

# Predictive QSPR Study of the Dissociation Constants of Diverse Pharmaceutical Compounds

Andrew G. Mercader<sup>1,2\*</sup>, Mohammad Goodarzi<sup>3</sup>, Pablo R. Duchowicz<sup>1</sup>, Francisco M. Fernández<sup>1</sup> and Eduardo A. Castro<sup>1</sup>

<sup>1</sup>Instituto de Investigaciones Físicoquímicas Teóricas y Aplicadas (INIFTA, UNLP, CCT La Plata-CONICET), Diag. 113 y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina

<sup>2</sup>PRALIB (UBA-CONICET), Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Junín 956, C1113AAD Buenos Aires, Argentina

<sup>3</sup>Young Researchers Club and Department of Chemistry, Faculty of Sciences- Islamic Azad University, Arak Branch, PO Box 38135–567 Arak, Markazi, Iran

\*Corresponding author: Andrew G. Mercader, amercader@inifta.unlp.edu.ar

**The objective of the article was to perform a predictive analysis, based on quantitative structure–property relationships, of the dissociation constants ( $pK_a$ ) of different medicinal compounds (e.g., salicylic acid, salbutamol, lidocaine). Given the importance of this property in medicinal chemistry, it is of interest to develop theoretical methods for its prediction. The descriptors selection from a pool containing more than a thousand geometrical, topological, quantum-mechanical, and electronic types of descriptors was performed using the enhanced replacement method. Genetic algorithm and the replacement method (RM) techniques were used as reference points. A new methodology for the selection of the optimal number of descriptors to include in a model was presented and successfully used, showing that the best model should contain four descriptors. The best quantitative structure–property relationships linear model constructed using 62 molecular structures not previously used in this type of quantitative structure–property study showed good predictive attributes. The root mean squared error of the 26 molecules test set was 0.5600. The analysis of the quantitative structure–property relationships model suggests that the dissociation constants depend significantly on the number of acceptor atoms for H-bonds and on the number of carboxylic acids present in the molecules.**

**Key words:** enhanced replacement method, pharmaceutical compounds,  $pK_a$ , QSPR

Received 30 September 2009, revised and accepted for publication 29 August 2010

Knowledge of the physicochemical properties of a drug compound, e.g., its acid–base properties, is important in the optimization stage of a drug development project (1). The dissociation constant ( $pK_a$ ) is a measure of the tendency of a molecule or ion to keep a proton at its ionization center(s) (2). In biological terms,  $pK_a$  is important in determining whether a molecule will be taken up by aqueous tissue components or lipid membranes and is related to  $\log P$  (the partition coefficient) (3). Because most drugs are ionized in physiological conditions,  $pK_a$  is particularly relevant to medicinal chemistry because it is major factor in the pharmacokinetics of drugs (3,4). Commonly, dissociation constants of drug compounds are determined by techniques such as titration by potentiometry and UV–Vis spectrometry (1). Although highly useful, these techniques typically need sample amounts in the order of a few mg for analysis (1). Moreover, with these techniques, there is no differentiation in analytical response between the analyte of interest and any analog impurity (1).

Therefore, it is of great interest to be able to predict the  $pK_a$  of compounds that have not yet been tested experimentally, as well as attempting to determine which structural parameters have an effect on the  $pK_a$  values. A generally accepted remedy for the lack of experimental data in complex chemical phenomena is the analysis based on quantitative structure–property relationships (QSPR) (5).

The ultimate role of the different formulations of the QSPR theory is to suggest mathematical models for estimating relevant properties of interest, especially when they cannot be experimentally determined for some reason. These studies simply rely on the assumption that the physicochemical properties of a compound are determined solely by its molecular structure. The molecular structure is therefore translated into the so-called molecular descriptors through mathematical formulae obtained from several theories, such as chemical graph theory, information theory, and quantum mechanics (6,7). Currently, there are thousands of theoretical descriptors available in the literature, and one usually faces the problem of selecting those which are the most representative of the property under consideration (8).

The main objective of the research presented in this paper was to develop a model for the prediction of the dissociation constants ( $pK_a$ ) of 88 (62 training set and 26 test set) drug compounds (e.g., salicylic acid, salbutamol, lidocaine) whose experimental data were collected from the literature and were not used in a predictive study before. Furthermore, a recently developed methodology for the determination of the optimal number of descriptors will be presented and applied.

## Materials and Methods

### Data Set

In this study, we used a training set of 62 compounds and a test set of 26 compounds with known dissociation constants ( $pK_a$ ) measured in zero ionic strength aqueous solutions (1,9–17). The training set was selected with the purpose of having a distribution of data as normal as possible. The test molecules were chosen randomly taking care that their experimental  $pK_a$  values were sufficiently representative of the whole span. This was achieved by taking a random selection of the test set and afterward checking that the selection was spread over the experimental values; if the selection was not properly dispersed, the process was repeated. Table 1 shows the compound names and their experimental  $pK_a$  values.

### Molecular Descriptors

The structures of the compounds were firstly pre-optimized with the molecular mechanics force field (MM+) procedure included in the Hyperchem 6.03 package<sup>a</sup>, and the resulting geometries were further refined by means of the semiempirical method PM3 (parametric method-3) using the Polak-Ribiere algorithm and a gradient norm limit of 0.01 kcal Å<sup>-1</sup>. We computed the molecular descriptors by means of the software Dragon 5.0<sup>b</sup>, including parameters of all types: Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology, and Atoms-Weighted Assembly), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), Molecular Walk Counts, BCUT descriptors, 2D-Autocorrelations, Aromaticity Indices, Randic Molecular Profiles, Radial Distribution Functions, Functional Groups, and Atom-Centered Fragments (8). Additionally, four quantum-chemical descriptors (molecular dipole moments, total energies, HOMO-LUMO energies), which are not provided by the program Dragon, were added to the pool so the selection methodology was able to choose the most suitable descriptors form a pool with higher diversity. The resulting pool contained  $D = 1294$  descriptors.

### Model Search

In our calculations, we employ the computer system Matlab 5.0.<sup>c</sup> It is our purpose to search the set  $\mathbf{D}$  of  $D$  descriptors, for an optimal subset  $\mathbf{d}$  of  $d \ll D$  ones with minimum standard deviation  $S$ ,

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

by means of the multivariable linear regression (MLR) technique. In this equation,  $N$  is the number of molecules in the training set, and  $\text{res}_i$  the residual for molecule  $i$ , the difference between the experimental property ( $\mathbf{p}$ ) and predicted one ( $\mathbf{p}_{\text{pred}}$ ). More precisely, we want to obtain the global minimum of  $S(\mathbf{d})$ , where  $\mathbf{d}$  is a point in a space of  $D!/[(d!(D-d)!)]$  ones. Each point is a possible model of  $d$  descriptors as discussed below. Taking into account that a full search (FS) of optimal variables is impractical because it requires  $D!/[(d!(D-d)!)]$  linear regressions, some time ago, we proposed the replacement method (RM) (18–21), and later the enhanced replacement method (ERM) (22), that

produce linear regression QSPR models that are quite close the FS ones with much less computational work. These alternative techniques approach the minimum of  $S$  by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of  $d$  descriptors  $\mathbf{d} = \{X_1, X_2, \dots, X_d\}$ . The RM gives models with better statistical parameters than the forward stepwise regression procedure (23) and variants of the more elaborated genetic algorithms (24). The ERM leads to even better statistical parameters with slightly more computational work (22).

A GA is a search technique based on natural evolution where variables play the role of genes (in this case, a set of descriptors) in an individual of the species. An initial group of random individuals (population) evolves according to a fitness function (in this case, the standard deviation) that determines the survival of the individuals. The GAs offer a combination of hill-climbing ability (natural selection) and a stochastic method (crossover and mutation) and explore many solutions in parallel, processing information in a very efficient manner. The practical application of GAs requires the tuning of some parameters such as population size, generation gap, crossover rate, and mutation rate. These parameters typically interact among themselves nonlinearly and cannot be optimized one at a time. There is considerable discussion about parameter settings and approaches to parameter adaptation in the evolutionary computation literature; however, there does not seem to be conclusive results on which may be the best (25).

The GA parameter optimization required several runs, leading to the following results: number of individuals = 250; generation gap = 0.9; single point crossover probability = 0.6; mutation probability =  $0.7/d$ . The implementation of GA was performed stopping each run when one individual occupied more than 90% of the population or when the number of generations reached 1500.

The Kubinyi function ( $FIT$ ) (26,27) is a statistical parameter that closely relates to the Fisher ratio ( $F$ ), but avoids the main disadvantage of the latter that is too sensitive to changes in small  $d$  values and poorly sensitive to changes in large  $d$  values. The  $FIT(\mathbf{d})$  criterion has a low sensitivity to changes in small  $d$  values and a substantially increasing sensitivity for large  $d$  values. The greater the  $FIT$  value the better the linear equation. It is given by the following expression:

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

where  $R$  is the correlation coefficient,  $N$  is the number of molecules in the training set, and  $d$  is the number of descriptors included in the model. It is expected that a plot of  $FIT$  vs.  $d$  presents a maximum from which it is possible to calculate the optimal number of molecular descriptors ( $d_{opt}$ ) to be included in the linear regression model. There are many occasions when the maximum is not reached after adding a reasonable number of descriptors in the model. For this reason, we recently proposed a variable  $FIT$  equation or  $VFIT$  that depends on an adjustable parameter  $k$  that gives more weight to  $d$  in the numerator of the  $FIT$  equation (28). It reads:

**Table 1:** Experimental and predicted (eqn 4) dissociation constants ( $pK_a$ ) and residuals

Number	Name of molecules	CAS	$pK_a$ exp.	$pK_a$ pred.	Residual
<i>Training set</i>					
1	Oxprenolol	6452-71-7	9.3	9.20	0.10
2	Protriptyline	438-60-8	10.7	10.22	0.48
3	Trimipramine	739-71-9	9.4	9.71	-0.31
4	Quinine	130-95-0	9.7	8.90	0.80
5	Salbutamol	34391-04-3	10.3	9.66	0.64
6	Tulobuterol	41570-61-0	10.4	9.53	0.87
7	Procainamide	51-06-9	9.2	9.13	0.07
8	Morphine	57-27-2	9.9	8.52	1.38
9	Codeine	76-57-3	8.2	8.27	-0.07
10	Lidocaine	137-58-6	7.9	8.89	-0.99
11	Sumatriptan	103628-46-2	9.6	8.30	1.30
12	Buspirone	36505-84-7	7.2	7.20	0.00
13	Bufuralol	57704-16-2	9	9.60	-0.60
14	Bupivacaine	2180-92-9	8.1	9.60	-1.50
15	Mepivacaine	22801-44-1	7.7	8.95	-1.25
16	Prilocaine	721-50-6	7.9	9.25	-1.35
17	Ketamine	6740-88-1	7.5	8.94	-1.44
18	Acetaminophen	103-90-2	9.5	8.07	1.43
19	Phenylpropanolamine	14838-15-4	9.44	9.51	-0.07
20	4-Aminophenol	123-30-8	10.46	8.78	1.68
21	Verapamil	52-53-9	9.04	9.36	-0.32
22	Norverapamil	67018-85-3	9.87	9.59	0.28
23	D-617	Ref. (14)	10.35	9.34	1.01
24	Phenobarbital	50-06-6	7.41	7.41	0.00
25	Barbital	57-44-3	7.91	7.09	0.82
26	Amobarbital	57-43-2	7.94	7.77	0.17
27	Diltiazem	42399-41-7	7.75	7.60	0.15
28	Rifampicin	13292-46-1	7.58	8.34	-0.76
29	Promazine	58-40-2	9.09	9.09	0.00
30	Indapamide	26807-65-8	9.16	7.81	1.35
31	Desipramine	50-47-5	10.28	9.71	0.57
32	Triflupromazine	146-54-3	8.56	7.26	1.30
33	Diazepam	3900-31-0	7.63	7.81	-0.18
34	Acetylsalicylic acid	50-78-2	3.74	3.85	-0.11
35	Benzoic acid	65-85-0	4.17	4.72	-0.55
36	4-Hydroxybenzaldehyde	123-08-0	7.58	7.86	-0.28
37	4-Hydroxybenzoic acid	99-96-7	4.44	4.60	-0.16
38	Nicotinic acid	59-67-6	4.84	3.06	1.78
39	Pyridine	110-86-1	5.27	7.80	-2.53
40	Salicylic acid	69-72-7	3.07	4.60	-1.53
41	Alminoprofen	39718-89-3	5.02	5.05	-0.03
42	Carprofen	53716-49-7	4.36	4.51	-0.15
43	Fenoprofen	31879-05-7	5.7	4.49	1.21
44	Flurbiprofen	5104-49-4	4.2	4.31	-0.11
45	Indoprofen	31842-01-0	4.25	4.00	0.26
46	Naproxen	22204-53-1	4.2	4.29	-0.09
47	Pirprofen	31793-07-4	4.64	4.31	0.33
48	Suprofen	40828-46-4	4.11	4.26	-0.15
49	Tiaprofenic acid	33005-95-7	3.8	4.22	-0.42
50	Imazapyr	81334-34-1	1.9	3.02	-1.12
51	Acifluorfen	62476-59-9	3.8	1.45	2.35
52	Imazethapyr	81385-77-5	2.1	3.43	-1.33
53	Nicosulfuron	111991-09-4	4.6	4.19	0.41
54	Thifensulfuron-methyl	79277-27-3	4	4.10	-0.10
55	Metsulfuron-methyl	74223-64-6	3.3	4.39	-1.09
56	Triasulfuron	82097-50-5	4.6	4.78	-0.18
57	Chlorsulfuron	64902-72-3	3.6	5.09	-1.49
58	Bensulfuron-methyl	83055-99-6	5.2	4.06	1.14
59	Flumetsulam	98967-40-9	4.6	4.33	0.27
60	Metosulam	139528-85-1	4.8	5.43	-0.63
61	Fomesafen	72178-02-0	2.7	3.78	-1.08

Table 1: (Continued)

Number	Name of molecules	CAS	$pK_a$ exp.	$pK_a$ pred.	Residual
62	Diclofop	40843-25-2	3.4	3.58	-0.18
<i>Test set</i>					
63	Acebutolol	37517-30-9	9.2	8.91	0.29
64	Procaine	59-46-1	8.9	8.63	0.27
65	Phenylephrine	59-42-7	8.9	8.90	0.00
66	Chlorpheniramine	132-22-9	9.14	8.94	0.20
67	Gallopamil	16662-46-7	9.01	9.19	-0.18
68	D-620	Ref. (14)	9.84	9.62	0.22
69	D-702	Ref. (14)	10.32	9.61	0.71
70	D-703	Ref. (14)	9.15	9.61	-0.46
71	D-715	Ref. (14)	9.88	9.86	0.02
72	Chlorpromazine	50-53-3	9.21	8.91	0.30
73	Levomopromazine	7104-38-3	9.15	8.92	0.23
74	Thioridazine	50-52-2	9.5	9.76	-0.26
75	Proprietary	2622-26-6	8.1	9.18	-1.08
76	Secobarbital	76-73-3	7.92	7.93	-0.01
77	Bupropion	34841-39-9	8.3	9.16	-0.86
78	Diphenhydramine	58-73-1	9.12	9.34	-0.22
79	Propranolol	318-98-9	9.55	9.52	0.03
80	Doxepin	1229-29-4	9.16	9.22	-0.06
81	Omeprazole	73590-58-6	6.15	7.06	-0.91
82	Alprenolol	13655-52-2	9.38	9.73	-0.35
83	Atenolol	29122-68-7	9.42	9.27	0.15
84	Metoprolol	51384-51-1	9.44	9.25	0.19
85	Ibuprofen	15687-27-1	4.55	5.15	-0.60
86	Ketoprofen	22071-15-4	4.18	4.51	-0.33
87	Fluazifop	83066-88-0	3.2	1.37	1.83
88	Imazaquin	81335-37-7	3.8	3.42	0.38

$$VFIT = \frac{R^2(N - kd - 1)}{(N + d^2)(1 - R^2)} \quad (3)$$

Using this equation, it is possible to obtain  $d_{opt}$  as the number of descriptors that yields a maximum ( $d_{max}$ ) value in a  $VFIT$  vs.  $d$  plot. A new technique to determine the parameter  $k$  is presented in this work; the procedure consists of taking incremental values of 0.5 in  $k$  until the maximum remains unchanged for two increments and complies with the rule of thumb that at least five data points should be present for each fitting parameter (29).

As a theoretical validation of all the models, we choose the well-known leave-one-out (*loo*) and the leave-more-out cross-validation procedures (*l-n%-o*) (30), where  $n\%$  represents the percentage of molecules removed from the training set. We generated 5 000 000 cases of random data removal for *l-n%-o*, where  $n\% = 11\%$  (seven compounds). In addition, with the purpose of demonstrating that the best model found does not result from happenstance, we resorted to a widely used approach to establish the model robustness: the so-called  $\gamma$ -randomization (31). It consists in scrambling the experimental property  $\mathbf{p}$  in such a way that the property value and the compound do not match; 5 000 000 cases of random scrambling were generated.

## Results and Discussion

To determine the optimal number of descriptors, we calculated different predictive relationships with the ability to link the molecular

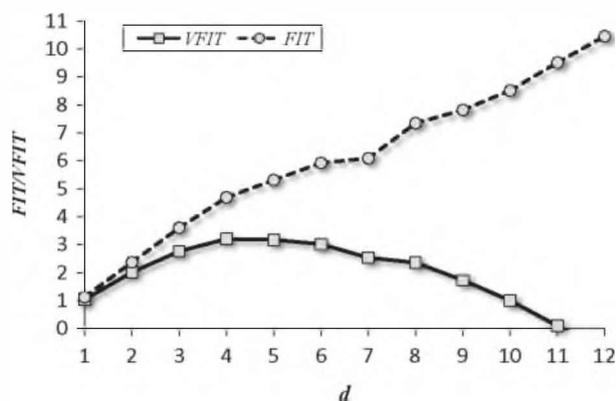
structure of the drug compounds with the dissociation constants ( $pK_a$ ), by means of linear regression models with 1–12 parameters ( $d$ ) that were selected by ERM from the pool of  $D = 1294$  descriptors.

As can be seen in Table 2, as  $k$  in  $VFIT$  is increased a first maximum appears at  $d = 8$  ( $k = 3.5$ ) that remains for one increment, a second one  $d = 6$  ( $k = 4.5$ ), a third one at  $d = 5$  ( $k = 5$ ), afterward a maximum at  $d = 4$  ( $k = 5.5$ ) that remains for six more increments ( $k = 6, 6.5, 7, 7.5, 8, 8.5$ ) and complies with the above-mentioned practical rule (29).

Thus, the resulting  $VFIT$  with  $k = 5.5$  increases with  $d$  up to a maximum value  $d = d_{max} = 4$  shown in Figure 1. We assume that this is the optimal value of descriptors in the model. Figure 1 also shows that  $FIT$  does not present a maximum in the interval of  $d$  between

**Table 2:** Incremental values of  $k$  and the resulting number of descriptors ( $d$ ) that present a maximum in  $VFIT$

$k$	$d$ (max. in $VFIT$ )	$k$	$d$ (max. in $VFIT$ )
1.5	–	5.5	4
2	–	6	4
2.5	–	6.5	4
3	–	7	4
3.5	8	7.5	4
4	8	8	4
4.5	6	8.5	4
5	5	9	3



**Figure 1:** *VFIT* and *FIT* as functions of the number of descriptors for the training set.

1 and 12. We thus conclude that the best QSPR model according to ERM is

$$\begin{aligned}
 pK_a = & 7.4368(\pm 0.4) + 5.854(\pm 0.7) \cdot R_{DF010m} \\
 & - 4.0272(\pm 0.3) \cdot nCOOH - 3.1513(\pm 0.5) \cdot nCOOHPh \quad (4) \\
 & - 0.5444(\pm 0.04) \cdot nHAcc
 \end{aligned}$$

$$N = 62, R = 0.9301, S = 0.9865, FIT = 4.685, p < 10^{-7}$$

$$R_{loo} = 0.9106, S_{loo} = 1.1144, R_{l-11\%-o} = 0.8433, S_{l-11\%-o} = 1.4781$$

$$R_{TS} = 0.9657, RMSE_{TS} = 0.5600$$

Here, the absolute errors of the regression coefficients are given in parentheses;  $p$  is the significance of the model, *FIT* the Kubinyi

**Table 3:** Linear QSPR models for the training set with  $N = 62$ . The best relationship appears in boldface

Model	Descriptors used	$R$	$S$	$RMSE_{TS}$
M1	<i>MATS1m</i>	0.7360	1.7728	1.4574
M2	<i>Me</i> , <i>nCOOH</i>	0.8523	1.3814	0.8563
M3	<i>nDB</i> , <i>nCOOH</i> , <i>nCOOHPh</i>	0.9029	1.1449	0.6678
<b>M4</b>	<b><i>RDF010m</i>, <i>nCOOH</i>, <i>nCOOHPh</i>, <i>nHAcc</i> (eqn 4)</b>	<b>0.9301</b>	<b>0.9865</b>	<b>0.5600</b>
M5	<i>nDB</i> , <i>RDF010e</i> , <i>H6m</i> , <i>nOHPh</i> , <i>O-057</i>	0.9444	0.8914	0.5891

**Table 4:** Meaning of the symbols for the molecular descriptors appearing in the different models

Molecular descriptor	Type	Description
<i>MATS1m</i>	2D Autocorrelations	Moran autocorrelation – lag 1/weighted by atomic masses
<i>Me</i>	Constitutional	Mean atomic Sanderson electronegativity (scaled on Carbon atom)
<i>nCOOH</i>	Constitutional	Number of carboxylic acids (aliphatic)
<i>nDB</i>	Constitutional	Number of double bonds
<i>nCOOHPh</i>	Constitutional	Number of carboxylic acids (aromatic)
<i>RDF010m</i>	Radial Distribution Function	Radial Distribution Function – 1.0/weighted by atomic masses
<i>nHAcc</i>	Constitutional	Number of acceptor atoms for H-bonds (N O F)
<i>RDF010e</i>	Radial Distribution Function	Radial Distribution Function – 1.0/weighted by atomic Sanderson electronegativities
<i>H6m</i>	GETAWAY	H autocorrelation of lag 6/weighted by atomic masses
<i>nOHPh</i>	Constitutional	Number of phenols
<i>O-057</i>	Atom-Centered Fragments	Phenol/enol/carboxyl OH

function, *loo* and *l-11%-o* stand for the leave-one-out and leave-more-out cross-validation techniques, respectively, RMSE stands for root mean squared errors, and *TS* stands for test set.

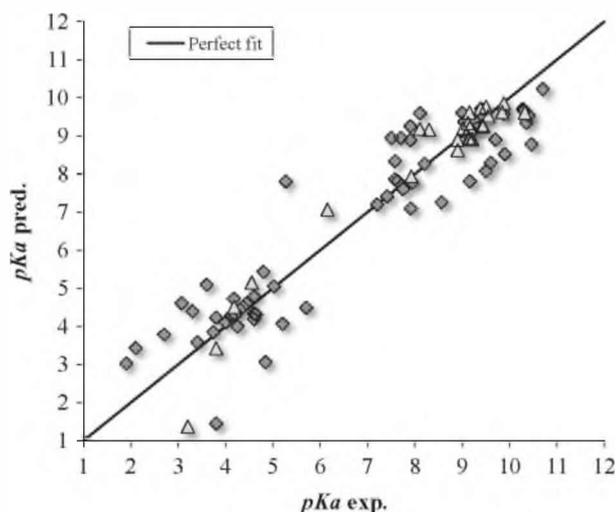
Following the same strategy, the RM yields the same model found by ERM. We also tried the GA on the same problem, the best four descriptors model obtained after twenty runs using the previously mentioned optimized parameters was also the model in eqn (4) found by ERM.

Table 3 shows a summary of the linear models with 1 to  $d_{opt} + 1$  parameters for ERM. As can be seen that the training set statistical parameters of the models improve through  $d = 5$  nevertheless RMSE of the test set only improves through  $d = 4$ . This indicates that the optimal number of descriptors is  $d = 4$ , corroborating the results obtained using the above-mentioned *VFIT* methodology, and that the model with  $d = 5$  is possibly over fitted. Table 4 displays the details of the molecular descriptors of Table 3.

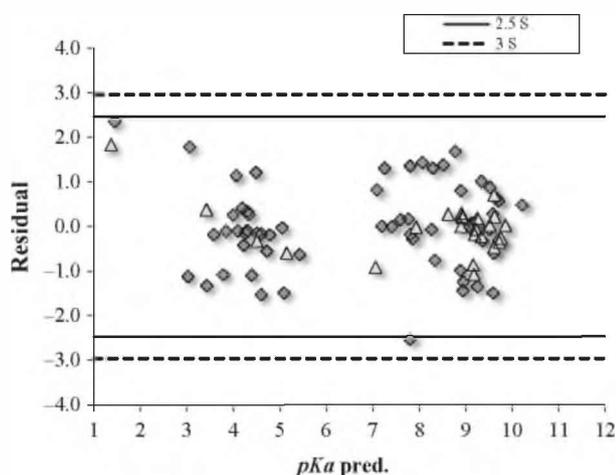
After analyzing 5 000 000 cases of  $\gamma$ -randomization on eqn (4), the smallest  $S$  value obtained was 2.0508 is considerably larger than the one found in the true calibration ( $S = 0.9865$ ). In this way, the robustness of the model could be further proved, showing that the calibration resulted in a true structure–property relationship and was not a fortuitous correlation.

The plot of predicted vs. experimental  $pK_a$  shown in Figure 2 suggests that the 62 training and 26 test set compounds approximately follow a straight line. Table 1 also includes the predicted dissociation constants ( $pK_a$ ) obtained via eqn (4) for the training and test sets and the corresponding residuals. Figure 3 shows that the behavior of the residuals in terms of the predictions follows a normal distribution. No molecule in the set exhibits a residual larger than 3S that can be considered as an outlier.

The correlation matrix shown in Table 5 reveals that the descriptors of the linear model are not seriously inter-correlated ( $R_{ij} < 0.3842$ ) and therefore, all the descriptors contain structural information that is not overlapped with any of the rest of the descriptors in the model, which justifies the appearance of all those parameters in the equation. The predictive power of the linear model is satisfactory as revealed by its stability upon the inclusion or exclusion of compounds, measured by the statistical parameters  $R_{loo} = 0.9106$



**Figure 2:** Experimental  $pK_a$  versus calculated  $pK_a$  using eqn (4) for the training set (rhombus) and test set (triangles).



**Figure 3:** Dispersion plot of the residuals for the training (rhombus) and test sets (triangles) according to eqn (4).

and  $R_{1-11\%-0} = 0.8433$ . According to the literature,  $R_{1-11\%-0}$  must be greater than 0.71 to have a validated model (32).

An adequate contrast of the presented model with previously reported ones is not feasible because they are based on sets of molecules of different nature and size. Nevertheless, a few examples are presented as a reference points: Jover *et al.* (33) constructed QSPR model based on the nonlinear more sophisticated neural network arriving to a model that showed an RMSE of the test set of 0.95; Lee and Crippen (4) have compiled many models in a recent review from which the only comparable model in the training set size is a linear free energy relationship found by Dixon and Jurs(34) that presented a RMSE of 0.471 on the test set, using a one family set of molecules in contrast to the broader set used in the present work; Harding *et al.* (2) presented several models based on a quantum topological molecular similarity (QTMS) study,

**Table 5:** Correlation matrix for the descriptors in eqn (4) ( $N = 62$ ) The highest correlation appears in boldface

	<i>RDF010m</i>	<i>nCOOH</i>	<i>nCOOHPh</i>	<i>nHAcc</i>
<i>RDF010m</i>	1	0.1769	0.2903	<b>0.3842</b>
<i>nCOOH</i>		1	0.1526	0.2005
<i>nCOOHPh</i>			1	0.0602
<i>nHAcc</i>				1

nevertheless no external validation was reported. Hence, as mentioned, an appropriate contrast is not possible; nonetheless, the presented model contrasts well with previously reported work using a data set with a high diversity of structures.

The molecular descriptors appearing in the linear eqn (4) combine different dimensional aspects of the molecular structure and can be classified as follows: (i) a radial distribution function: *RDF010m*, weighted by atomic masses; and three constitutional descriptors: *nCOOH*, number of aliphatic carboxylic acids, *nCOOHPh*, number of aromatic carboxylic acids, and *nHAcc*, number of acceptor atoms for H-bonds (N O F). The combination of the four selected descriptors is the best one for predicting the property under study ( $pK_a$ ), leading to a model with standard deviation  $S$  that is lower than that achieved by any other 4-descriptors equation obtained from the pool **D**.

A radial distribution function (RDF) (35) of an ensemble of atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of certain radius, also incorporating different atomic properties to differentiate the contribution of each atom to the property under study. For the case of *RDF010m*, the sphere radius is of 10.0 Å and atomic masses are employed to distinguish their nature.

Constitutional descriptors are OD-descriptors, independent from molecular connectivity and conformations (8). The descriptor *nCOOH* is determined by counting the number of aliphatic carboxylic acids present in a molecule. The descriptor *nCOOHPh* is determined by counting the number carboxylic acids in an aromatic ring present in a molecule. The descriptor *nHAcc* is determined by counting the number of acceptor atoms for H-bonds (with N, O, and F) (36).

The standardization of the regression coefficients of eqn (4) (23) allows assigning a greater importance to the molecular descriptors that exhibit larger absolute standardized coefficients. The descriptor order according to the standardized coefficients shown between parentheses is:

$$nHAcc(0.6642) > nCOOH(0.6364) > RDF010m(0.4594) > nCOOHPh(0.3331) \quad (5)$$

The first descriptor depends on the number of H-bond acceptors present in different functional groups; the second depends on importance and an additional descriptor depends on the number of carboxylic acid groups, this suggests that the dissociation constants ( $pK_a$ ) have a significant dependence on the number of carboxyl acid present in the molecule which may not be surprising.

## Conclusions

In this paper, we constructed a predictive QSPR model for dissociation constants ( $pK_a$ ), an important parameter in the optimization stage of a drug development project, from 62 different medicinal compounds using four molecular descriptors that take into account 2D- and 3D-aspects of the molecular structure. The model showed good predictive ability established by the theoretical and test set validations. We presented a successful strategy to determine the optimal number of descriptors in a QSPR model. Our results showed that in this case, the ERM gives identical results as GA and RM. The analysis of the QSPR model suggests that the dissociation constants depend significantly on the number of acceptor atoms for H-bonds and on the number of carboxylic acids present in the molecules.

We expect the presented model to be a useful tool in the prediction of  $pK_a$ , in a fast and costless manner, for any future studies that may require an estimation of this important physicochemical property.

## References

- Örnkov E., Linusson A., Folestad S. (2003) Determination of dissociation constants of labile drug compounds by capillary electrophoresis. *J Pharm Biomed Anal*;33:379–391.
- Harding A.P., Wedge D.C., Popelier P.L.A. (2009) pKa Prediction from "Quantum Chemical Topology" Descriptors. *J Chem Inf Model*;49:1914–1924.
- Milletti F., Storchi L., Sforza G., Cruciani G. (2007) New and Original pKa Prediction Method Using Grid Molecular Interaction Fields. *J Chem Inf Model*;47:2172–2181.
- Lee A.C., Crippen G.M. (2009) Predicting pKa. *J Chem Inf Model*;49:2013–2033.
- Hansch C., Leo A. (1995) Exploring QSAR. Fundamentals and Applications in Chemistry and Biology. Washington, DC: American Chemical Society.
- Trinajstić N. (1992) Chemical Graph Theory. Boca Raton, FL: CRC Press.
- Katritzky A.R., Lobanov V.S., Karelson M. (1995) QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. *Chem Soc Rev*;24:279–287.
- Todeschini R., Consonni V. (2000) Handbook of Molecular Descriptors. Weinheim, Germany: Wiley VCH.
- Laganà A., Fago G., Marino A., Penazzi V.M. (2000) Liquid chromatography mass spectrometry tandem for multiresidue determination of selected post-emergence herbicides after soil column extraction. *Anal Chim Acta*;415:41–56.
- Marín A., García E., García A., Barbas C. (2002) Validation of a HPLC quantification of acetaminophen, phenylephrine and chlorpheniramine in pharmaceutical formulations: capsules and sachets. *J Pharm Biomed Anal*;29:701–714.
- Pehourcq F., Jarry C., Bannwarth B. (2003) Potential of immobilized artificial membrane chromatography for lipophilicity determination of arylpropionic acid nonsteroidal anti-inflammatory drugs. *J Pharm Biomed Anal*;33:137–144.
- Burgot G., Burgot J.L. (2002) Protometric thermometric titrations of sparingly soluble compounds in water in the presence of n-octanol. *J Pharm Biomed Anal*;30:625–634.
- Popovic G.V., Sladic D.M., Stefanovic V.M., Pfendt L.B. (2003) Study on protolytic equilibria of lorazepam and oxazepam by UV and NMR spectroscopy. *J Pharm Biomed Anal*;31:693.
- Wallis M., Mullett W.M., Levens K., Borlak J., Wunsch G., Pawliszyn J. (2002) Verapamil drug metabolism studies by automated in-tube solid phase microextraction. *J Pharm Biomed Anal*;30:307–319.
- Needham S.R., Brown P.R. (2000) The high performance liquid chromatography electrospray ionization mass spectrometry analysis of diverse basic pharmaceuticals on cyanopropyl and pentafluorophenylpropyl stationary phases. *J Pharm Biomed Anal*;23:597–605.
- Cherkaoui S., Veuthey J.L. (2002) Use of negatively charged cyclodextrins for the simultaneous enantioseparation of selected anesthetic drugs by capillary electrophoresis–mass spectrometry. *J Pharm Biomed Anal*;27:615–626.
- Jia Z., Ramstad T., Zhong M. (2002) Determination of protein–drug binding constants by pressure-assisted capillary electrophoresis (PACE)/frontal analysis (FA). *J Pharm Biomed Anal*;30:405–413.
- Duchowicz P.R., Castro E.A., Fernández F.M., González M.P. (2005) A New Search Algorithm of QSPR/QSAR Theories: Normal Boiling Points of Some Organic Molecules. *Chem Phys Lett*;412:376–380.
- Duchowicz P.R., Castro E.A., Fernández F.M. (2006) Alternative Algorithm for the Search of an Optimal Set of Descriptors in QSAR-QSPR Studies. *MATCH Commun Math Comput Chem*;55:179–192.
- Duchowicz P.R., Fernández M., Caballero J., Castro E.A., Fernández F.M. (2006) QSAR of Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase. *Bioorg Med Chem*;14:5876–5889.
- Helguera A.M., Duchowicz P.R., Pérez M.A.C., Castro E.A., Cordeiro M.N.D.S., González M.P. (2006) Application of the Replacement Method as Novel Variable Selection Strategy in QSAR. 1. Carcinogenic Potential. *Chemom Intell Lab Syst*;81:180–187.
- Mercader A.G., Duchowicz P.R., Fernandez F.M., Castro E.A. (2008) Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories. *Chemom Intell Lab Syst*;92:138–144.
- Draper N.R., Smith H. (1981) Applied Regression Analysis. New York: John Wiley&Sons.
- So S.S., Karplus M. (1996) Evolutionary Optimization in Quantitative Structure-Activity Relationship: An Application of Genetic Neural Networks. *J Med Chem*;39:1521–1530.
- Melanie M. (1998) An Introduction to Genetic Algorithms. Cambridge, Massachusetts, London, England: A Bradford Book The MIT Press.
- Kubinyi H. (1994) Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution. *Quant Struct Act Relat*;13:393–401.
- Kubinyi H. (1994) Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quant Struct Act Relat*;13:285–294.

28. Mercader A.G., Duchowicz P.R., Fernández F.M., Castro E.A., Wolcan E. (2008) QSPR Study of solvent quenching of the  $^5D_0 \rightarrow ^7F_2$  emission of Eu(6,6,7,7,8,8,8-heptafluoro-2,2-dimethyl-3,5-octanedionate)<sub>3</sub>. *Chem Phys Lett*;462:352–357.
29. Hansch C. (1990) *Comprehensive Drug Design*. New York: Pergamon Press.
30. Hawkins D.M., Basak S.C., Mills D. (2003) Assessing Model Fit by Cross-Validation. *J Chem Inf Model*;43:579–586.
31. Wold S., Eriksson L. (1995) *Chemometrics Methods in Molecular Design*. Weinheim: VCH.
32. Golbraikh A., Tropsha A. (2002) Beware of q<sup>2</sup>! *J Mol Graphics Modell*;20:269–276.
33. Jover J., Bosque R., Sales J. (2007) Neural Network Based QSPR Study for Predicting pK<sub>a</sub> of Phenols in Different Solvents. *QSAR Comb Sci*;26:385–397.
34. Dixon S.L., Jurs P.C. (1993) Estimation of pK<sub>a</sub> for organic oxyacids using calculated atomic charges. *J Comput Chem*;14:1460–1467.
35. Consonni V., Todeschini R., Pavan M. (2002) Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies. *J Chem Inf Model*;42:693.
36. Viswanadhan V.N., Ghose A.K., Revankar G.R., Robins R.K. (1989) Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J Chem Inf Comput Sci*;29:163–172.

## Notes

<sup>a</sup>HYPERCHEM. 6.03 (Hypercube) <http://www.hyper.com>.

<sup>b</sup>DRAGON. 5.0 Evaluation Version <http://www.disat.unimib.it/chm>.

<sup>c</sup>Matlab. 5.0 The MathWorks Inc. <http://www.mathworks.com>.