Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography B, 804 (2004) 31–41
JOURNAL OF CHROMATOGRAPHY B
www.elsevier.com/locate/chromb

Multivariate analysis of the selectivity for a pentachlorophenol-imprinted polymer

C. Baggiani*, L. Anfossi, C. Giovannoli, C. Tozzi

Dipartimento di Chimica Analitica, Università di Torino, via P. Giuria 5, 10125 Torino, Italy

Abstract

A pentachlorophenol (PCP)-imprinted polymer (MIP) was obtained by thermal polymerization of a mixture of template, 4-vinylpyridine and ethylene glycol dimethacrylate with molar ratio 1 + 3 + 27, using as porogenic solvent methanol–water (3 + 1 (v/v)). The polymer was packed in an HPLC column and selectivity towards 52 PCP-related phenols (22-chloro-, 21-alkyl-, 4-aryl-, 3-methoxy- and 6-polyphenols) was measured using acetonitrile–acetic acid (99 + 1 (v/v)) as mobile phase. The same was made for a reference polymer obtained without pentachlorophenol (NIP). The molecular recognition properties of the imprinted polymer were expressed in terms of selectivity index (SI), calculated for each phenol as k_{NIP}/k_{MIP}. Sixteen molecular descriptors were calculated for each molecule: qO, the partial charge of the phenolic oxygen atom; qH, the partial charge of the phenolic hydrogen atom; Δq, the absolute value of the difference qO − qH; HOMO, the highest occupied molecular orbital; LUMO, the lowest unoccupied molecular orbital; Δorb, absolute value of the difference HOMO − LUMO; μ^2, the square of total dipole moment; MW, the molecular weight; SAS, the solvent-accessible molecular surface area; hSAS, the hydrophobic solvent-accessible molecular surface area; Svdw, the van der Waals molecular surface area; hSvdw, the hydrophobic part of Svdw; MOv, the molecular ovality; RG, the radius of gyration; log *P*, the logarithm of *n*-octanol–water partition coefficient; p*K*, the phenolic dissociation constant. Correlations between selectivity index and these descriptors were searched utilizing multivariate principal component analysis (PCA). The multivariate model obtained by regression on the principal components correlate collectively several of the calculated descriptors with the polymer selectivity. The magnitude of the model's parameters shows that selectivity is strongly influenced by molecular descriptors having structural character, such as MW, hSvdw and log *P*, while the effect of molecular descriptors having electronic character, such as qO and p*K*, is much less marked.

Keywords: Multivariate analysis; Selectivity; Pentachlorophenol; Molecularly imprinted polymers

1. Introduction

The binding selectivity of a molecularly imprinted polymer can be related to the spatial orientation of functional monomers around the template during the polymerization process. In fact, the number and type of functional groups in a template molecule (i.e. carboxy, hydroxy, amido, amino, etc.) able to form non-covalent interactions with functional monomers conditionate not only the strength of the resulting molecular recognition, but also the possibility to bind efficiently other molecules related to the template [1]. As significative examples, methacrylic acid-*co*-ethylene glycol dimethacrylate polymers imprinted with L-phenylalanine derivatives shown an increasing enantiomeric selectivity when template changed from L-phenylalanine ethyl ester—able to establish only an ion-pair interaction between the carboxylate and the amine and a weak hydrogen bond between the carboxyl and the ethyl ester—to L-phenylalanine ethyl amide and to L-phenylalanine anilide—able to establish the same kind of ion-pair interaction and a strong cyclic hydrogen bond between the carboxyl and the amide [2]. In a work on molecular imprinting of several peracetylated phenylgalactosides, it was shown that the presence of an amino function on the aglycon part of the template strongly influenced the polymer selectivity adding an additional interaction with the polymer [3]. Again, a large study on acrylamide-*co*-ethylene glycol dimethacrylate polymers imprinted with several protected amino acids, shown that materials imprinted with L-tyrosine derivatives better rebound the related template than L-phenylalanine derivatives, as effect of the presence of an additional interaction between the phenolic hydroxyl and the binding cavity [4].

* Corresponding author. Tel.: +39-011-6707622; fax: +39-011-6707615.
E-mail address: claudio.baggiani@unito.it (C. Baggiani).

1570-0232/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2004.01.037

It should be considered that all the reported examples are based on templates characterized by complex molecular structures, able to raise strong molecular recognition effects in an imprinted polymer on the basis of multiple non-covalent interactions. When template complexity decreases, multiple interactions are less frequent and the polymer selectivity should be based mainly on shape similarity between the template and the related molecules. As example, different porogenic solvents such as benzene, toluene and xylenes produce imprinted polymers able to recognize with a certain degree of selectivity the related solvent [5], while polymers imprinted with chlorinated phenoxyacids shown molecular recognition properties directly related to position of the chlorine atoms on the aromatic rings [6].

Even if such templates have simple structures, in the molecular recognition process it is difficult to clearly discriminate between the steric contribute due to the shape of the template, and the contribute due to the non-covalent interaction between template and functional monomer. Very few systematic studies are reported in literature. As a significative example, a recent study on the resolution of enantiomeric pairs of substituted chiral amines revealed that steric and spatial interactions markedly influenced the molecular recognition properties of molecularly imprinted polymers in a quite predictable manner [7].

Recently, several papers have been published describing the successful application as solid phase extraction materials of molecular imprinted polymers obtained using as templates very simple molecule such as phenols, and a certain degree of selectivity has been shown [8–11]. Thus, it is of practical interest to understand how simple template structure such as phenols are able to conditionate the molecular recognition properties of the resulting imprinted polymers, and if the resulting selectivity could be controlled, increased or decreased by a careful choice of the pre-polymerization mixture.

In this work, we put our attention on molecular imprinting of the dangerous pollutant pentachlorophenol (PCP), a toxic substance largely diffused in many environments. In an effort to better understand the influence of the template structure on the selectivity of a PCP-imprinted polymer and obtain insights into the mechanisms governing the chromatographic separation mechanism, we used quantitative structure–retention relationship (QSRR) analysis to correlate the chromatographic retention behavior of several related phenols to structural molecular parameters determined by molecular mechanics or semi-empirical quantum chemical techniques. QSRR analysis was chosen because it is a useful technique capable of relating chromatographic retention behavior to the chemical structure of a solute. In addition, it can facilitate insight into the mechanisms governing chromatographic separation. In fact, by regression analysis, QSRR correlates retention parameters with structural properties of molecules, either determined from experiment or computed from molecular mechanics or semi-empirical quantum chemical techniques.

Structural molecular parameters useful for QSRR were selected and evaluated using principal component analysis (PCA) combined with principal component regression (PCR). The main advantage of this technique is to represent in an economic way the location of the m objects in a reduced coordinate system where instead of n variables, corresponding to the molecular descriptors, only p (with $p < n$) usually can be used to describe the original $m \times n$ dataset with maximum possible information. The new variables are called principal components and they are given by the linear combination of the n real variables, which coefficients are called “loadings”, while the new values corresponding to each principal component for every object are called “scores”. The relation between original dataset, loading and scores is defined by the matricial equation:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{e}$$

where $\boldsymbol{X}$ is the $m \times n$ matrix representing the dataset, $\boldsymbol{T}$ the $m \times p$ matrix of the scores, $\boldsymbol{P}'$ the $n \times p$ transposed matrix of the loadings and $\boldsymbol{e}$ is the $m \times n$ matrix of residuals.

Bi- or three-dimensional loadings plots give an indication of the relative importance of the corresponding variables in the principal components considered, while scores plots are very useful as a display tool for examining the relationships between objects, looking for trends, grouping or outliers. Moreover, PCA/PCR is preferable in comparison with other possible approaches, such as multiple linear regression (MLR), because it works well with the experimental variables without to be affected by the presence of non-orthogonal, i.e. correlated, molecular descriptors [12,13].

2. Experimental

2.1. Materials

2,3,4,5,6-Pentachlorophenol (PCP), all others phenol considered in this study, ethylene dimethacrylate and 4-vinylpiridine were from Sigma–Aldrich–Fluka (Milano, Italy), all others reagents and organic solvents were supplied by Merck (Darmstadt, Germany).

4-Vinylpyridine and ethylene glycol dimethacrylate were distilled at reduced pressure immediately before use. Phenol stock solutions were prepared by dissolving 20 mg of substance in 20 ml of acetonitrile and stored in the dark at −20 °C.

The HPLC apparatus (pump L-6200, UV-Vis detector L-4200 and integrator D-2500) came from Hitachi–Merck (Darmstadt, Germany).

2.2. Polymer preparation

In a 10 ml thick wall glass test tube a solution was prepared by dissolving 0.200 g (0.751 mmoles) of PCP into 4.0 ml of 3 + 1 ((v/v)) methanol–water.

Then, 0.234 ml (2.25 mmoles) of 4-vinylpiridine, 3.83 ml (20.28 mmoles) of ethylene glycol dimethacrylate and 0.040 g of 2,2′-azobis-(2-methylpropionitrile) were added. The mixture was purged with nitrogen and sonicated in a water-bath for 5 min. The vial was sealed, then the mixture was left to polymerize overnight at 60 °C. The polymer obtained was broken with a steel spatula, mechanically ground in a mortar and wet-sieved to 30–90 μm particle size. The particulate was extensively washed with 9 + 1 (v/v) methanol–acetic acid. No efforts were made to measure the amount of template molecule recovered. A non-imprinted polymer (NIP) was prepared and treated in the same manner, omitting PCP.

2.3. Column packing

An adequate amount of polymer was suspended in a 1 + 1 (v/v) methanol–water mixture and the slurry packed in a 100 mm stainless-steel HPLC column (i.d. 3.9 mm, geometrical volume 1.19 cm^3). The packing of the stationary phase was performed by gradually adding the slurry of the polymer to the column and eluting it with the mobile phase (1 + 1 (v/v) ethanol–water) at constant pressure of 10 MPa. The packed column was washed at 1 ml/min with 9 + 1 ((v/v)) ethanol–acetic acid until a stable baseline was reached (280 nm). After equilibration, the pressure in the column was of 2–5 MPa using organic solvents as a mobile phase and at a flow rate of 1 ml/min.

2.4. Liquid chromatography

Columns were equilibrated at a flow rate of 1 ml/min with 40 ml of acetonitrile–acetic acid 99 + 1 (v/v). Then, 20 μl of stock solution of PCP (or related substance) diluted 1 + 9 (v/v) with acetonitrile were injected and eluted at 1 ml/min, and the absorbance recorded at 280 nm. Each elution was repeated three times to assure the chromatogram reproducibility. Column void volume was measured by eluting 20 μl of acetone 0.05% (v/v) in acetonitrile.

The retention factor (k) was calculated as $(t - t_O)/t_O$, where t is the retention time of the eluted substance, and t_O the retention time corresponding to the column void volume. The selectivity index (SI) is defined as an index of polymer selectivity due to imprinting effects towards analogues of the template molecule. It was calculated as k_{NIP}/k_{MIP}.

2.5. Molecular descriptors

The geometries of 53 molecules (PCP and 52 related phenols) were subjected to molecular modeling with full geometry optimization (gradient were always less than 0.05 kcal $Å^{-1}$ mol^{-1}) using semi-empirical quantum-chemical method AM1 included in the HyperChem 5.01 package with the extension ChemPlus (HyperCube, Waterloo, Canada). On the basis of the optimized geometries, fourteen molecular descriptors related with electronic and steric properties of the molecules were calculated. These descriptors consist of the partial charge of the phenolic oxygen atom (qO), the partial charge of the phenolic hydrogen atom (qH), the absolute value of the difference of charge between the phenolic oxygen atom and the phenolic hydrogen atom (Δq), the highest occupied molecular orbital (HOMO), the lowest unoccupied molecular orbital (LUMO), the absolute value of the difference between the highest occupied molecular orbital and the lowest unoccupied molecular orbitals (Δorb), the square of total dipole moment (μ^2), the molecular weight (MW), the solvent-accessible molecular surface area (SAS), the hydrophobic solvent-accessible molecular surface area (hSAS), the van der Waals molecular surface area (Svdw), the hydrophobic part of the van der Waals molecular surface area (hSvdw), the molecular ovality (MOv, calculated from the van der Waals molecular surface area and volume, in according to [14]) and the radius of gyration (RG, calculated from the moments of inertia). In addition, calculated values for the logarithm of *n*-octanol–water partition coefficient (log P) and the phenolic dissociation constant (pK) were obtained using Clog P (BioByte Corp., Pomona, California, USA) and Sparc (http://www.ibmlc2.chem.uga.edu/sparc) programs. Calculated molecular descriptors are reported in Table 1.

2.6. Multivariate statistical analysis

Before to perform multivariate analysis, the initial dataset (selectivity factor and molecular descriptors) was transformed converting each single value in the Euclidean distance calculated from the corresponding value for pentachlorophenol. This type of scaling is justified by the needs to compare how phenols are differently recognized by the imprinted polymer respect to the template molecule. Thus, the Euclidean scaling can be considered a measure of similarity/dissimilarity, and it indicates how two object are different (far) each other.

The transformed variables in the dataset were autoscaled, mean-centered and subjected to multivariate analysis, that was performed in Mathcad environment (Mathcad 2000, MathSoft, Cambridge, MA, USA) using home-written PCA and PCR routines based on the generalized inversion matrix method [15].

Multivariate models were obtained by PCR utilizing iteratively a criterion involving the absolute correlation between respective eigenvectors and dependent variables [16,17]. The method used can be described as follows:

Step 1: All the principal components of the $m \times n$ dataset are calculated by using PCA.

Step 2: The principal component showing the highest adjusted regression coefficient (R_{adj}) with the selectivity factors is selected.

Step 3: From this best single-PC sub-set the best two-PC sub-set is identified as the one providing the highest R_{adj} with the selectivity factors from all two-PC possible combinations containing the best single-PC sub-set: the

Table 1
Selectivity index (SI) and molecular descriptors calculated for 2,3,4,5,6-pentachlorophenol (entry #0) and 52 related phenols (entries #1 to #52)

Phenols		SI	qO	qH	Δq	HOMO	LUMO	Δorb	μ^2	MW	SAS	hSAS	Svdw	hSvdw	MOv	RG	log *P*	p*K*
0	2,3,4,5,6-Pentachlorophenol	0.247	−0.205	0.209	0.414	−0.789	−9.136	8.347	1.219	270.7	352.7	315.7	193.7	175.3	7.1	5.305	5.12	4.50
1	Phenol	0.781	−0.228	0.196	0.424	−0.291	−9.175	8.884	1.304	95.9	255.8	206.6	124.3	104.5	11.4	3.333	1.46	9.99
2	2-Chlorophenol	0.752	−0.221	0.203	0.424	−0.023	−9.210	9.188	0.446	128.7	276.4	234.1	139.1	120.3	9.5	3.838	2.15	8.56
3	3-Chlorophenol	0.752	−0.224	0.198	0.422	−0.023	−9.276	9.253	3.771	128.6	280.7	230.6	139.2	119.3	10.5	3.988	2.50	9.11
4	4-Chlorophenol	0.758	−0.224	0.198	0.422	0.049	−9.010	9.058	1.943	130.8	280.4	230.5	138.7	118.8	10.6	3.814	2.39	9.35
5	2,3-Dichlorophenol	0.800	−0.222	0.200	0.422	−0.185	−9.038	8.853	1.362	165.8	299.5	250.3	154.8	135.2	9.6	4.338	2.84	7.70
6	2,4-Dichlorophenol	0.690	−0.218	0.204	0.422	−0.243	−9.093	8.850	0.271	165.8	299.5	257.3	153.5	134.8	8.7	4.454	3.06	7.89
7	2,5-Dichlorophenol	0.690	−0.217	0.204	0.421	−0.315	−9.210	8.895	1.182	165.8	300.9	258.8	153.5	134.9	8.6	4.365	3.06	7.51
8	2,6-Dichlorophenol	0.690	−0.210	0.206	0.416	−0.247	−9.193	8.946	1.780	163.0	299.9	262.9	153.3	134.9	8.5	4.333	2.75	6.79
9	3,4-Dichlorophenol	0.685	−0.221	0.200	0.421	−0.217	−9.055	8.838	4.248	165.8	300.6	250.6	153.0	133.1	9.8	4.335	3.33	8.58
10	3,5-Dichlorophenol	0.621	−0.221	0.201	0.422	−0.271	−9.526	9.255	1.772	165.8	305.1	255.1	154.3	134.4	9.8	4.677	3.62	8.18
11	2,3,4-Trichlorophenol	0.671	−0.216	0.205	0.421	−0.411	−9.115	8.704	0.391	200.8	316.4	274.3	167.0	148.4	8.1	4.599	3.77	6.87
12	2,3,5-Trichlorophenol	0.552	−0.211	0.204	0.415	−0.469	−9.115	8.646	2.528	197.5	323.4	278.8	168.2	148.7	8.9	4.884	3.84	6.62
13	2,3,6-Trichlorophenol	0.546	−0.208	0.207	0.415	−0.486	−9.307	8.820	2.421	197.5	319.3	282.3	166.6	148.2	7.9	4.732	3.77	5.72
14	2,4,5-Trichlorophenol	0.575	−0.215	0.205	0.420	−0.488	−9.093	8.605	1.336	197.5	319.7	277.1	166.9	148.1	8.3	4.857	3.72	6.86
15	2,4,6-Trichlorophenol	0.575	−0.207	0.207	0.414	−0.442	−9.125	8.684	1.004	200.8	322.6	285.5	167.8	149.4	7.9	5.023	3.69	6.23
16	3,4,5-Trichlorophenol	0.559	−0.219	0.201	0.420	−0.395	−9.114	8.719	2.836	165.8	321.7	271.7	167.8	147.9	9.2	5.225	4.01	7.84
17	2,3,4,5-Tetrachlorophenol	0.391	−0.210	0.204	0.414	−0.584	−9.051	8.467	3.122	235.7	340.2	298.1	181.6	162.1	8.4	5.001	4.21	6.05
18	2,3,4,6-Tetrachlorophenol	0.571	−0.208	0.206	0.414	−0.625	−9.077	8.452	0.466	235.7	339.3	302.5	181.3	163.1	7.2	5.131	4.45	5.15
19	2,3,5,6-Tetrachlorophenol	0.420	−0.206	0.208	0.414	−0.682	−9.193	8.511	0.968	235.7	338.9	301.9	180.9	162.5	7.5	5.096	3.88	4.91
20	2-Chloro-5-methylphenol	0.662	−0.216	0.199	0.415	0.032	−8.909	8.941	4.982	145.3	309.5	265.0	158.1	138.6	9.2	4.158	2.90	8.38
21	4-Chloro-2-methylphenol	0.662	−0.223	0.198	0.421	0.079	−8.925	9.004	1.774	145.3	307.4	265.1	158.1	139.2	8.6	4.237	2.63	9.71
22	4-Chloro-3-methylphenol	0.725	−0.225	0.198	0.423	0.086	−8.937	9.023	2.332	145.3	305.3	255.3	156.3	136.4	9.5	4.174	3.10	9.55
23	2-Methylphenol (*o*-cresol)	0.775	−0.228	0.197	0.425	0.291	−9.036	9.327	0.808	110.4	283.1	240.4	144.0	124.9	9.4	3.689	1.95	10.30
24	3-Methylphenol (*m*-cresol)	0.730	−0.228	0.196	0.424	0.278	−9.115	9.394	0.857	110.4	285.3	235.9	143.5	123.7	10.0	3.784	1.96	10.09
25	4-Methylphenol (*p*-cresol)	0.775	−0.228	0.196	0.424	0.323	−8.953	9.276	1.388	110.4	285.5	235.5	142.8	122.9	10.2	3.691	1.94	10.26
26	2,3-Dimethylphenol	0.685	−0.235	0.201	0.436	0.291	−8.985	9.276	2.713	124.9	303.2	263.3	158.1	140.1	7.6	4.005	2.48	10.54
27	2,4-Dimethylphenol	0.952	−0.228	0.196	0.424	0.330	−8.844	9.175	0.960	124.9	311.6	268.9	163.8	144.7	8.7	4.070	2.30	10.26
28	2,5-Dimethylphenol	0.685	−0.227	0.196	0.423	0.304	−8.910	9.214	1.357	139.4	312.5	270.1	161.9	143.1	8.4	3.839	2.33	10.41
29	2,6-Dimethylphenol	0.694	−0.232	0.202	0.434	0.293	−8.963	9.256	1.402	124.9	307.7	274.3	160.8	143.6	7.0	3.992	2.36	10.62
30	3,4-Dimethylphenol	0.781	−0.228	0.195	0.423	0.329	−8.881	9.209	2.190	124.9	308.3	258.2	161.0	141.1	9.3	4.075	2.23	10.43
31	3,5-Dimethylphenol	0.714	−0.229	0.196	0.425	0.298	−9.037	9.335	1.464	124.9	314.2	264.3	162.8	143.0	9.3	4.225	2.35	10.19
32	2,3,5-Trimethylphenol	0.671	−0.228	0.196	0.424	0.308	−8.877	9.185	0.922	139.4	334.9	293.2	179.8	161.0	7.7	4.385	2.87	10.59
33	2,4,6-Trimethylphenol	0.671	−0.232	0.202	0.434	0.328	−8.776	9.104	1.367	139.4	336.6	303.2	179.3	162.1	6.5	4.385	2.73	10.9
34	2,3,5,6-Tetramethylphenol	0.592	−0.236	0.202	0.438	0.300	−8.812	9.111	1.407	153.9	351.7	319.4	194.6	177.7	5.8	4.566	3.32	10.85
35	2-Ethylphenol	0.730	−0.234	0.202	0.436	0.275	−9.079	9.354	1.910	124.8	305.5	267.4	159.3	141.0	7.9	4.058	2.47	10.20
36	3-Ethylphenol	0.730	−0.229	0.196	0.425	0.293	−9.107	9.400	0.830	124.8	310.9	340.3	162.0	142.4	9.1	4.157	2.4	10.12
37	4-Ethylphenol	0.730	−0.228	0.196	0.424	0.326	−9.010	9.336	1.423	124.8	310.4	260.8	164.9	144.8	9.5	4.075	2.58	10.21
38	2-Allylphenol	0.862	−0.229	0.198	0.427	0.279	−9.036	9.315	1.263	134.2	332.0	333.2	176.3	157.6	7.8	4.543	2.5	10.29
39	4-Tertbutylphenol	0.752	−0.227	0.196	0.423	0.347	−8.997	9.344	1.421	153.6	350.6	301.5	196.8	177.1	7.9	4.574	3.31	10.29
40	2-Phenylphenol	0.595	−0.218	0.195	0.413	0.128	−9.091	9.219	1.508	173.7	371.5	331.1	207.4	188.3	7.0	4.870	3.09	10.54
41	4-Phenylphenol	0.581	−0.225	0.197	0.422	−0.337	−8.642	8.306	1.243	173.7	368.3	319.2	200.3	180.6	7.7	4.964	3.2	9.58
42	1-Naphtol	0.671	−0.227	0.199	0.426	−0.357	−8.544	8.187	0.815	144.2	321.6	278.7	171.8	153.0	7.9	4.349	2.85	9.34

Table 1 (*Continued*)

43	2-Naphtol	0.847	−0.227	0.196	0.423	−0.444	−8.654	8.211	1.874	146.9	324.0	274.6	172.6	152.8	8.8	4.337	2.7	9.57
44	2-Methoxyphenol	0.775	−0.224	0.207	0.431	0.312	−8.860	9.172	4.381	125.7	298.5	246.1	153.8	127.9	16.4	4.018	1.32	9.9
45	3-Methoxyphenol	0.741	−0.227	0.196	0.423	0.285	−8.944	9.229	3.791	125.7	299.7	238.9	155.0	128.9	16.8	4.125	1.34	10.22
46	4-Methoxyphenol	0.877	−0.227	0.195	0.422	0.229	−8.720	8.949	4.757	124.1	299.5	238.4	155.1	128.4	17.5	4.018	1.58	10.12
47	Catecol	0.826	−0.206	0.192	0.398	0.204	−8.827	9.031	2.135	111.6	268.6	169.2	134.2	95.6	41.4	3.642	0.88	9.49
48	4-Chlorocatecol	0.752	−0.241	0.209	0.450	−0.036	−8.926	8.890	5.954	146.6	291.1	200.3	148.5	110.0	38.2	4.205	1.66	8.67
49	4-Methylcatecol	0.800	−0.244	0.207	0.451	0.257	−8.791	9.047	3.565	126.1	295.4	204.7	153.4	114.8	36.9	4.026	1.37	9.62
50	1,3-Dihydroxybenzene	0.990	−0.226	0.198	0.424	0.321	−9.129	9.450	4.567	111.6	268.6	169.2	132.5	92.8	43.4	3.739	0.8	9.15
51	1,4-Dihydroxybenzene	0.971	−0.227	0.195	0.422	0.174	−8.762	8.936	4.924	111.6	268.1	168.9	132.9	93.2	43.8	3.652	0.59	10.13
52	1,3,5-Trihydroxybenzene	0.990	−0.228	0.201	0.429	0.316	−9.194	9.510	0.000	127.3	279.9	131.9	143.2	83.6	94.6	4.153	0.16	8.45
	r, coefficient of correlation with SI		0.555	0.526	0.305	0.665	0.258	0.540	0.199	0.817	0.568	0.670	0.602	0.671	0.531	0.753	0.851	0.611
	s, standard error of estimate		0.138	0.141	0.158	0.124	0.160	0.139	0.162	0.095	0.136	0.123	0.132	0.123	0.140	0.109	0.087	0.131

qO, partial charge of the phenolic oxygen atom; qH, partial charge of the phenolic hydrogen atom; Δq, absolute value of the difference between qH and qO; HOMO, highest occupied molecular orbital; LUMO, lowest unoccupied molecular orbitals; Δorb, absolute value of the difference between HOMO and LUMO; μ^2, square of total dipole moment; MW, molecular weight; SAS, solvent-accessible molecular surface area; hSAS, hydrophobic solvent-accessible molecular surface area; Svdw, van der Waals molecular surface area; hSvdw, hydrophobic part of the van der Waals molecular surface area; MOv, molecularovality; RG, radius of gyration; log*P*, logarithm of *n*-octanol–water partition coefficient; p*K*, dissociation constant.

process continues till all the principal components that increase the model's R_{adj} are included.

Step 4: The partial regression coefficient are calculated for all the variables of the model, and a new refined $m \times (n-1)$ dataset is obtained excluding the variable with the lowest partial regression coefficient.

Step 5: The procedure is iterated to step 1 until the R_{adj} of the reduced model falls down the value of the R_{adj} for the full-variables model.

It should be noted that this approach was used because of the more popular technique of principal component selection based on the criterium of eigenvectors decreasing magnitude [18] does not assure that eigenvectors highly correlating with the dependent variables will be selected. In fact, this approach could use eigenvectors irrelevant to the prediction property of the model, but which best explain the dataset variability. Thus, this kind of regression may yield non-optimal solutions.

The leverage of the objects used to calculate the PCR model was obtained considering the diagonal elements *h* of the leverage matrix $\boldsymbol{H}$ (hat matrix) obtained by the following relation:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

where $\boldsymbol{X}$ is the $m \times n$ matrix representing the dataset.

3. Results and discussion

The phenols chosen to study the selectivity of the PCP-imprinted polymer were selected considering the nature of their substituents. Apart from all the possible chlorinated phenols, several simple alkylated phenols and a restricted group of aryl-, methoxy- and polyphenols were considered. Due to the difficulties to describe accurately the ionization of carboxyl and amino groups with the available semi-empirical quantum-chemical methods, these kind of molecules were excluded. The same was made for a small number of nitro-substituted phenols, which showed anomalous high retention times measured on the NIP column ($k_{NIP} > 3$). This kind of selection was made in an effort to take into the account a homogeneous set of data, in which big changes in molecular structures will be limited. In fact, it is unrealistic to describe correctly the multivariate behavior of a set of data containing many different classes of molecules (characterized by very different values for the corresponding molecular descriptors) but with a limited number of objects ($m = 53$).

From Fig. 1, it is clear that the PCP-imprinted column shows a significative level of selectivity as direct effect of the imprinting process. In fact, the point representing the recognition of the template molecule PCP is quite isolated and located in an area of the plot characterized by high values for k_{MIP}. On the contrary, the main part of the phenols crowd in a well defined part of the plot, corresponding to

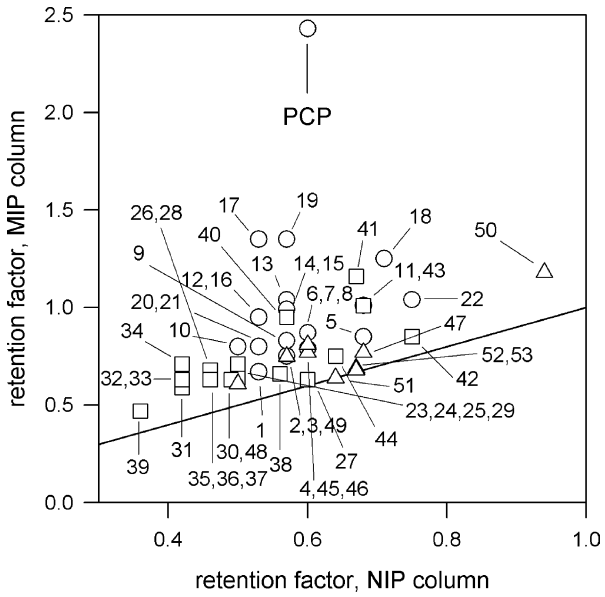

Fig. 1. Retention factors of phenols on the MIP column vs. NIP column. Open circles: chlorophenols; open squares: alkylphenols; open triangles: other phenols. The solid line indicates the area corresponding to the absence of imprinting effect ($k_{MIP} = k_{NIP}$) for the interaction between the stationary phase and a phenol.

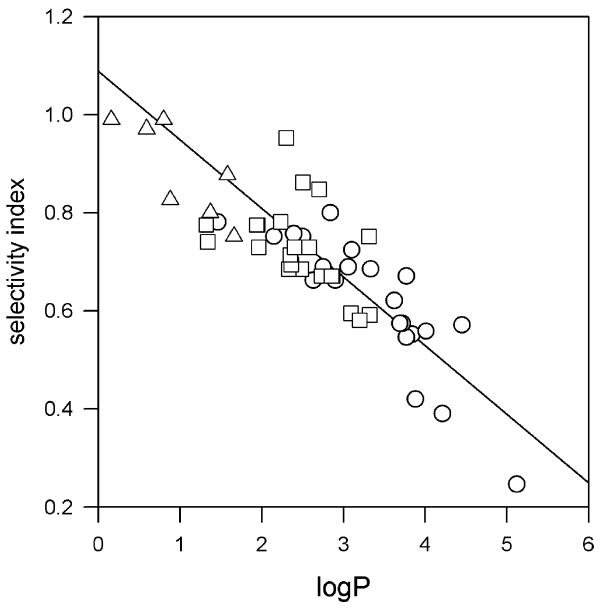

Fig. 2. Plot of selectivity factor vs. $\log P$. The parameter calculated for the linear regression equation $SI = -0.118 \log P - 1.014$ were: coefficient of correlation (R^2) = 0.851, residual sum of squares (SS_r) = 0.279, standard error of estimate (s^2) = 0.073. Open circles: chlorophenols; open squares: alkylphenols; open triangles: other phenols.

low values for k_{MIP} and k_{NIP}. It should be noted anyway that all the phenols, except few, show values of k_{MIP} greater than k_{NIP}, i.e. that, for the considered system, the imprinting process has a direct influence on its molecular recognition properties. Chlorinated phenols are better recognized than other phenols, and this fact could be interpreted in terms of hydrophobic, steric or electronic (acidity) effects. Anyway, some other phenols, characterized by bulky or multiple substituents (for instance: 2,3,5,6-tetramethylphenol, 2-phenylphenol, 4-phenylphenol) are well recognized anyway. Of consequence, apart the apparent preference for the chlorinated molecules, it is difficult to see any clear relation between the nature of the phenols and the selectivity of the imprinted polymer.

The situation does not change considering relations between selectivity index and molecular descriptors taken one by one. When these univariate models are examined, selectivity index does not show marked relationships with the most part of molecular descriptors. In fact, except for MW ($r = 0.817$) and $\log P$ ($r = 0.851$), reported in Fig. 2, the other descriptors show poor correlation coefficients with the selectivity factor. It should be stressed that MW and $\log P$ are well correlated each other ($r = 0.921$) in the set of data considered, and that for halogenated or alkylated phenols hydrophobicity expressed as $\log P$ can be simply calculated from steric descriptors such as molecular mass, solvent-accessible surface or van der Waals surface with an good level of precision [19,20]. Thus, the univariate analysis indicates only a possible hydrophobic effect and it does not let to obtain better informations on the nature of the molecular recognition than a visual examination of the k_{MIP}/k_{NIP} plot.

3.1. Principal component analysis

As anticipated in the introduction, the dataset corresponding to the molecular descriptors was examined using the principal component analysis technique essentially to obtain insights of the relative importance of the variables. It should be considered that PCA does not consider the existence of dependent variables (the selectivity index), but operate on the independent variables (the molecular descriptors) only.

For the principles on which PCA is based, a part of the calculated variance should be attributed to noise, while a reduced sub-set of principal components explains fully the calculated variance due to the descriptors variability. In a multivariate problem to know how many are the significative principal components is of paramount importance when it is necessary to discriminate what are the most useful descriptors. This can be approached using many different methods, calculating them directly from the eigenvalue distribution, or using cross-validation techniques [21,22]. A good estimate of the number of significative principal components can be easily calculated using the so-called "*K*-correlation index" [23]:

$$K := \frac{\sum_{i=1}^{n} \left| \left(V_i / \sum_{i=1}^{n} V_i \right) - (1/n) \right|}{(2/n)(n-1)} \quad \text{with} \quad 0 < K < 1$$

where n is the number of descriptors, and V the calculated eigenvalues. The value of K, that is an estimate of how much the descriptors are correlated each other, it is possible to calculate KL, the maximum number of potentially

Table 2
Principal component analysis of the molecular descriptor dataset: eigenvalues greater than 0.001 and corresponding cumulative % of explained variance

	Eigenvalues										
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
Explained variance (eigenvalue size)	8.236	3.119	1.558	1.087	0.732	0.600	0.249	0.134	0.100	0.075	0.043
Cumulative % explained variance	51.4	71.0	80.7	87.5	92.1	95.8	97.4	98.2	98.8	99.3	99.6

significant principal components, and KP, the minimum number of potentially significant principal components, as

$$\mathrm{KP} = \mathrm{round}[1 + (n-1)(1-K)] \quad \text{and}$$
$$\mathrm{KL} = \mathrm{round}[n^{(1-K)}]$$

where round is the greater of the nearest integer.

For the complete dataset of calculated molecular descriptors the value for the *K*-correlation index is 0.667, corresponding to a medium degree of correlation between descriptors. The explained variance of the dataset is 95.8 and 80.7% when are considered the upper (KP = 6) and the lower (KL = 3) limits (see Table 2). It should be noted that the upper limit converge with the maximum significant eigenvalue measured using the popular "scree plot" (see Fig. 3). A PCA model involving only three principal components has the advantage to be easily represented in a graphical form utilizing three bi-dimensional scores and loadings plots, without loss of information because more than 4/5 of the molecular descriptors variability is retained.

Scores plot is reported in Fig. 4. Considering models ranging from 1 to 6 principal components, for three objects (numbered #48, 4-chlorocatechol, #49, 4-methylcatechol and #52, 1,3,5-trihydroxybenzene). the statistical analysis of the matrix of the residuals show that the residual variance calculated for each object is greater than the residual variance calculated for the entire dataset. Thus, these object can be considered potential outlayers and prudently excluded in the successive PCR analysis. It should be also noted that objects are grouped in some sub-sets, corresponding to chlorinated, alkyl/arylphenols and metoxy/polyphenols, and that the chlorophenols subclass is well described by the first principal component alone (i.e. it has a positive and constant value for the second and third principal component), while alkyl/aryl- and polyphenols are described collectively by all the three principal components. Thus, a substantial difference between chlorinated and non-chlorinated phenols can be seen, even if PCA alone do not let to see if such a difference could influence correlations between principal components (i.e. molecular descriptors) and selectivity factors.

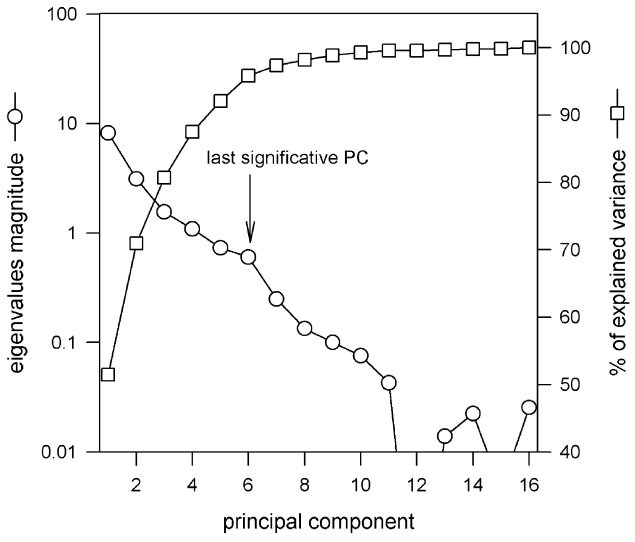

Fig. 3. Principal component analysis of the molecular descriptor dataset: scree plot for eigenvalues greater than 0.001 and corresponding cumulative % of explained variance.

As regards the loadings (see Fig. 5), no molecular descriptor shows very high (>0.6) or very low (<0.1) values for the three first principal components taken together. Thus, there are not any descriptors able to influence strongly the model, while it is plausible that all these descriptors—or a large part of them—are useful to account of the whole dataset variability.

It is remarkable that in PC1 the main contribute is given from steric molecular descriptors, such as MW, GR and log *P*, while in PC2 and PC3 the main contribute is given from electronic molecular descriptors, such as qO, qH, Δq, LUMO, μ^2 e p*K*. Considering that PC1 accounts for more than 50% of the dataset variability, this indicates that steric descriptors play a more significant role to describe the dataset properties than electronic molecular descriptors.

The close proximity in which SAS and Svdw are in the loadings plot indicates that these descriptors show a marked redundance, explaining the same dataset variability. Thus, one of them could be deleted without affecting multivariate model.

Considering models ranging from 1 to 6 principal components, the statistical analysis of the residuals show that a model including the principal components PC1–PC3 the explained variance calculated for each variable (Fig. 6) is greater for the steric than electronic molecular descriptors (except for MOv), while to include variance depending from electronic molecular descriptors it is necessary to expand the model to five principal components, and variance explained by μ^2 is included in a model with six principal components only. Since PC1–PC3 principal components explain 4/5 of the molecular descriptors variability, this confirms that steric descriptors play a more significant role to describe the dataset properties. Anyway, the PCA cannot show if these

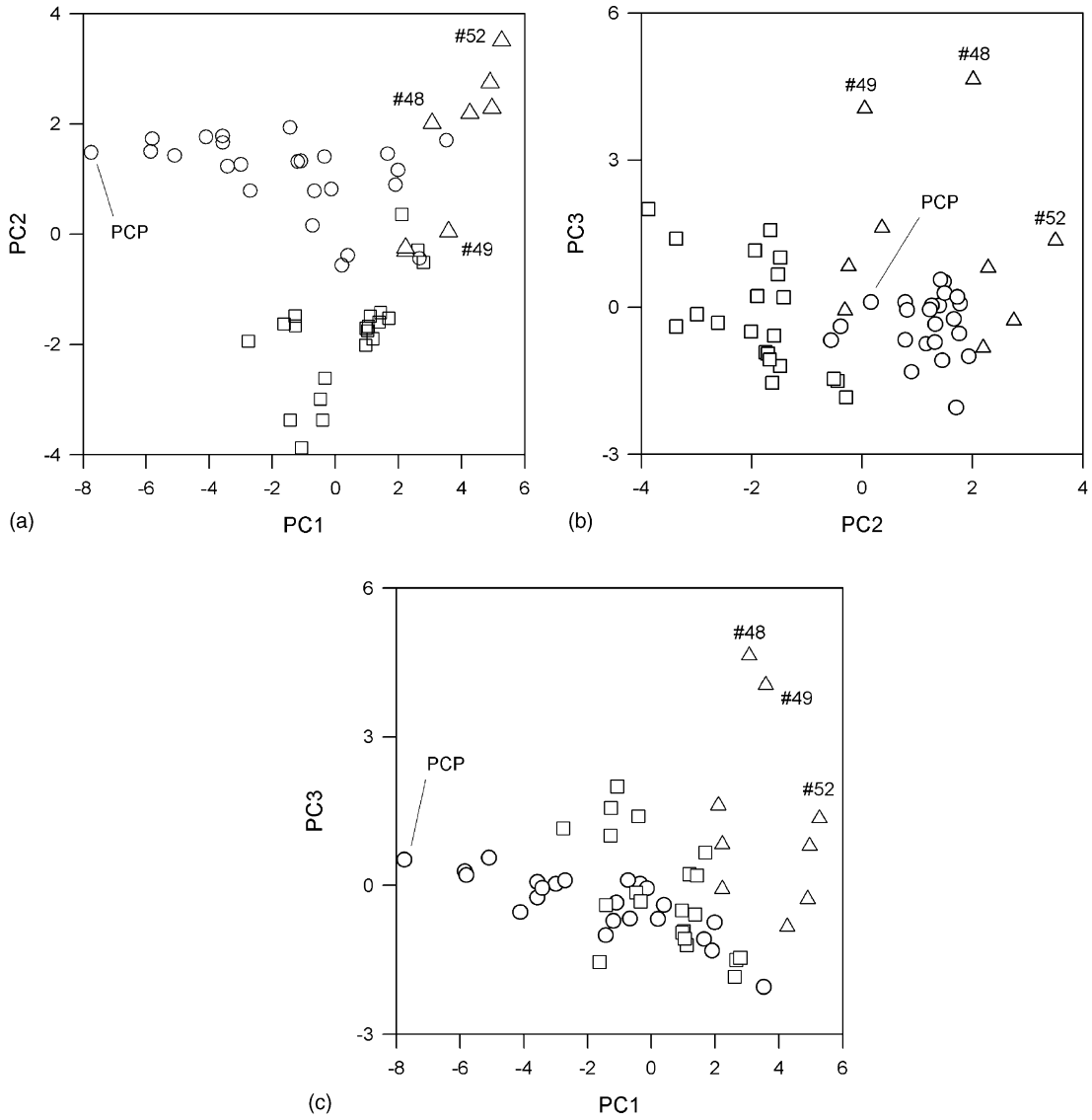

Fig. 4. (a–c) Principal component analysis of the molecular descriptor dataset: scores plots. Open circles: chlorophenols; open squares: alkyl/arylphenols; open triangles: other phenols. #48, #49 and #52 are outlayers.

descriptors better correlate collectively with selectivity factors. To do this it is necessary to consider the regression made on principal components.

3.2. Principal component regression

A first-level principal component regression model was calculated including all the molecular descriptors except for SAS (highly correlated with Svdw through the covariance matrix) and excluding the outlayering objects #48 (4-chlorocatechol) #49 (4-methylcatechol) and #52 (1,3,5-trihydroxybenzene).

Then, refined principal component regression models with decreasing dimensionality (i.e. decreasing number of molecular descriptors) were calculated till obtaining a model with the same adjusted regression coefficient of the first-level model, but with only seven residual variables (Fig. 7). This model was named "minimum dimensionality model" (MDM):

$$\mathrm{SI} = 0.243\,\mathrm{qO} - 0.190\,\Delta q + 0.946\,\mathrm{MW} - 0.285\,\mathrm{hSvdw} + 0.255\,\mathrm{MOv} + 0.343\log P - 0.452\,\mathrm{p}K;$$

$$m = 50,\ n = 7,\ p = 6,\ R_{\mathrm{adj}} = 0.756,\ s^2 = 0.255$$

The influence of single objects on the model is reported in the leverage plot (Fig. 8). No objects can be considered strong outlayers ($|\mathrm{SI}_{\mathrm{calc}} - \mathrm{SI}_{\mathrm{obs}}/s(1-h)^{0.5}| > 3$), while only four objects, numbered #40 (2-phenylphenol), #47 (catecol), #50 (1,3-dihydroxybenzene) and #51 (1,4-dihydroxybenzene), show an high leverage ($h > 3h_{\mathrm{mean}}$) and can be considered to have a strong influence on the model. Thus, MDM represents quite well the objects considered in this work. The

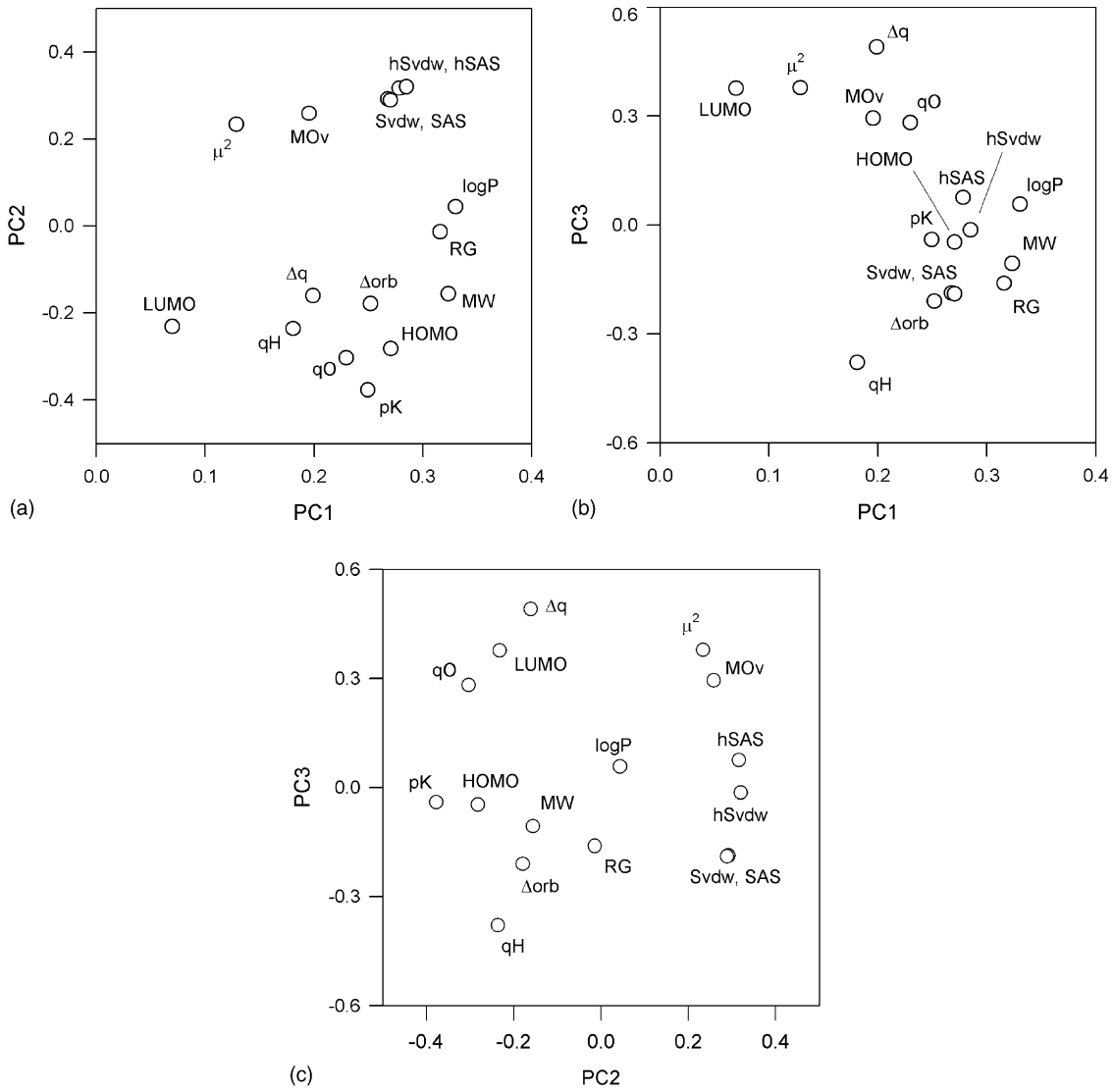

Fig. 5. (a–c) Principal component analysis of the molecular descriptor dataset: loadings plots.

same it can be seen by the SI_{calc} versus SI_{obs} plot (Fig. 9). It should be noted that the distribution of the calculated residuals is homoscedastic, thus the variance of the model is distributed in a quite homogeneous form along all the objects.

The most significative feature of the MDM is the high numerical value for the MW parameter, twice bigger than the second in magnitude, p*K*. Since all the variables were mean-centered and autoscaled before of the multivariate analysis, the high value of the MW parameter indicated that steric factors are decisive to conditionate the molecular recognition, an extent significantly greater than electronic factors.

Related to the first, it is another significative feature: the loss of the majority of electronic molecular descriptor. In fact, the partial charge at phenolic hydrogen (qH), orbital-related (HOMO, LUMO, Δorb) and dipole-related (μ^2) descriptors are let fall, while are preserved the main part of the structural molecular descriptors, such as MW, hSvdw and MOv. Anyway, molecular descriptors referring to hydrophobicity (log *P*) and acidity (p*K*) are preserved. It makes sense because—as previously underlined—for substituted phenols MW and log *P* are strictly related, while it has been shown that partial charge at phenolic oxygen and the p*K* value are linearly correlated when partial charge is calculated using AM1 or PM3 semi-empirical quantum-mechanical methods [24].

As regards the algebraic sign of the MDM parameters, it is interesting that a variable related to the molecular shape (MOv) reinforces the effect of the biggest parameter in the model, MW, while hSvdw—that take into the account the hydrophobic fraction of the molecular surface—weakens this effect. Thus, dimension and shape differences are decisive to determine its recognition by the polymer, but change in the hydrophobic part of the molecule can modify this. As regards the effect of the algebraic sign on the electronic

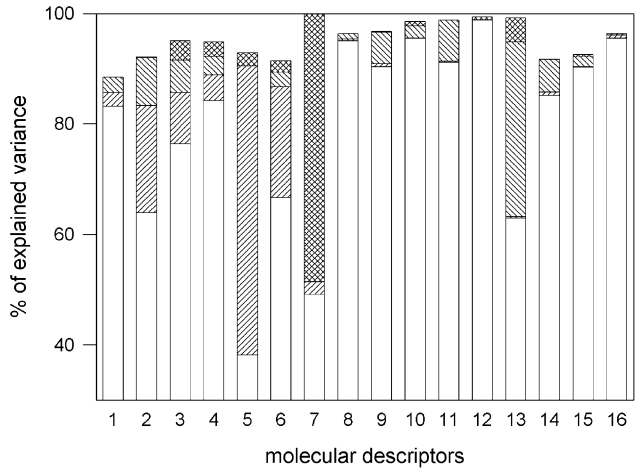

Fig. 6. Principal component analysis of the molecular descriptor dataset: % of variance explained for each variables. (1) qO; (2) qH; (3) Δq; (4) HOMO; (5) LUMO; (6) Δorb; (7) μ^2; (8) MW; (9) SAS; (10) hSAS; (11) Svdw; (12) hSvdw; (13) Mov; (14) RG; (15) log *P*; (16) p*K*. Open bars: PC1–PC3 model. Forward diagonals bars: PC1-PC4 model. Backward diagonals bars: PC1–PC5 model. Diagonal crossed bars: PC1–PC6 model.

parameters, it is remarkable that the positive sign for qO is almost counterbalanced be the negative sign for Δq. So, the main significative electronic parameter became p*K*, which has negative algebraic sign. Thus, its influence is opposed to that of steric factors.

Considering the whole model in terms of polymer selectivity index, it can be seen that steric and electronic molecular descriptors have opposite effects on the molecular recognition of PCP-analogs: steric descriptors decrease the recognition by enhancing the selectivity index, while electronic descriptors increase the recognition—even if in a lesser extent—by reducing the selectivity index.

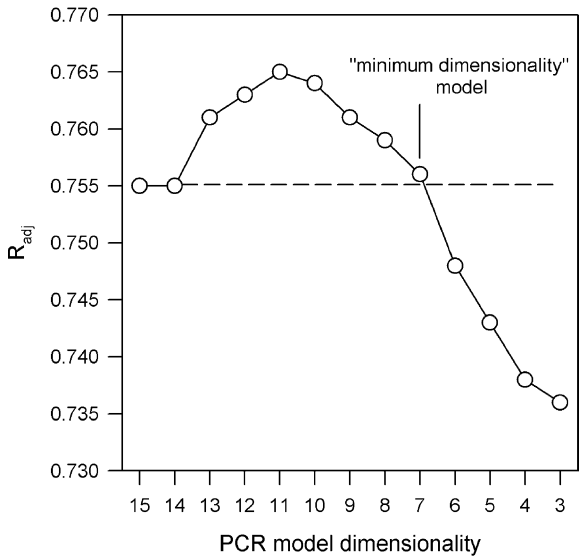

Fig. 7. Principal component regression of the molecular descriptor dataset: dimensionality of regression models. Dashed line indicates the value corresponding to the adjusted regression coefficient for the 15-variables model.

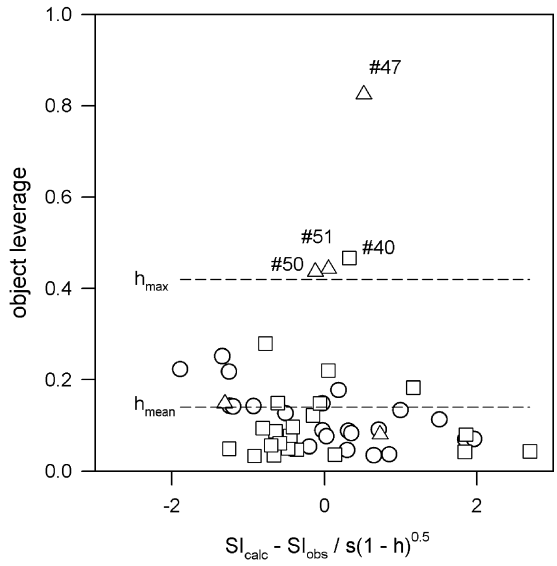

Fig. 8. Principal component regression of the reduced molecular descriptor dataset: leverage plot. Open circles: chlorophenols; open squares: alkylphenols; open triangles: other phenols.

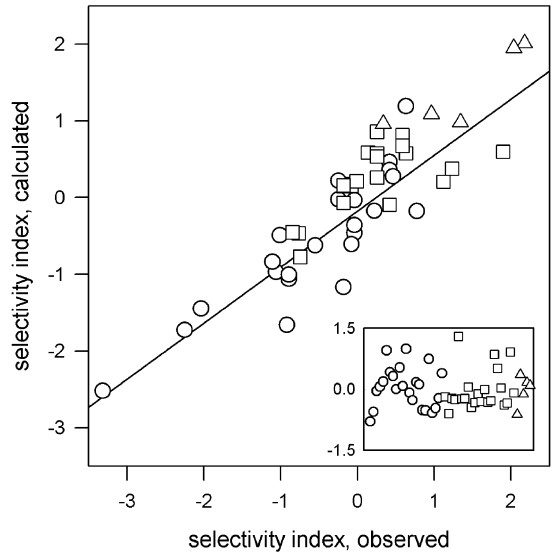

Fig. 9. Principal component regression of the reduced molecular descriptor dataset: calculated selectivity index vs. observed selectivity index (mean-centered and autoscaled values). In the inset: differences between observed and calculated selectivity index. Open circles: chlorophenols; open squares: alkylphenols; open triangles: other phenols.

4. Conclusions

Correlate column selectivity with molecular descriptors is difficult using the traditional statistical tools, because the most part of these descriptors correlate each other, and this make the use of multiple linear regression improper, if not impraticable. The application of chemometric methods of multivariate analysis such as principal component analysis and principal component regression make it possible. The results obtained in this work show that a polymer obtained

by molecular imprinting with pentachlorophenol show a pattern of selectivity towards several other related phenols that can be rationalized in terms of steric and electronic molecular descriptors.

References

[1] I.A. Nicholls, K. Adbo, H.S. Andersson, P.O. Andersson, J. Ankarloo, J. Hedin-Dahlström, P. Jokela, J.G. Karlsson, L. Olofsson, J. Rosengren, S. Shoravi, J. Svenson, S. Wikman, Anal. Chim. Acta 435 (2001) 9.
[2] M. Lepistö, B. Sellergren, J. Org. Chem. 54 (1989) 6010.
[3] K.G.I. Nilsson, K. Sakaguchi, P. Gemeiner, K. Mosbach, J. Chromatogr. A 707 (1995) 199.
[4] C. Yu, K. Mosbach, J. Org. Chem. 62 (1997) 4057.
[5] K. Yoshizako, K. Hosoya, Y. Iwakoshi, K. Kimata, N. Tanaka, Anal. Chem. 70 (1998) 386.
[6] C. Baggiani, G. Giraudi, C. Giovannoli, G. Giraudi, A. Vanni, J. Chromatogr. A 883 (2000) 119.
[7] D.A. Spivak, J. Campbell, Analyst 126 (2001) 793.
[8] N. Masqué, R.M. Marcé, F. Borrull, P.A.G. Cormack, D.C. Sherrington, Anal. Chem. 72 (2000) 4122.
[9] M. Janotta, R. Weiss, B. Mizaikoff, O. Brüggemann, L. Ye, K. Mosbach, Int. J. Environ. Anal. Chem. 80 (2001) 75.
[10] E. Caro, N. Masqué, R.M. Marcé, F. Borrull, P.A.G. Cormack, D.C. Sherrington, J. Chromatogr. A 963 (2002) 169.
[11] E. Caro, R.M. Marce, P.A.G. Cormack, D.C. Sherrington, F. Borrull, J. Chromatogr. A 995 (2003) 233.
[12] S. Wold, K. Esbensen, P. Geladi, Chemometr. Intell. Lab. Syst. 2 (1987) 37.
[13] T. Næs, B.H. Mevik, J. Chemometr. 15 (2001) 413.
[14] N. Bodor, M.J. Huang, J. Pharmaceut. Sci. 81 (1992) 272.
[15] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman. Chemometrics: A Textbook, Elsevier, Amsterdam, 1988 (Chapter 13).
[16] Y.L. Xie, J.H. Kalivas, Anal. Chim. Acta 348 (1997) 19.
[17] S.Z. Fairchild, J.H. Kalivas, J. Chemometr. 15 (2001) 615.
[18] J.M. Sutter, J.H. Kalivas, P.M. Lang, J. Chemometr. 6 (1992) 217.
[19] K. Iwase, K. Komatsu, S. Hirono, S. Nakagawa, I. Moriguchi, Chem. Pharm. Bull. 33 (1985) 2114.
[20] N. Bodor, P. Buchwald, J. Phys. Chem. B 101 (1997) 3404.
[21] S. Wold, Technometrics 20 (1978) 378.
[22] E.R. Malinowski, J. Chemometr. 3 (1988) 49.
[23] R. Todeschini, Anal. Chim. Acta 348 (1997) 419.
[24] T. Hanai, J. Chromatogr. A 985 (2003) 343.