

Modeling the Octanol–Water Partition Coefficient of Substituted Phenols by the Use of Structure Information

LORENTZ JÄNTSCHI,¹ SORANA-DANIELA BOLBOACĂ²

¹ Technical University, Cluj-Napoca, 15 Constantin Daicoviciu Street,
400020 Cluj-Napoca, Romania

² Iuliu Hatieganu University of Medicine and Pharmacy, 13 Emil Isac Street,
400023 Cluj-Napoca, Romania

Received 14 August 2006; accepted 29 November 2006

Published online 3 January 2007 in Wiley InterScience (www.interscience.wiley.com).

DOI 10.1002/qua.21292

ABSTRACT: This work presents the abilities in estimation and prediction of the octanol–water partition coefficient of some para-substituted phenols through the integration of complex structures information by the use of an original molecular descriptors family on the structure–property relationship approach. The proposed approach uses the complex information obtained from para-substituted phenols structure in order to generate and calculate the molecular descriptors family. The structure–property relationship models were built based on the generated descriptors. The obtained multi-varied models (model with two and four descriptors, respectively) were validated through the assessment of the cross-validation leave-one-out score. The comparison between the multi-varied model with two and four descriptors was performed using Steiger’s Z-test. The analysis of the statistical characteristics of the obtained models demonstrated that the model with four descriptors has greater ability to estimate and predict compared with the model with two descriptors. This observation was also sustained by the results of correlated-correlation analysis. The multi-varied model with four descriptors revealed that the octanol–water partition coefficient of studied para-substituted phenols is likely to be of geometry nature, it is strongly dependent on the partial charges of compounds and group electronegativity, and it is in relation to the elastic force. © 2007 Wiley Periodicals, Inc. *Int J Quantum Chem* 107: 1736–1744, 2007

Key words: molecular descriptors family on structure–property relationships (MDF-SPR); octanol–water partition coefficient; para-substituted phenols

Correspondence to: L. Jäntschi; e-mail: lori@academicdirect.org
Contract grant sponsor: UEFISCSU Romania.
Contract grant number: ET36/2005.

Introduction

The octanol–water partition coefficient, defined as the ratio of the concentration of a chemical in octanol and in water at equilibrium and at a specified temperature [1] is used by many researchers in quantitative structure–property relationship studies. Partition coefficients are used in medicinal chemistry [2], drug design [3], toxicology [4], and environmental chemistry [5].

The literature reported various methods that are able to predict the octanol–water partition coefficient [6] by applying the fragment constant methods [7], by computing van der Waals molecular volume and surface area through analytical and numerical techniques [8], by the use of fuzzy [9], and the neural network approach [10].

An original approach to molecular descriptor family on structure–property relationships (MDF-SPR), method that proved to be able to estimate and predict properties, has been developed [11]. Starting from the successful results obtained by the use of the MDF-SPR methodology on estimation and prediction of retention chromatography index [12], octanol/water partition coefficients [13, 14], water activated carbon adsorption [15], and molar refraction [16], the aim of the research was to study the abilities of the MDF-SPR methodology in estimation and prediction of octanol–water partition coefficient of some para-substituted phenols.

Materials and Methods

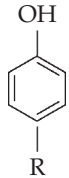
PARA-SUBSTITUTED PHENOLS

A number of 30 para-substituted phenols, studied previously by Schultz [17] were included in the study. The generic structure of compounds, their abbreviation (Abb.), the substituent from the para position (R), and associated octanol–water partition coefficient, expressed in logarithmic scale are presented in Table I.

MOLECULAR DESCRIPTORS FAMILY ON STRUCTURE–PROPERTY RELATIONSHIP METHODOLOGY

The octanol–water partition coefficient of para-substituted phenols was modeled by the use of the MDF-SPR methodology. The steps followed in the

TABLE I
Abbreviation, substituent, and associated octanol–water partition coefficient for para-phenols.

	Abb.	Substituent R	Log K_{ow}
	pph_01	CONH ₂	0.00
	pph_02	NHCOCH ₃	0.32
	pph_03	CH ₂ CH ₂ OH	0.72
	pph_04	CH ₂ CN	0.90
	pph_05	OCH ₃	1.34
	pph_06	CHO	1.35
	pph_07	COCH ₃	1.35
	pph_08	H	1.49
	pph_09	COC ₂ H ₅	1.55
	pph_10	CN	1.60
	pph_11	F	1.77
	pph_12	OC ₂ H ₅	1.87
	pph_13	NO ₂	1.91
	pph_14	CH ₃	1.94
	pph_15	Cl	2.39
	pph_16	C ₂ H ₅	2.58
	pph_17	Br	2.59
	pph_18	I	2.91
	pph_19	OC ₄ H ₉	3.04
	pph_20	CH(CH ₃) ₂	3.05
	pph_21	COC ₆ H ₅	3.07
	pph_22	C ₃ H ₇	3.18
	pph_23	N=NC ₆ H ₅	3.18
	pph_24	C ₆ H ₅	3.20
	pph_25	C(CH ₃) ₃	3.31
	pph_26	OC ₆ H ₅	3.56
	pph_27	CH ₂ CH(CH ₃) ₂	3.60
	pph_28	Cyclopentyl	3.63
	pph_29	CH ₂ C ₆ H ₅	3.69
	pph_30	CH ₂ C(CH ₃) ₃	4.03

modeling process, described in detail in Ref. [11] are:

1. *3D representation of compounds*: the three-dimensional representations of para-substituted phenols were built up by using *HyperChem* software [18].
2. *Creation of measured properties file*: the octanol–water partition coefficient for each para-substituted phenol, expressed in logarithmic scale was stored in *phenols.txt* file.
3. *Molecular descriptors family generation and computing*: All 30 compounds were used in the construction and generation of the molecular descriptors family. The algorithm generates the list of molecular descriptors family and associated values of para-substituted phenols, strictly based

on complex information obtained from the compound structure. To discard redundant information, a bias method with a significance level of 10^{-9} was applied after generation of the molecular descriptors family. Each calculated descriptor has an individual seven-letter name that expresses the modality of construction:

- a. Compound characteristic relative to its geometry (*g*) or topology (*t*)—the 7th letter
- b. Atomic property: cardinality (*C*), number of directly bonded hydrogen's (*H*), atomic relative mass (*M*), atomic electronegativity (*E*), group electronegativity (*G*), and the partial charge, semi-empirical extended Hückel model, single-point approach (*Q*)—the 6th letter
- c. Atomic interaction descriptor—the 5th letter
- d. Overlapping interaction model—the 4th letter
- e. Fragmentation criterion: the minimal fragments (*m*), the maximal fragments (*M*), the Szeged fragments criterion (*D*), and the Cluj fragments criterion (*P*) [19, 20]—the 3rd letter
- f. Cumulative method of fragmentation properties (nineteen functions)—the 2nd letter:
 - i. Conditional group (four functions): smallest fragmental descriptor value from the array (*m*), highest value (*M*), smallest absolute value (*n*), and highest absolute value (*N*)
 - ii. Average group (five functions): sum of descriptor values (*S*), average mean for valid fragments (*A*), average mean for all fragments (*a*), average mean by atom (*B*), average mean by bond (*b*)
 - iii. Geometric group (five functions): multiplication of descriptor values (*P*), geometric mean for valid fragments (*G*), geometric mean for all fragments (*g*), geometric mean by atom (*F*), and geometric mean by bond (*f*)
 - iv. Harmonic group (five functions): harmonic sum of values (*s*), harmonic mean for valid fragments (*H*), harmonic mean for all fragments (*h*), harmonic mean by atom (*l*), and harmonic mean by bond (*i*)
- g. Linearization procedure applied in global molecular descriptor generation: identity (*l*), inverse (*i*), absolute (*A*), an inverse of absolute (*a*), natural logarithm of absolute value (*L*), and simple natural logarithm (*l*)—1st letter.

4. *Identification of best performing MDF-SPR models:* The criteria imposed in searching for the best-performing models were: the model significance, the values for the correlation and squared correlation coefficients (they were considered performing models if the correlation and/or squared correlation coefficients were closed to +1 to -1), the standard error and the significance of the coefficients.
5. *Validation of the MDF-SPR models:* The analysis of the predictive abilities of the MDF-SPR models was performed through model validation analysis by computing: the cross-validation leave-one-out (loo) score, the Fisher parameter and its significance for leave-one-out analysis, and the standard error for leave-one-out analysis. In leave-one-out analysis, the property of each compound was predicted by the regression equation calculated based on all the other compounds by using the leave-one-out analysis application [21].
6. *Analysis of the MDF-SPR models:* The chosen MDF-SPR models were analyzed through computing and interpreting of a number of seven statistical characteristics of the models. Comparison between the multi-varied model with four descriptors and the model with two descriptors was performed through a correlated correlation analysis using Steiger's test [22] at a significance level of 5%. The estimation ability of the model with the highest squared correlation coefficient was analyzed in training and test sets using the training vs test application [23]. Nine situations were analyzed, starting with sample sizes in training sets from 15 to 30 and corresponding sample sizes in test sets from 15 to 7.

Results

Two multi-varied MDF-SPR models with two and four descriptors, respectively, proved to have abilities in estimation and prediction of the octanol-water partition coefficient for studied para-substituted phenols. The MDF-SPR models were:

1. The MDF-SPR model with two descriptors:

$$\hat{Y}_{2D} = 1.07 + 3.38 \cdot 10^{-3} \cdot isDDkGg - 0.40 \cdot IMmrKQg. \quad (1)$$

2. The MDF-SPR model with four descriptors:

TABLE II

Descriptors of MDF-SPR models, their values, and estimated octanol-water partition coefficients by the model with two (\hat{Y}_{2D}), and four (\hat{Y}_{4D}) descriptors, respectively.

Abb.	Two descriptors		Four descriptors		\hat{Y}_{2D}	\hat{Y}_{4D}
	isDDkGg	IMmrKQg	IPMDKQg	IFMMKQg		
pph_01	700.18	6.3359	-207.27	-10.13	0.8964	0.1807
pph_02	833.32	7.4800	-120.71	2.40	0.8880	0.2822
pph_03	664.20	5.6843	-147.47	-6.77	1.0360	0.5529
pph_04	644.55	5.9387	-90.69	-10.64	0.8674	1.1734
pph_05	527.06	3.7241	-57.18	-5.86	1.3584	1.3856
pph_06	597.00	5.7597	-100.68	-13.71	0.7783	1.1295
pph_07	693.19	5.8970	-133.89	-13.79	1.0488	1.3002
pph_08	317.52	0.5873	-84.64	-11.54	1.9078	1.7111
pph_09	761.22	4.6825	-168.54	-11.15	1.7666	1.6524
pph_10	763.19	5.7093	-115.17	-8.64	1.3612	1.5422
pph_11	459.29	2.7259	-61.45	-7.48	1.5295	1.5107
pph_12	657.84	3.7713	-72.80	-6.61	1.7823	2.0041
pph_13	776.25	5.6198	-28.18	-2.48	1.4413	1.9908
pph_14	382.22	0.5686	-103.37	-12.09	2.1344	1.9448
pph_15	541.18	1.5883	-68.53	-7.18	2.2634	2.3491
pph_16	505.83	0.5686	-142.29	-14.59	2.5529	2.4615
pph_17	487.59	0.5495	-71.23	-8.47	2.4988	2.5588
pph_18	557.79	0.5593	-85.06	-10.46	2.7326	2.9703
pph_19	908.30	3.7837	-147.14	-11.15	2.6253	3.0464
pph_20	647.10	0.5647	-164.12	-15.01	3.0328	3.0759
pph_21	1280.30	6.0440	-334.61	-18.01	2.9777	2.9444
pph_22	637.23	0.5767	-180.75	-16.55	2.9946	2.9797
pph_23	1053.10	2.1567	-429.28	-29.15	3.7686	3.2778
pph_24	893.42	0.5804	-300.93	-20.12	3.8606	3.5500
pph_25	769.84	0.5609	-182.80	-14.95	3.4499	3.5791
pph_26	1055.90	3.8434	-3.08	8.62	3.1011	3.6558
pph_27	763.21	0.7108	-204.49	-16.75	3.3673	3.4163
pph_28	797.66	0.5631	-258.12	-19.81	3.5433	3.4029
pph_29	1026.80	2.0231	-179.73	-3.15	3.7331	3.5077
pph_30	900.22	0.7438	-229.89	-17.48	3.8180	3.9817

\hat{Y}_{2D} = estimated log K_{ow} by the model with two variables, \hat{Y}_{4D} = estimated log K_{ow} by the model with four variables.

$$\begin{aligned} \hat{Y}_{4D} = & 8.69 \cdot 10^{-2} + 5.56 \cdot 10^{-3} \cdot isDDkGg \\ & - 4.16 \cdot 10^{-1} \cdot IMmrKQg \\ & + 9.41 \cdot 10^{-3} \cdot IPMDKQg \\ & - 7.80 \cdot 10^{-2} \cdot IFMMKQg. \quad (2) \end{aligned}$$

The molecular descriptors used by the models, their calculated values, the estimated value of the octanol-water partition coefficient obtained with each model (\hat{Y}_{2D} , estimated octanol-water partition coefficient by the model with two descriptors; \hat{Y}_{4D} , estimated octanol-water partition coefficient by the model with four descriptors), and the values of

residuals (defined as differences between measured octanol-water partition coefficient and estimated by the multi-varied model with two variables: $R_{\hat{Y}_{2D}}$ and by the multi-varied model with four variables $R_{\hat{Y}_{4D}}$, respectively) are presented in Table II.

Graphical representations of residuals obtained with the MDF-SPR models with two and four descriptors, respectively, are shown in Figure 1. The statistical characteristics of the MDF-SPR models are presented in Table III, and the quality characteristics of the regression models are shown in Table IV. The plot of the estimated log K_{ow} by multi-varied MDF-SPR model with

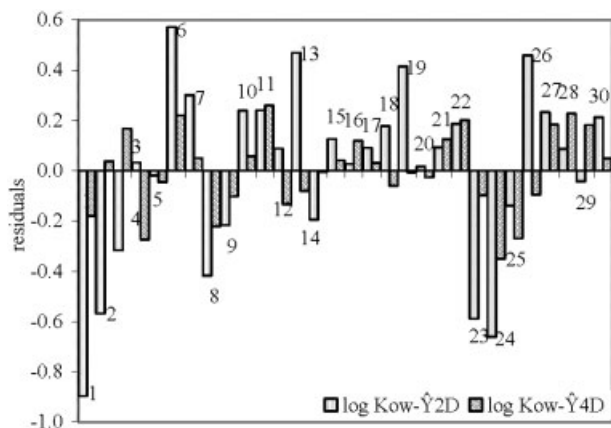


FIGURE 1. Plot of residuals obtained by MDF-SPR models with two and four descriptors, respectively.

four descriptors versus measured $\log K_{ow}$ is presented in Figure 2.

A correlated correlation analysis was applied to verify the hypothesis that the correlation coefficient obtained by the model with four descriptors was not statistically different, at a significance level of 5%, compared with the correlation coefficient ob-

tained by the model with two descriptors. The results are presented in Table V.

Validation of the multi-varied MDF-SPR model with four descriptors was performed by spilling the sample of para-substituted phenols in training and test sets. The characteristics of the regression models and their performances are shown in Table 6. The following are included in Table VI: the coefficients of the regression models (using the generic model: $\hat{Y} = a_0 + a_1 \cdot isDDkGg + a_2 \cdot IMmrKQg + a_3 \cdot IPMDKQg + a_4 \cdot IFMMKQg$), the number of compound included in training (No_{tr}) and test (No_{ts}) sets, the multiple correlation coefficient of each training (r_{tr}) and test (r_{ts}) sample and associated 95% confidence intervals (95% $Cl_{r_{tr}}$, for training sets; and 95% $Cl_{r_{ts}}$, for test sets), Fisher parameter and its significance, at a significance level of 5%, for training (F_{tr}) and test (F_{ts}) models, and Fisher Z-test of comparison between the correlation coefficient obtained in training set and the correlation coefficient obtained in corresponding test set ($Z_{r_{tr-ts}}$).

The estimation and prediction abilities of the multi-varied MDF-SPR model with four descriptors obtained in training versus test analysis, when the number of compounds in training set was equal

TABLE III
MDF-SPR models: statistical characteristics.

Parameter	Value	
	Two descriptors ($n = 30, v = 2$)	Four descriptors ($n = 30, v = 4$)
r (correlation coefficient)	0.9457	0.9890
Cl_r [lower, upper] (95% confidence intervals for r)	[0.8897,0.9740]	[0.9767,0.9948]
r^2 (squared correlation coefficient)	0.8943	0.9781
r^2_{adj} (adjusted correlation coefficient)	0.8865	0.9745
s_{est} (standard error)	0.3671	0.1739
F_{est} (Fisher parameter of regression model)	114 [†]	279 [†]
r^2_{cv-loo} (cross-validation loo squared correlation coefficient)	0.8660	0.9680
s_{loo} (standard error on cross-validation loo analysis)	0.4139	0.2101
F_{pred} (Fisher parameter on loo analysis)	87 [†]	189 [†]
$r^2 - r^2_{cv-loo}$ (model stability)	0.0284	0.0100
r^2 (<i>isDDkGg</i> , <i>IMmrKQg</i>)*	0.0760	0.0760
r^2 (<i>isDDkGg</i> , <i>IPMDKQg</i>)*	n.a.	0.3346
r^2 (<i>isDDkGg</i> , <i>IFMMKQg</i>)*	n.a.	0.0083
r^2 (<i>IMmrKQg</i> , <i>IPMDKQg</i>)*	n.a.	0.0232
r^2 (<i>IMmrKQg</i> , <i>IFMMKQg</i>)*	n.a.	0.1624
r^2 (<i>IPMDKQg</i> , <i>IFMMKQg</i>)*	n.a.	0.6214

n , number of components; v , number of descriptors.

* Squared correlation coefficient between descriptors; n.a., not applicable.

[†] $p < 0.0001$.

TABLE IV
Characteristics of MDF-SPR models.

	SE	r^2 (Y, desc)	t-stat	95% CI _{coefficient}
Multi-varied MDF-SPR model with two descriptors				
Intercept	0.2392	n.a.	4.4674*	[0.578, 1.559]
isDDkGg	0.0003	0.1813	10.2639*	[0.003, 0.004]
IMmrKQg	0.0297	0.4821	-13.4982*	[-0.462, -0.340]
Multi-varied MDF-SPR model with four descriptors				
Intercept	0.1710	n.a.	0.5072*	[-0.265, 0.439]
isDDkGg	0.0003	0.1813	20.375*	[0.005, 0.006]
IMmrKQg	0.0157	0.4821	-26.447*	[-0.449, -0.384]
IPMDKQg	0.0010	0.1392	9.4236*	[0.007, 0.012]
IFMMKQg	0.0109	0.1205	-7.1719	[-0.100, -0.056]

SE, standard error; Y = $\log K_{ow}$; desc, molecular descriptor; n.a., not applicable.

* $p < 0.0001$.

with 2/3 from the total number of compounds, are presented in Figure 3.

Discussion

The MDF-SPR methodology proved a useful method in estimation and prediction of the octanol-water partition coefficient for studied para-substituted phenols, this property being in relationship with complex information obtained from the compounds structure. The best estimation and predic-

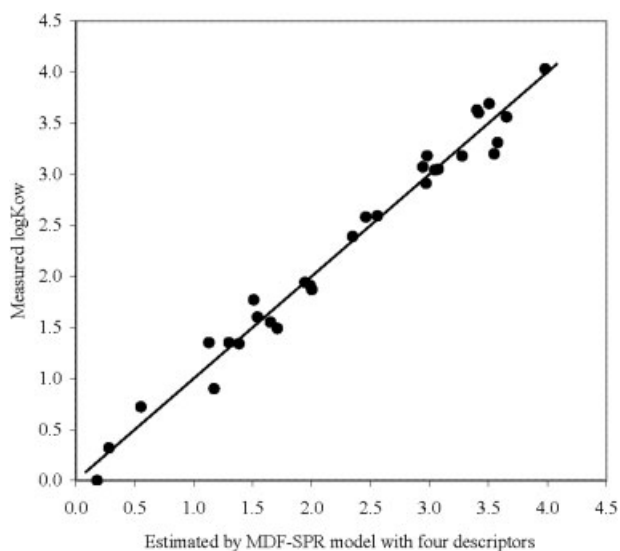


FIGURE 2. Estimated by MDF-SPR model with four descriptors versus measured $\log K_{ow}$.

tion abilities were obtained by the multi-varied MDF-SPR models with two and four descriptors [Eqs. (1) and (2)].

The analysis of the MDF-SPR model with two descriptors (Eq. (1)) revealed that the octanol-water partition coefficient of studied para-substituted phenols was strongly related with molecular geometry (*isDDkGg*, *IMmrKQg*), being dependent on the partial charges (*IMmrKQg*) and group electronegativity (*isDDkGg*), directly related with the elastic force (*IMmrKQg*) and inverse related with the property potential (*isDDkGg*). For one descriptor (*isDDkGg*), its intercept had positive regression coefficients, while the other (*IMmrKQg*) had a negative one. The intercept of one descriptor (*isDDkGg*) had positive regression coefficients while the other (*IMmrKQg*) had a negative one.

The analysis of the performances of the model with two descriptors concluded that this was statistically significant in estimation as well as in prediction (see the squared correlation coefficient, adjusted values, and leave-one-out score, Table III).

TABLE V
Correlated correlation analysis results.

Parameter	Value
$r(\log K_{ow}, \hat{Y}_{4D})$	0.9890
$r(\log K_{ow}, \hat{Y}_{2D})$	0.9457
$r(\hat{Y}_{4D}, \hat{Y}_{2D})$	0.9562
Steiger's z-test	4.3501
p	<0.001

TABLE VI
Statistical characteristics of MDF-SPR models in training versus test analysis.

a ₀	Models coefficients				Training set				Test set				
	a ₁	a ₂	a ₃	a ₄	No _{tr}	r _{tr}	95% Clr _{tr}	F _{tr}	No _{ts}	r _{ts}	95% Clr _{ts}	F _{ts}	Z _{tr-ts}
249 · 10 ⁻²	5.69 · 10 ⁻³	-4.31 · 10 ⁻¹	9.54 · 10 ⁻³	-8.11 · 10 ⁻²	15	0.996	[0.987, 0.999]	299*	15	0.981	[0.942, 0.994]	52*	1.87
-2.19 · 10 ⁻¹	5.93 · 10 ⁻³	-4.20 · 10 ⁻¹	1.21 · 10 ⁻²	-1.12 · 10 ⁻¹	16	0.989	[0.968, 0.996]	124*	14	0.985	[0.953, 0.995]	35*	1.19†
3.04 · 10 ⁻²	5.66 · 10 ⁻³	-4.17 · 10 ⁻¹	1.06 · 10 ⁻²	-9.34 · 10 ⁻²	17	0.986	[0.960, 0.995]	104*	13	0.988	[0.959, 0.996]	80*	0.63†
-2.01 · 10 ⁻²	5.58 · 10 ⁻³	-3.83 · 10 ⁻¹	1.06 · 10 ⁻²	-9.40 · 10 ⁻²	18	0.988	[0.966, 0.995]	130*	12	0.987	[0.954, 0.996]	43*	0.02†
7.92 · 10 ⁻²	5.50 · 10 ⁻³	-4.12 · 10 ⁻¹	9.04 · 10 ⁻³	-7.40 · 10 ⁻²	19	0.992	[0.979, 0.997]	221*	11	0.981	[0.925, 0.995]	29*	1.05†
2.74 · 10 ⁻¹	5.36 · 10 ⁻³	-4.33 · 10 ⁻¹	8.86 · 10 ⁻³	-6.96 · 10 ⁻²	20	0.990	[0.974, 0.996]	186*	10	0.985	[0.936, 0.996]	39*	0.46†
1.93 · 10 ⁻¹	5.52 · 10 ⁻³	-4.41 · 10 ⁻¹	8.94 · 10 ⁻³	-6.88 · 10 ⁻²	21	0.992	[0.979, 0.997]	241*	9	0.983	[0.920, 0.997]	20**	0.75†
1.03 · 10 ⁻¹	5.41 · 10 ⁻³	-4.04 · 10 ⁻¹	9.95 · 10 ⁻³	-8.85 · 10 ⁻²	22	0.986	[0.965, 0.994]	145*	8	0.992	[0.955, 0.999]	34**	0.60†
4.49 · 10 ⁻²	5.67 · 10 ⁻³	-4.19 · 10 ⁻¹	9.31 · 10 ⁻³	-7.41 · 10 ⁻²	23	0.988	[0.972, 0.995]	186*	7	0.988	[0.919, 0.998]	20**	0.01†

*p < 0.0001; **0.0001 < p < 0.05; †p > 0.05.

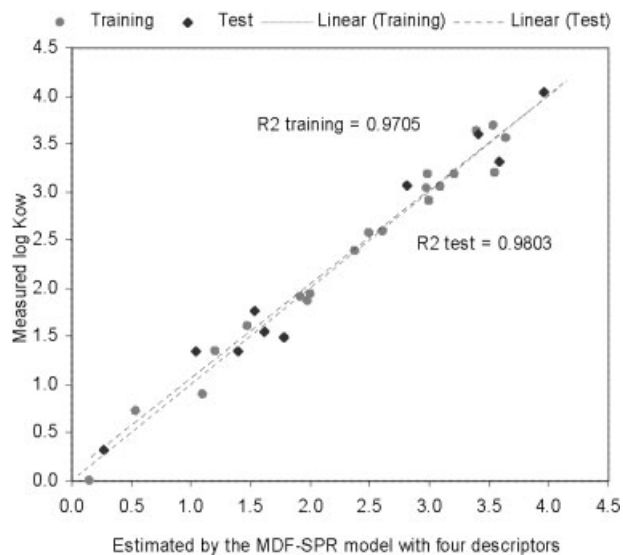


FIGURE 3. Prediction and estimation abilities of the multi-varied MDF-SPR model with four descriptors in training versus test analysis when the number of compounds in training set was equal to 2/3n.

Almost 90 percent of the octanol–water partition coefficient for studied para-substituted phenols can be explained by its linear relationship with the variation of *isDDkGg* and *IMmrKQg* descriptors (model with two descriptors, Table III). The goodness-of-fit of the MDF-SPR model with two descriptors is sustained by the correlation coefficient, which is equal with 0.9457, its validity by the significance of the model and standard error, while its predictive abilities by the cross-validation leave-one-out squared correlation coefficient, and by the Fisher parameter and its significance in leave-one-out analysis, which is <0.0001. The multi-varied MDF-SPR model with two descriptors proved a valid and stable model ($r_{cv-100}^2 = 0.8660$; $r^2 - r_{cv-100}^2 = 0.0284$).

The first thing that can be observed by analyzing the multi-varied MDF-SPR model with four descriptors [Eq. (2)] refers to the molecular descriptors used by the model: two are the descriptors used by the MDF-SPR model with two descriptors. The analysis of the molecular descriptors used by the multi-varied model with four descriptors suggests that the octanol–water partition coefficient of studied para-substituted phenols is strongly related to molecular geometry (*isDDkGg*, *IMmrKQg*, *IPMD-KQg*, *IFMMKQg*), partial charges (*IMmrKQg*, *IPMD-KQg*, *IFMMKQg*) and group electronegativity (*is-DDkGg*), it is in relation to the elastic force

(*IMmrKQg*, *IPMDKQg*, *IFMMKQg*), and inverse related with the property potential (*isDDkGg*).

The estimation abilities of the multi-varied MDF-SPR model with four descriptors are sustained by the value of the correlation coefficient ($r = 0.9890$, Table III), confidence boundaries associated with the regression coefficients and probabilities associated with Student test applied for the regression coefficients (for all coefficients <0.0001 , see Table IV). Almost 99% from the variation of the octanol-water partition coefficient of studied para-substituted phenols can be explained by its linear relationship with the variation of the four molecular descriptors used in the model [Eq. (2), Table III]. The value of the Fisher parameter ($F_{\text{pred}} = 189$) and its significance, which is <0.0001 , support the prediction abilities of the model. The stability of the multi-varied MDF-SPR model with four descriptors is sustained by the values of difference between the correlation coefficient and the cross-validation leave-one-out correlation score ($r^2 - r_{\text{cv-100}}^2 = 0.0100$), the value of the cross-validation score being very close to the value of the squared correlation coefficient. The power of the model with four descriptors in prediction of octanol-water partition coefficient of studied para-substituted phenols is sustained by the absence of co-linearity between descriptors (see the squared correlation coefficients between pairs of descriptors, which is <0.33 , with one exception, Table III) and/or between $\log K_{\text{ow}}$ and descriptors (see the squared correlation coefficients in Table IV, which are <0.48).

The comparison between multi-varied MDF-SPR models with two and four descriptors, respectively, can be performed by analyzing the residuals and/or the correlation coefficients. As far as the residuals are concerned, their values obtained by the MDF-SPR model with two descriptors varies from -1.1771 to 1.1861 , while the values obtained by the model with four descriptors varies from -0.8964 to 0.5717 . The analysis of the absolute value of residuals obtained by the MDF-SPR models reveals that the minimum values were obtained in 19 cases by the MDF-SPR model with four descriptors. The comparison between MDF-SPR models revealed that the model with four descriptors obtained a significantly greater correlation coefficient compared with the model with two descriptors ($p < 0.0001$, Table V). The regression model with two descriptors, as well as the model with four descriptors, respects the specification of Hawkins [24] regarding the number of descriptors according to sample size.

The goodness-of-fit of the multi-varied MDF-SPR model with four descriptors and its internal predictivity was assessed in training versus test analysis. The analysis was performed by splitting the sample of compounds into training and test sets, the allocation of a compound into a set or into another being performed through randomization.

The analysis of the results concluded that, with two exceptions, the values of coefficients of the models in training sets did not exceeded the 95% confidence intervals of the multi-varied MDF-SPR model with four descriptors. With one exception, when the value was greater than the upper 95% confidence interval boundary, the correlation coefficients obtained in training and test sets did not exceed the 95% confidence intervals associated with the correlation coefficient of the multi-varied MDF-SPR model with four descriptors (see values in Tables III and VI). As noted in Table VI, with one exception (for sample size in training set equal with 15), the correlation coefficients obtained in training sets were not statistical significant different, at a significance level of 5%, compared with the values obtained in test sets ($p > 0.05$, Table V).

The multi-varied MDF-SPR model with four descriptors can be used to predict the octanol-water partition coefficient of para-substituted phenols without any experiments and measurements. By using the MDF SPR predictor application [25], the property of a new para-substituted phenol can be obtained in a short time, provided that its structure is a *.hin file.

Conclusions

The octanol-water partition coefficient of para-substituted phenols proved to be strongly related with compounds geometry, partial charges and elastic force and in relation with group electronegativity and inverse related with the property potential.

The goodness-of-fit of the multi-varied MDF-SPR model with four descriptors and internal validation results sustain that the model is both stable and valid. Future studies on new external para-substituted phenols are necessary in order to assess the robustness and predictivity of the multi-varied MDF-SPR model with four descriptors.

References

1. Sangster, J. Octanol–Water Partition Coefficients: Fundamentals and Physical Chemistry; Wiley & Sons, Chichester, UK, 1997.
2. Padmanabhan, J.; Parthasarathi, R.; Subramanian, V.; Chattaraj, P. K. *Bioorg Med Chem* 2006, 14, 1021.
3. Kim, I.-H.; Morisseau, C.; Watanabe, T.; Hammock, B. D. *J Med Chem* 2004, 47, 2110.
4. Dimitrov, S.; Koleva, Y.; Schultz, T. W.; Walker, J. D.; Mekenyanyan, O. *Environ Toxicol Chem*, 2004, 23, 463.
5. Puzyn, T.; Rostkowski, P.; Świeczkowski, A.; Jedrusiak, A.; Falandysz, J. *Chemosphere* 2006, 62, 1817.
6. Leo, A. J. *Chem Rev* 1993, 93, 1281.
7. Poulin, P.; Krishnan, K. *Toxicol Method* 1996, 6, 117.
8. Bodor, N.; Buchwald, P. *J Phys Chem B* 1997, 101, 3404.
9. Pannier, A. K.; Brand, R. M.; Jones, D. D. *Pharm Res* 2003, 20, 143.
10. Zheng, G.; Huang, W. H.; Lu, X. H. *Anal Bioanal Chem* 2003, 376, 680.
11. Jäntschi, L. *Leonardo Electronic Journal of Practices and Technologies* 2005, 6, 76.
12. Jäntschi, L. *Leonardo J Sci* 2004, 4, 67.
13. Jäntschi, L. *Appl Med Inform* 2004, 15, 48.
14. Jäntschi, L., Bolboacă, S. *Leonardo Elect J Pract Technol* 2006, 8, 71.
15. Jäntschi, L. *Leonardo J Sci* 2004, 5, 63.
16. Jäntschi, L.; Bolboacă, S. *Leonardo Elect J Pract Technol* 2005, 7, 55.
17. Schultz, T. W. *Bull Environ Contam Toxicol* 1987, 38, 994.
18. ***, HyperChem, Molecular Modelling System [online]; ©2003, Hypercube [cited 2005 Nov]. Available from: URL: <http://hyper.com/products/>.
19. Jäntschi, L.; Katona, G.; Diudea, V. M. *Commun Math Comput Chem (MATCH)* 2000, 41, 151.
20. Diudea, M.; Gutman, I.; Jäntschi, L. *Molecular Topology*; 2nd ed.; Nova Science: Huntington, NY, 2002.
21. ***, Leave-one-out Analysis. ©2005, Virtual Library of Free Software [cited 2006 March]. Available from: URL: http://vl.academicdirect.org/molecular_topology/mdf_findings/loo.
22. Steiger, J. H. *Psychol Bull* 1980, 87, 245.
23. ***, Training vs. Test Experiment. ©2005, Virtual Library of free Software [cited 2006 March], Available from: URL: http://vl.academicdirect.org/molecular_topology/qsar_qspr_s/.
24. Hawkins, D. M. *J Chem Inf Comput Sci* 2004, 44, 1.
25. ***, MDF SPR-SAR Predictor, ©2005, Virtual Library of Free Software [cited 2006 March]. Available from: URL: http://vl.academicdirect.org/molecular_topology/mdf_findings/sar.