4
37	2,6-dichlorophenol	3.46	3.32	4.1
38	2,6-dimethyl-4-nitrophenol	2.66	3.65	27.2
39	2,6-dimethylphenol	3.44	3.24	6.1
40	2,6-dinitrophenol	2.98	2.56	16.5
41	2-aminophenol	2.02	1.26	60.9
42	2-bromo-4-methylphenol	3.50	3.71	5.7
43	2-bromoaniline	3.60	3.38	6.6
44	2-bromophenol	3.01	3.17	5.1
45	2-chlorophenol	2.90	2.85	1.9
46	2-chlorotoluene	4.97	4.93	0.9
47	2-ethylphenol	3.20	3.42	6.5
48	2-hydroxybenzamide	2.01	1.75	14.6
49	2-methoxyphenol	2.67	2.62	1.9
50	2-methylacetophenone	3.74	3.73	0.4
51	2-methylaniline	2.96	2.76	7.4
52	2-methylanisole	4.55	4.45	2.2
53	2-methylbenzaldehyde	3.62	3.63	0.4
54	2-naphthol	3.18	3.17	0.3
55	2-nitroaniline	3.07	3.01	2.2
56	2-phenyl-2-propanol	2.92	2.96	1.5
57	2-phenylethanol	2.61	2.28	14.7
58	2-phenylethyl bromide	4.78	4.74	0.9
59	2-phenylethyl chloride	4.59	4.62	0.7
60	2-phenylphenol	3.90	3.77	3.5
61	2-phenyltoluene	5.87	5.72	2.6
62	3,4,5-trichlorophenol	4.23	4.44	4.7
63	3,4-dichlorophenol	3.52	3.86	8.7
64	3,4-dimethylphenol	3.07	2.96	3.7
65	3,5-dichlorophenol	3.84	3.98	3.5
66	3,5-dimethylphenol	3.16	3.11	1.6
67	3-bromoaniline	3.42	3.33	2.8
68	3-bromophenol	3.16	3.27	3.4
69	3-bromotoluene	5.11	5.12	0.2
70	3-chlorotoluene	4.95	4.81	3.0
71	3-hydroxyacetophenone	2.16	1.94	11.4
72	3-hydroxybenzaldehyde	2.03	1.74	16.9
73	3-hydroxybenzonitrile	2.36	2.35	0.5
74	3-methoxyphenol	2.45	2.83	13.5
75	3-methylacetophenone	3.72	3.67	1.4
76	3-methylaniline	2.95	2.76	7.0
77	3-methylanisole	4.35	4.40	1.1
78	3-methylbenzaldehyde	3.65	3.61	1.0
79	3-methylbenzamide	2.12	2.22	4.4
80	3-methylbenzoic acid	3.03	2.92	3.7
81	3-methylbenzonitrile	3.67	3.65	0.6
82	3-nitroaniline	2.85	2.51	13.4
83	3-nitrobenzyl alcohol	2.50	1.98	26.5
84	3-nitrophenol	2.62	2.60	0.8
85	3-nitrotoluene	4.01	4.12	2.6
86	3-phenyl-1-propanol	2.92	2.76	6.0
87	3-phenyl-1-propene	5.04	4.99	1.0
88	3-phenyl-1-propionamide	2.09	2.26	7.6
89	3-phenyl-1-propionitrile	3.38	3.40	0.5
90	3-phenyl-1-propyl bromide	5.38	5.29	1.6
91	3-phenyl-1-propyl chloride	5.17	5.15	0.4
92	3-phenylphenol	3.75	3.67	2.2
93	3-phenyltoluene	5.94	5.71	4.0
94	4-aminophenol	1.17	0.73	61.1
95	4-bromoaniline	2.98	3.45	13.7
96	4-bromobenzoic acid	3.47	3.30	5.1
97	4-bromophenol	3.14	3.26	3.8
98	4-bromotoluene	5.10	5.10	0.0
99	4-chloro-2-methylphenol	3.51	3.32	5.6
100	4-chloroaniline	2.82	2.99	5.7
101	4-chlorobenzoic acid	3.31	2.99	10.8
102	4-chlorophenol	3.00	2.95	1.8
103	4-chlorotoluene	4.94	4.82	2.4
104	4-ethylphenol	3.20	3.30	3.1
105	4-fluoroaniline	2.38	2.51	5.1
106	4-fluorophenylacetic acid	2.60	2.61	0.4
107	4-hydroxyacetophenone	1.93	2.00	3.3
108	4-hydroxybenzaldehyde	1.73	1.78	3.0
109	4-hydroxybenzamide	0.77	0.53	46.7
110	4-hydroxybenzonitrile	2.12	2.27	6.4
111	4-iodophenol	3.40	3.55	4.3
112	4-methoxyphenol	2.26	2.64	14.5
113	4-methylacetophenone	3.69	3.62	2.0
114	4-methylaniline	2.96	2.75	7.5
115	4-methylbenzamide	2.17	2.22	2.5
116	4-methylbenzoic acid	3.02	2.98	1.3
117	4-methylphenol	2.84	2.74	3.8
118	4-nitroaniline	2.10	2.56	18.0
119	4-nitrobenzyl alcohol	2.40	2.04	17.5
120	4-nitrophenacyl bromide	3.49	3.33	4.9
121	4-nitrophenol	2.39	2.55	6.4
122	4-nitrotoluene	3.95	4.09	3.4
123	4-phenyl-1-butanol	3.34	3.18	4.9
124	4-phenyl-1-butyronitrile	3.82	3.82	0.1
125	4-phenyl-2-butanone	3.62	3.65	0.8
126	4-phenylphenol	3.75	3.64	3.0
127	4-phenyltoluene	6.00	5.97	0.5
128	4-*tert*-butylphenol	3.89	3.95	1.5
129	5-phenyl-1-pentanol	3.71	3.55	4.4
130	aniline	2.51	2.32	8.4
131	benz-α-anthracene	7.02	6.63	5.8
132	benzaldehyde	3.11	2.98	4.3
133	benzamide	1.69	1.78	5.1
134	benzoic acid	2.63	2.49	5.7
135	benzonitrile	3.09	3.30	6.3
136	benzyl-2-bromoacetate	3.90	4.26	8.5
137	benzyl alcohol	2.42	2.09	16.0
138	benzyl bromide	4.41	4.55	3.0
139	benzyl chloride	4.13	4.07	1.5
140	benzyl cyanide	3.20	3.21	0.2
141	biphenyl	5.45	5.35	1.9
142	bromobenzene	4.55	4.74	3.9
143	butan-1-ol	2.34	2.62	10.6
144	butylbenzene	6.13	6.02	1.9
145	butyrophenone	4.19	4.29	2.3
146	chlorobenzene	4.38	4.45	1.7
147	chrysene	6.60	6.68	1.2
148	decan-1-ol	5.65	5.81	2.8
149	dimethyl-phthalate	3.18	3.55	10.5
150	dodecan-1-ol	6.92	6.34	9.2

Table 1 (Continued)

no.	compound	exptl	calc	error[b]
151	ethyl benzoate	4.23	4.26	0.6
152	ethyl phenylacetate	3.91	4.02	2.7
153	ethylbenzene	4.99	4.94	1.1
154	fluorobenzene	3.59	3.87	7.3
155	hexachlorobenzene	7.51	7.07	6.3
156	hexan-1-ol	3.37	3.61	6.7
157	hexanophenone	5.24	5.08	3.3
158	isobutylbenzene	6.12	6.15	0.5
159	isopropylbenzene	5.46	5.39	1.3
160	methyl-2-hydroxybenzoate	3.89	3.56	9.4
161	methyl-2-methylbenzoate	4.10	4.30	4.8
162	methyl-3-hydroxybenzoate	2.48	2.60	4.5
163	methyl-3-methylbenzoate	4.13	4.32	4.3
164	methyl-3-phenylpropionate	3.99	4.04	1.3
165	methyl-4-hydroxybenzoate	2.39	2.71	11.7
166	methyl-4-methylbenzoate	4.12	4.26	3.4
167	methyl-4-phenylbutyrate	4.40	4.45	1.1
168	methyl-phenylacetate	3.55	3.62	1.8
169	methyl phenylethyl ether	3.94	4.05	2.6
170	*N*-ethylaniline	3.93	3.50	12.4
171	*N*-methylbenzamide	1.89	2.28	17.3
172	octan-1-ol	4.49	4.47	0.6
173	pentachlorophenol	5.27	5.44	3.2
174	pentan-1-ol	2.83	3.23	12.4
175	phenacyl bromide	3.66	3.67	0.3
176	phenol	2.36	2.32	1.6
177	phenylacetaldehyde	3.10	3.32	6.7
178	phenylacetamide	1.78	1.66	7.4
179	propiophenone	3.70	3.77	1.8
180	thymol	4.20	4.36	3.6
181	valerophenone	4.70	4.64	1.4
182	α−4-dibromoacetophenone	4.30	4.26	1.0
183	*n*-heptane	7.17	7.05	1.7
184	*n*-octane	7.79	7.51	3.8
185	*n*-pentane	5.94	5.96	0.3
186	*n*-propyl-4-hydroxybenzoate	3.17	3.67	13.7
187	*s*-butylbenzene	5.99	6.07	1.4
188	*tert*-butylbenzene	5.79	5.78	0.2
189	1-bromo-2-nitrobenzene[a]	3.90	4.21	7.3
190	1-phenyl-1-propanol[a]	3.10	3.02	2.7
191	1-phenyl-1-propene[a]	5.19	4.95	4.9
192	2,3,4-trichlorophenol[a]	3.97	4.23	6.2
193	2,4,6-trimethylphenol[a]	3.88	3.57	8.6
194	2,4-dimethylphenol[a]	3.20	3.19	0.4
195	2,6-dibromophenol[a]	3.80	4.02	5.5
196	2-aminophenyl[a]	3.81	3.85	1.0
197	2-bromotoluene[a]	5.11	5.11	0.1
198	2-chloro-5-methylphenol[a]	3.31	3.49	5.3
199	2-hydroxyacetophenone[a]	3.44	2.88	19.4
200	2-hydroxybenzaldehyde[a]	3.43	2.68	28.2
201	2-hydroxybenzonitrile[a]	1.97	2.37	17.1
202	2-methylbenzamide[a]	2.00	1.88	6.6
203	2-methylbenzonitrile[a]	3.59	3.77	4.8
204	2-methylphenol[a]	2.89	2.90	0.3
205	2-nitrotoluene[a]	3.88	4.03	3.7
206	2-phenyl-1-propanol[a]	2.91	2.63	10.6
207	3-aminophenol[a]	1.41	0.88	59.6
208	3-chlorophenol[a]	3.06	3.05	0.2
209	3-methylphenol[a]	2.82	2.75	2.6
210	4-chloro-3-methylphenol[a]	3.38	3.58	5.6
211	4-iodoaniline[a]	3.21	3.54	9.4
212	4-methylbenzaldehyde[a]	3.58	3.63	1.4
213	4-methylbenzonitrile[a]	3.62	3.63	0.4
214	acetophenone[a]	3.14	3.18	1.3
215	anthracene[a]	6.12	5.53	10.7
216	benzene[a]	3.85	4.06	5.2
217	benzyl acetate[a]	3.60	3.69	2.5
218	ethyl-3-phenylpropionate[a]	4.40	4.42	0.4
219	heptan-1-ol[a]	3.94	4.22	6.6
220	hepthanophenone[a]	5.78	5.51	5.0
221	iodobenzene[a]	4.88	4.99	2.1
222	methoxybenzene[a]	3.86	3.92	1.5
223	methyl benzoate[a]	3.59	3.78	5.0
224	*N*,*N*-dimethylbenzamide[a]	2.37	2.28	4.0
225	naphthalene[a]	4.82	4.76	1.2
226	nitrobenzene[a]	3.37	3.65	7.6
227	pentachlorobenzene[a]	6.95	6.59	5.5
228	phenanthrene[a]	5.73	5.65	1.5
229	propylbenzene[a]	5.56	5.44	2.2
230	pyrene[a]	6.74	5.85	15.2
231	toluene[a]	4.44	4.55	2.4
232	*n*-hexane[a]	6.51	6.42	1.4
233	*n*-hexylbenzene[a]	7.19	7.02	2.4

[a] Prediction set. [b] Error: absolute valor of 100[(calc−exptl)/exptl].

procedures provide collinearity control (i.e., any two descriptors intercorrelated above 0.8 are never involved in the same model) and implement heuristic algorithms for the rapid selection of the best correlation, without testing all possible combinations of the available descriptors. After the heuristic reduction the pool of descriptors was decreased to nearly 200. The goodness of the correlation is tested by the regression coefficient (R^2), the F-test, the standard deviation (SD), the relative standard deviation (rsd), that is, SD divided by the mean p value, and by the mean of relative errors expressed in absolute values (error). The stability of the correlations was tested against the cross-validated coefficient, R_{cv}^2, which describes the stability of a regression model obtained by focusing on the sensitivity of the model to the elimination of any single data point. The t-test and the level of significance of each coefficient as well as the standardized regression coefficients (beta) are also reported. To further validate the model other tests were performed for the descriptors, the pairwise correlations, and the variance inflation factors (VIF). The VIF values, defined as $(1\text{-}R^2)^{-1}$, were calculated to identify whether excessively high multicollinear coefficients existed among the descriptors; a VIF greater than 10 is indicative of multicollinearity.

The model which passed the statistical diagnosis with as few descriptors as possible was chosen. When adding another descriptor did not improve significantly the statistics of a model, it was determined that the optimum subset size had been achieved. The optimum model size in this study was five descriptors. Validation of the model was performed on the external prediction set of compounds withheld from the working set.

RESULTS AND DISCUSSION

The QSPR analysis of the p values for the 188 compounds of the working set (see Table 1) resulted in the five-descriptor model summarized in Table 2. The regression statistics show that the obtained correlation is good, with an overall standard deviation of 0.41, which represents a relative standard deviation of 10.9%, and an error of 9.8%.

This model contains one constitutional descriptor, the number of benzene rings; two topological descriptors, the average information content (order 0) and the Kier&Hall index (order 2); one electrostatic descriptor, the relative positive charged surface area, SA; and one descriptor related to the hydrogen bonding characteristics of the molecules,

Table 2. Five-Descriptor Correlation Equation for the Working Set[a]

descriptor	coeff	SD	*t*-test
intercept	5.87	0.27	22.12
min(#HA, #HD)	−1.00	0.06	−18.08
Kier&Hall Index (order 2)	1.02	0.06	17.44
average information content (order 0)	−1.70	0.13	−13.58
number of benzene rings	−0.61	0.07	−8.22
RPCS relative positive charged SA (Zefirov)	−0.11	0.01	−7.59

[a] $R^2 = 0.904$; $F = 344$; $n = 188$; SD = 0.41; $R_{cv}{}^2 = 0.895$.

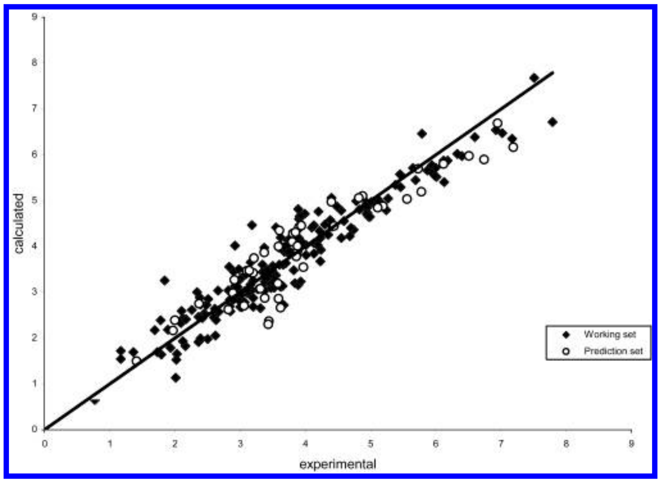

Figure 1. Plot of calculated (Table 2) vs experimental *p* of the working and prediction sets.

the minimum (#HA, #HD). The average information content descriptors[26] are defined on the basis of the Shannon information theory and are calculated as follows

$$^{c}IC = -\sum_i \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

where n_i is the number of atoms in the *i*th class and *n* is the total number of atoms in the molecule. The division of atoms into different classes depends on the coordination sphere taken into account. This leads to the indices of different order *k*. The Kier&Hall index (order 2), $^2\chi^v$, belongs to the well-known valence connectivity indices.[27] They account for the presence of heteroatoms and the hybridization of atoms in the molecule. The descriptor RPCS, [28,29] relative positive charged surface area, is the product of the solvent accessible surface area of the most positive atom by the relative positive charge (RPCG). The chemical charges in the molecule are calculated using the approach proposed by Zefirov,[30] based on the Sanderson's electronegativity scale**.** The minimum (#HA, #HD) descriptor is the minimum value of the count of hydrogen-acceptor sites and the count of hydrogen-donor sites.

Good results are also obtained with the prediction set, showing the high prediction capacity of the model. The statistical parameters are as follows: $R^2 = 0.858$; $F = 261$; $n = 45$; SD = 0.46; rsd = 11.4% and the error is 11.0%. Figure 1 shows a plot of the calculated versus observed values for all the solutes studied, the working and prediction sets.

Despite statistical results obtained being satisfactory for many QSPR purposes, from a practical point of view, standard deviations about 0.5 are too high for a useful retention factors prediction**.** The well-known dependence of chromatographic retention on solute hydrophobicity[1,31] led us to test the relationship between *p* and log $P_{o/w}$ for a set of 146 solutes for which log $P_{o/w}$ was available. Despite the evident dependence, the correlation was poor and, at least, a term expressing the hydrogen bond ability of the solute had to be added to get a reliable model.[6]

Table 3. Five-Descriptor Correlation Equation for the Working Set[a]

descriptor	coeff	SD	beta	*t*-test	sig	VIF
intercept	0.815	0.187		4.37	0.000	
log $P_{o/w}$	0.818	0.021	0.704	38.34	0.000	1.69
HDCA-2 (Zefirov)	−1.546	0.123	−0.260	−12.59	0.000	2.12
HOMO−LUMO	0.143	0.018	0.129	7.72	0.000	1.40
Pol/d^2	−1.626	0.213	−0.142	−7.64	0.000	1.72
DPSA-1 (quantum)	1.041E-03	2.63E-04	0.074	3.96	0.000	1.76

[a] $R^2 = 0.964$; $F = 964$; $n = 188$; SD = 0.25; $R_{cv}{}^2 = 0.961$.

Solute hydrogen bond capabilities are expressed by a number of descriptors generated by CODESSA, but log $P_{o/w}$ is not a primary descriptor included in it. Published QSPR models to describe log $P_{o/w}$ require more than 10 descriptors to obtain acceptable correlations for most sets of compounds.[13] Thus, to improve the QSPR correlation, we have added solute log $P_{o/w}$ value as an external descriptor in the pool of descriptors. The experimental log $P_{o/w}$ values are known for most solutes,[32] but when they are not available, values calculated with the ACD-Labs[33] software are used instead.

When log $P_{o/w}$ is added to the pool of descriptors, a new best five-descriptor QSPR model is generated with the working set. It is given in Table 3. The correlation obtained is very good, with a standard deviation of 0.25, which represents a relative standard deviation of 6.7%, and the error is 5.9%. The level of significance associated to the *t*-student coefficient shows that there is a linear relation between *p* and the descriptors. The pairwise correlations for the five descriptors ranged from 0.037 to 0.643, with an average value of 0.27, and a mean VIF value of 1.7. According to the beta values, log $P_{o/w}$ and HDCA-2 are the most significant descriptors.

Besides the log $P_{o/w}$ descriptor, this model contains two electrostatic descriptors, polarity parameter/square distance and DPSA-1, and one quantum descriptor, HOMO−LUMO gap. The fifth descriptor of the model belongs to a kind of descriptors defined to account for the possible hydrogen bonding interactions between the molecules, HDCA-2.

The first electrostatic descriptor is the polarity parameter/square distance, pol/d[2,34], which is calculated as the difference between the maximum (q_{max}) and the minimum (q_{min}) charges factorized by the square of the distance between the atoms that bear minimum and maximum partial charges. The second electrostatic descriptor is DPSA-1, and it belongs to the charged partial surface area (CPSA) descriptors proposed by Jurs et al.[28,29] These descriptors encode the features responsible for polar interactions between molecules. In general, they are calculated from the contributions related to the atomic partial charges and the molecular-accessible surface area. DPSA-1 is the difference between the partial positive, PPSA-1, and negative, PNSA-1, surface areas, which are

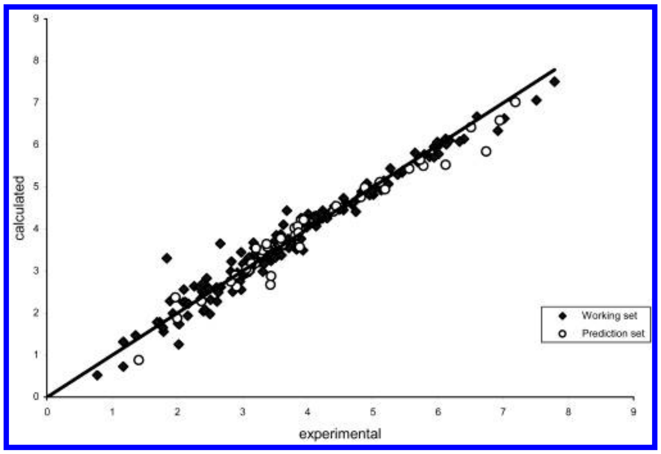

Figure 2. Plot of calculated (Table 3) vs experimental *p* of the working and prediction sets.

defined as the sum of the positively or negatively charged solvent-accessible atomic surface areas, S_a, in the molecule. In this case the charges are obtained from MOPAC calculations. The quantum descriptor is the energy gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) of the compound. The hydrogen-acidity related descriptor is HDCA-2. It also belongs to the CPSA descriptors and is defined as the area-weighted surface charge of any hydrogen bonding donor atom in the molecule with a positive partial charge. It is calculated as

$$\text{HDCA} - 2 = \sum_{\text{D}} \frac{q_{\text{D}}\sqrt{S_{\text{D}}}}{\sqrt{S_{\text{TOT}}}}$$

where q_{D} is the partial charge of the hydrogen bonding donor atom, S_{D} denotes the surface area for this atom, and S_{TOT} is the total molecular surface area calculated from the van der Waals radii of the atoms within the approximation of overlapping spheres. In this case, the charges are calculated from the Zefirov approach.

Very good results are obtained with the prediction set, showing the high prediction capacity of the model. The statistics are as follows: $R^2 = 0.954$; $F = 889$; $n = 45$; SD $= 0.27$; rsd $= 6.7\%$, and the error is 6.6%. Figure 2 shows a plot of the calculated versus observed values for all the solutes studied when log $P_{o/w}$ is included as a descriptor. Thus, the model containing log $P_{o/w}$ improves significantly the prediction of the solute polarity parameter. Calculated *p* values as well as the associated unsigned relative error are given in Table 1 for each compound.

To evaluate again the significance of selected descriptors of the model (Table 3), the variation of statistical parameters of the forward correlation, that is introducing the descriptors one by one, for the 188 solutes of the working set is given in Table 4. With log $P_{o/w}$ as the unique descriptor the correlation shows $R^2 = 0.828$ and SD $= 0.54$. The addition of the hydrogen bond descriptor HDCA-2 improves the correlation significantly up to $R^2 = 0.921$, $F = 1112$, and SD $= 0.37$. These results show the evident dependence of *p* with the hydrogen bond acidity ability of the solute. The addition of the third and the fourth descriptors, HOMO−LUMO gap and pol/d^2, respectively, improves the correlation to $R^2 = 0.960$, $F = 1267$, and SD $= 0.26$. The fifth descriptor, DPSA-1, scarcely increases the statistical parameters. Then, four descriptors are enough to explain the solute polarity parameter in QSPR model, according to the following equation

$$p_{\text{MeCN}} = 0.857 \log P_{o/w} - 1.498\,(\text{HDCA} - 2) + 0.181\,(\text{HOMO-LUMO}) - 1.627\,(\text{pol/d}^2) + 0.432 \tag{3}$$

Table 4. Statistical Parameters for Successive Models with Increasing Number of Descriptors

model	R^2	SD	R^2 change	F change
a	0.828	0.54	0.828	898
b	0.921	0.37	0.092	214
c	0.949	0.30	0.028	101
d	0.960	0.26	0.012	54
e	0.964	0.25	0.003	16

[a] Descriptors: (intercept), log $P_{o/w}$. [b] Descriptors: (intercept), log $P_{o/w}$, HDCA-2. [c] Descriptors: (intercept), log $P_{o/w}$, HDCA-2, HOMO−LUMO. [d] Descriptors: (intercept), log $P_{o/w}$, HDCA-2, HOMO−LUMO, pol/d^2. [e] Descriptors: (intercept), log $P_{o/w}$, HDCA-2, HOMO−LUMO, pol/d^2, DPSA-1.

As mentioned, the solute polarity parameter was studied by means of the Abraham general solvation equation,[6,35] using a set of 146 compounds from Smith and Burr. The molecular parameters embodied in this equation are an excess molar refraction, *E*, the solute dipolarity/polarizability, *S*, the effective hydrogen bond acidity and basicity, *A* and *B*, respectively, and the McGowan volume, *V*.

The descriptors of the QSPR model (eq 3) generated by CODESSA encode information close to that of Abraham parameters. Thus, HDCA-2, which is one of the hydrogen bond acidity descriptors included in CODESSA, is strongly related to *A*. CODESSA pol/d^2 descriptor is related to molecular polarity, whereas the physical meaning of the *S* parameter reflects the difficulty to evaluate separately dipolarity and polarizability. The HOMO−LUMO gap descriptor stands for the molecular polarizability since the electron distribution can be distorted readily if the LUMO energy lies close to the HOMO energy, generating then, a large polarizability. Descriptor *E* stands also for molecular polarizability according to the well proved dependence between molar refraction and polarizability.

To test the consistency of both approaches the correlation between *p* and log $P_{o/w}$, *A*, *S*, and *E* has been studied using those compounds in Table 1, for which the Abraham parameters are known ($n = 225$). However, the stepwise approach has pointed out that only three descriptors should be included in the correlation to keep the level of significance previously defined (<0.05). Therefore, the following three-descriptor model has been obtained

$$p_{\text{MeCN}} = 0.917 \log P_{o/w} - 1.441A - 0.367S + 2.190 \tag{4}$$

with $R^2 = 0.995$, $F = 2002$, and SD $= 0.25$. The comparison of eqs 3 and 4 shows a nice agreement in coefficients of log $P_{o/w}$ terms and between those of HDCA-2 and *A* terms. The coefficient of pol/d^2 (eq 3) is higher than that of *S* (eq 4), although both terms show negative sign. Therefore, eq 3 requires a positive polarizability term that is not necessary

Table 5. Parameters and Statistics of Regression between Calculated and Experimental *p* Values[a]

method	intercept	slope	R^2	*F*	*n*	SD
eq 4	0.13(0.05)	0.96(0.01)	0.965	6062	225	0.25
QSPR model	0.13(0.04)	0.96(0.01)	0.967	6613	225	0.25

[a] Standard deviation in parentheses.

in eq 4, which embodies dipolarity and polarizability in the S descriptor. Both models avoid any descriptor related to the hydrogen bond basicity and volume. This fact is explained for the strong correlation between log $P_{o/w}$ and these two descriptors.[36] Statistical parameters for eqs 3 and 4 are very similar showing again the consistency of both approaches.

According to eq 1, the higher the *p* value, the higher the retention of the solute. Therefore, the proposed models confirm the experimental facts that the solute hydrophobicity favors retention, whereas hydrogen bond acidity diminishes it.[1,6,28,37,38]

Despite the satisfactory results given by eq 3, a slightly more accurate estimation of *p* can be achieved using the five-descriptor QSPR model (Table 3), and this is the selected approach for practical purposes. Thus, results from QSPR model have been compared with those calculated by means of eq 4. Very similar statistical parameters are obtained by means of both methods (Table 5). The slope (0.96) of the correlation plots of calculated vs experimental *p* for both models accounts for their goodness, although the predicted values are slightly underestimated for the most hydrophobic compounds. The results confirm again the goodness of the QSPR model derived in this work, which shows the practical advantage to avoid the use of experimentally determined descriptors.

CONCLUSIONS

A quantitative structure−property relationship was derived to predict the solute polarity parameter *p*, for a diverse set of 233 compounds. The model contains five descriptors, four of them are calculated from the molecular structure of compounds, and the fifth is the hydrophobicity parameter, log $P_{o/w}$, which is known for many compounds or can be easily derived by commercial programs. Therefore, the solute polarity parameter for a wide range of compounds can be accurately predicted. The model has a squared correlation coefficient of 0.964 and a standard error of 0.25 for the working set and of 0.954 and 0.27 for the prediction set, respectively. In consistency with previous studies, the descriptors that better explain the values of *p* are log $P_{o/w}$ and the hydrogen bond acidity, HDCA-2. The addition of HOMO−LUMO energy gap and pol/d^2, encoding information about the polarizability and polarity of the compounds respectively, leads to a reliable four-descriptor model with satisfactory statistical parameters. The fifth descriptor of the model, DPSA-1, improves only slightly the correlation.

It is worthy to note that the descriptors contained in the proposed QSPR model are able to explain properly the physicochemical dependence of the solute polarity parameter *p*. This parameter is embodied in a general equation to predict retention exclusively on the basis of mobile phase/analyte/stationary phase polarity interactions. Equations and procedures to determine polarity of both chromatographic phases had been successfully developed. Therefore, proposed QSPR model for *p* estimation becomes a very useful tool in RP-HPLC optimization of procedures and methods in the everyday analytical work.

ACKNOWLEDGMENT

Financial support from the Catalan Government (Grants 2001, SGR 00052 and 00055), Valencian Government (Project CTIDIB/2002/226), and from the Spanish Government and FEDER funds (Project BQU2001-2882 and BQU2001-3047) is acknowledged. J.R.T.L. thanks the MCYT for a Ramón y Cajal position.

REFERENCES AND NOTES

(1) Poole, C. F. *The Essence of Chromatography*; Elsevier: Amsterdam, 2003.

(2) Rosés, M.; Bosch, E. Linear solvation energy relationships in reversed-phase liquid chromatography. Prediction of retention from a single solvent and a single solute parameter. *Anal. Chim. Acta* **1993**, *274*, 247−162.

(3) Bosch, E.; Bou, P.; Rosés, M. Linear description of solute retention in reversed-phase liquid chromatography by a new mobile phase polarity parameter. *Anal. Chim. Acta* **1994**, *299*, 219−229.

(4) Rosés, M.; Bou, P.; Bosch, E.; Siigur, K. A new solvent parameter for prediction of retention on HPLC. *Org. React.* **1995**, *29*, 51−53.

(5) Torres-Lapasió, J. R.; García-Alvarez-Coque, M. C.; Rosés, M.; Bosch, E. Prediction of the retention in reversed-phase liquid chromatography using solute-mobile phase-stationary phase polarity parameters. *J. Chromatogr. A* **2002**, *955*, 19−34.

(6) Torres-Lapasió, J. R.; García-Alvarez-Coque, M. C.; Rosés, M.; Bosch, E.; Zissimos, A. M.; Abraham, M. H. Analysis of a solute polarity parameter in reversed-phase liquid chromatography on a linear solvation relationships basis. Submitted for publication.

(7) Smith, R. M.; Burr, C. M. Retention prediction of analytes in reversed-phase high performance liquid chromatography based in molecular structure: I Monosubstituted aromatic compounds. *J. Chromatogr.* **1989** *475*, 57−74.

(8) Smith, R. M.; Burr, C. M. Retention prediction of analytes in reversed-phase high performance liquid chromatography based in molecular structure: II Long-term reproducibility of capacity factors and retention. *J. Chromatogr.* **1989**, *475*, 75−83.

(9) Smith, R. M.; Burr, C. M. Retention prediction of analytes in reversed-phase high performance liquid chromatography based in molecular structure: III Monosubstituted aliphatic compounds. *J. Chromatogr.* **1989**, *481*, 71−84.

(10) Smith, R. M.; Burr, C. M. Retention prediction of analytes in reversed-phase high performance liquid chromatography based in molecular structure: IV Branched and unsaturated alkylbenzenes. *J. Chromatogr.* **1989**, *481*, 85−95.

(11) Smith, R. M.; Burr, C. M. Retention prediction of analytes in reversed-phase high performance liquid chromatography based in molecular structure: V Cripes (Chromatographic retention index prediction expert system). *J. Chromatogr*. **1989**, *485*, 325−340.

(12) Smith, R. M.; Burr, C. M. Retention prediction of analytes in reversed-phase high performance liquid chromatography based in molecular structure: VI Disubstituted aromatic compounds. *J. Chromatogr.* **1991**, *550*, 335−356.

(13) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure−Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18.

(14) (a) Kaliszan, R. *Quantitative Structure-Chromatographic Retention Relationships;* Wiley: New York, 1987. (b) Kaliszan, R. *Structure and Retention in Chromatography. A Chemometric Approach;* Harwood: Amsterdam, 1997. (c) Forgács, E.; Cserháti, T. *Molecular Basis of Chromatographic Separation*; CRC: Boca Ratón, 1997.

(15) Baczek, T.; Kaliszan, R. Predictive approaches to gradient retention based on analyte structural descriptors from calculation chemistry. *J. Chromatogr. A* **2003**, *987*, 29−37, and references therein.

(16) Ledesma, E. B.; Wormat, M. J. QSRR Prediction of Chromatographic Retention of Ethynyl-Substituted PAH from semiempirically Computed Solute Descriptors. *Anal. Chem.* **2000**, *72*, 5437−5443.

(17) Katritzky, A. R.; Perumal, S.; Petrukhin, R.; Kleimpeter, E. CODESSA-Based Theoretical QSPR Model for Hydantoin HPLC−RT Lipophilicities. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 569−574.

(18) Hanai, T.; Hubert, J. Dependence of the retention of phenols upon van der Waals volume, π-energy and hydrogen bonding effects. *J. Chromatogr.* **1984**, *302*, 89−94.

(19) Hanai, T.; Hubert, J. Retention versus wan der Waals volume and π energy in liquid chromatography. *J. Chromatogr.* **1984**, *290*, 197−206.

(20) Hanai, T.; Hubert, J. Prediction of retention time of phenols in liquid chromatography. *J. High Resolut. Chromatogr. Chromatogr. Commun.* **1983**, *6*, 20−26.

(21) Kaibara, A.; Hirose, M.; Nakagawa, T. Effect of the polar functional group of the solute on hydrophobic interaction with the stationary ligand in reversed-phase high-performance liquid chromatography. *Chromatographia* **1990**, *29*, 551−556.

(22) Katritzky, A. R.; Lovanov, V. S.; Karelson, M. *CODESSA, Reference Manual V 2.13*; Semichem and the University of Florida. 1997.

(23) SPSS for Windows, version 10.0.06. SPSS Inc.

(24) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(25) Stewart, J. *P. P. MOPAC 6.0, Quantum Chemistry Program Exchange;* QCPE, No. 455, Indiana University, Bloomington, IN. 1989.

(26) Basak, S. C.; Harris, D. K.; Magnuson, V. R. Comparative study of lipohilicity versus topological molecular descriptors in biological correlations. *J. Pharm. Sci.* **1984**, *73*, 429−437.

(27) Kier, L. B.; Hall, L. H. The Nature of structure−activity relationships and their relation to molecular connectivity. *Eur. J. Med. Chem.* **1977**, *12*, 307−312.

(28) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.

(29) Stanton, D. T.; Egolf, L. M.; Jurs. P. C. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306−316.

(30) Zefirov, N. S.; Kirpichenok, M. A.; Izmailov, F. F.; Trofimov, M. I. Calculation schemes for atomic electronegativities in molecular graphs within the framework of Sanderson principle. *Dokl. Akad. Nauk SSSR* **1987**, *296*, 883−887.

(31) Abraham, M. H.; Chadha, H. S.; Leo, A. J. Hydrogen bonding: XXXV. Relationships between high performance liquid chromatography factors, and water-octanol partition coefficients. *J. Chromatogr. A* **1994**, *685*, 203−211.

(32) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR, hydrophobic, electronic and steric constants*; ACS Professional Reference Book, Washington, 1995.

(33) Advanced Chemistry Development Inc. in SciFinder, Chemical Abstracts Service, http:// www.cas.org/SCIFINDER.

(34) Kier, L. B. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science Publishers: New York, 1990.

(35) Abraham, M. H. Scales of Solute Hydrogen-Bonding: Their Construction and Application to Physicochemical and Biochemical Processes. *Chem. Soc. Rev.* **1993**, 73−83.

(36) Abraham, M. H. In *Quantitative Treatments of Solute/Solvent Interactions*; Politzer, P., Murray, J. S., Eds.; Elsevier: Amsterdam, 1994.

(37) Abraham, M. H.; Rosés, M. Hydrogen bonding. 38. Effect of solute structure and mobile phase composition on reversed-phase high performance liquid chromatographic capacity factors. *J. Phys. Org. Chem.* **1994**, *7*, 672−684.

(38) Abraham, M. H.; Rosés, M.; Poole, C. F.; Poole, S. K. Hydrogen bonding. 42. Characterization of reversed-phase high performance liquid chromatographic C_{18} stationary phases. *J. Phys. Org. Chem.* **1997**, *10*, 358−368.

CI0340660