# QSPR study of the acidity of carbon acids in aqueous solutions

**Pablo R. Duchowicz and Eduardo A. Castro***

*CEQUINOR, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, C.C. 962, La Plata 1900, Argentina. E-mail: castro@dalton.quimica.unlp.edu.ar*

10.1070/MC2002v012n05ABEH001588

**The molecular set of this study comprises 32 carbon acids with their corresponding p$K_a$ values in water ranging from 25 to –6.2; molecular and topological parameters comprise $\Delta E$, $\Sigma\Delta q_X$, atom and bond descriptors.**

According to the Brönsted definition, any compound which has a hydrogen atom is an acid, since it may be lost as an acidic proton. Depending on the molecule, this process requires more or less energy, and the process may be spontaneous. As proton transfer reactions are crucial in chemistry, it is important to quantify the tendency of the molecule to lose its hydrogen atom as an acidic proton. This is the role played by the quantity defined as p$K_a$. The equilibrium of dissociation of a Brönsted acid depends on the interaction of the acid and its conjugate base with the solvent molecules. Therefore, the p$K_a$ value depends on the solvent medium where the measurement took place, and any reference to the p$K_a$ value of a certain compound will be meaningful only if the solvent is specified. Experimentally, the most studied medium is water, which justifies the choice for the medium used in this paper. Although water is itself a Brönsted acid, the processes studied will not be affected by the solvent self-ionization because even the least acidic compound considered is still approximately $1\times10^9$ times more acidic than pure water. The experimental p$K_a$ values of several compounds, mainly organic acids in water[1,2] are determined through very well established methods,[3] such as spectroscopy, potentiometry, conductimetry, competitive reactions, *etc*. A detailed discussion of the importance of p$K_a$ in chemistry, as well as the role of the proton in organic chemistry can be found in two seminal papers.[4,5]

The experimental p$K_a$ measurements of organic compounds always involve some artifice and approximations.[6] Theoretical models to calculate p$K_a$ data for compounds in solution have been proposed.[4,7–14] However, they all furnish p$K_a$ values which differ, in a greater or lesser extent, from the experimental results.[14]

With the exception of some work on a small number of benzoic acids, carboxylic acids have been conspicuously absent from the QSAR literature.[15–17] The aim of this work was to study the acidity of 32 carbon acids in water, with p$K_a$ values ranging from 25.6 to –6.2 resorting to a QSAR analysis based upon rather simple molecular descriptors and comparing results with other theoretical and experimental data in order to test the possibility to obtain confident theoretical estimations.

Recently, various theoretical molecular descriptors were investigated in the search of correlations with gas-phase acidity of compounds of the form MeZ.[18] With a modest HF/3-21G method, a reasonable correlation ($r = 0.961$) between the gas-phase acidity of the compounds and their calculated deprotonation enthalpies was obtained. Interestingly, the employment of larger basis sets and polarization [6-31G(*d*)] or diffuse functions [6-31+G(*edf*,2*p*)] did not improve the correlations, in agreement with similar remarks.[19,20] Soon afterwards, results derived in ref. 18 were extended to include the acid behaviour in water of a larger set of compounds, exhibiting greater structural variations. The p$K_a$ values in water of 32 different carbon acids of the form CHWYZ, where W, Y, and/or Z are electron-withdrawing groups, were correlated with theoretical descriptors reflecting charge and energy variations upon deprotonation, yielding reasonable correlations with readily obtained descriptors ($r \geq 0.95$).[21] Experimental data were correlated with two-parameter equations, involving the theoretical descriptors $\Delta E$ and $\Sigma\Delta q_X$ calculated with HF and DFT methods. We have chosen the same molecular set to make a QSAR analysis based on simple molecular descriptors, and it is shown in Table 1 together with available experimental data in water.

The topological indices (or topological descriptors) are numerical quantities derived from molecular graphs representing molecules. The algorithms transforming the mathematical representations of molecular graphs into topological indices can be divided into three groups: simple, combinatorial and complex. Into the first group one can include algorithms performing simple functions on matrix elements or polynomial coefficients such as counting, multiplying and squaring. The second group includes algorithms performing, additionally, a combinatorial analysis over the elements of graph representations. The algorithms of the third group are based on complicated transformations (diagonalization) of the graph matrix representations.[22] Topological indices have been used so far in the correlation and prediction of a host of molecular properties, such as physicochemical, thermodynamic, biophysical, and physiological properties. To date, more than 1000 different topological descriptors have been put forward in the chemical literature, though only a handful of them have been widely employed for correlative or/and predictive studies.[23]

Despite the existence of such a large number of molecular descriptors, it has been observed that most properties and physical chemistry indices of organic molecules correlate well with the simplest and intuitively sensible topological indices: the atom

**Table 1** Molecular set of carbon acids CHWYZ with their corresponding p$K_a$ values in water.

| Molecule number | W | Y | Z | p$K_a$ |
|---|---|---|---|---|
| 1 | H | H | $CO_2Et$ | 25.6 |
| 2 | H | H | CN | 25.0 |
| 3 | H | H | $SO_2Me$ | 23.0 |
| 4 | H | H | COSEt | 21.0 |
| 5 | H | H | COMe | 19.3 |
| 6 | H | H | COPh | 18.3 |
| 7 | H | Cl | COMe | 16.5 |
| 8 | Cl | Cl | COMe | 15.0 |
| 9 | H | $SO_2Me$ | $SO_2Me$ | 14.0 |
| 10 | H | $CO_2Et$ | $CO_2Et$ | 13.3 |
| 11 | F | F | $NO_2$ | 12.4 |
| 12 | H | CN | CN | 11.2 |
| 13 | Me | COMe | COMe | 11.0 |
| 14 | H | COMe | $CO_2Me$ | 10.7 |
| 15 | H | H | $NO_2$ | 10.2 |
| 16 | Cl | F | $NO_2$ | 10.1 |
| 17 | H | COMe | $SO_2Me$ | 10.0 |
| 18 | H | COMe | COMe | 9.0 |
| 19 | H | Me | $NO_2$ | 8.6 |
| 20 | H | Cl | $NO_2$ | 7.2 |
| 21 | Cl | Cl | $NO_2$ | 6.0 |
| 22 | COMe | COMe | COMe | 5.9 |
| 23 | H | $CONH_2$ | $NO_2$ | 5.2 |
| 24 | H | COMe | $NO_2$ | 5.1 |
| 25 | Cl | $NO_2$ | $NO_2$ | 3.8 |
| 26 | H | $NO_2$ | $NO_2$ | 3.6 |
| 27 | $CH_2CN$ | $NO_2$ | $NO_2$ | 2.3 |
| 28 | $CONH_2$ | $NO_2$ | $NO_2$ | 1.3 |
| 29 | $NO_2$ | $NO_2$ | $NO_2$ | 0.1 |
| 30 | $CO_2Me$ | CN | CN | –2.8 |
| 31 | CN | CN | CN | –5.1 |
| 32 | CN | $NO_2$ | $NO_2$ | –6.2 |

number and the chemical bonds.[24–31] The regression equation adopts the mathematical form

$$\text{Property} = \sum_{X}^{\text{atoms}} a_X A_X + \sum_{X-Y}^{\text{bonds}} b_{X-Y} B_{X-Y} + W \quad (1)$$

where $A_X$ is the number of X atoms, $B_{X-Y}$ is the number of X–Y bonds, $W$ is a constant term and $a_X$, $b_{X-Y}$ are coefficients associated to the X atom and X–Y bond, respectively. $a_X$, $b_{X-Y}$ and $W$ are determined through a multiple regression analysis. Each bond is characterised in the usual chemical sense so that, for example, X–Y and X=Y bonds are considered different.

Rezende[21] correlated p$K_a$ values for 32 carbon acids in water with two and three parameters equations

$$pK_a = a\Delta E + b\sum\Delta q_X + cq_c + d \quad (2)$$

$$pK_a = e\Delta E + f\sum\Delta q_X + g \quad (3)$$

where descriptors $\Delta E$, $\Delta q_c$ and $\sum\Delta q_X$ were calculated with the HF/3-21G and the hybrid DFT B3LYP/6-31G(*d*) methods; *a*, *b*, *c*, *d*, *e*, *f* and *g* are regression coefficients. Thus, in order to look for better p$K_a$ predictions, we have employed the following fitting equations:

$$pK_a = A\Delta E + \sum_{X}^{\text{atoms}} a_X A_X + \sum_{X-Y}^{\text{bonds}} b_{X-Y} B_{X-Y} + W \quad (4)$$

$$pK_a = A\Delta E + A'(\Delta E)^2 + \sum_{X}^{\text{atoms}} a_X A_X + \sum_{X}^{\text{atoms}} a'_X A_X^2 + \sum_{X-Y}^{\text{bonds}} b_{X-Y} B_{X-Y} + \sum_{X-Y}^{\text{bonds}} b'_{X-Y} B_{X-Y}^2 + W', \quad (5)$$

where coefficients have the same meaning as before.

Many correlations need not be linear. In general, one should also test multivariate regression analysis for larger than linear polynomial order and if warranted for other functional dependence.[32] We have computed several fitting polynomial orders and have found that it is not necessary to go beyond the second order to improve significantly the final results.

Calculations were performed resorting to the standard MATHEMATICA® software[33] and the fitting procedure in the multivariate regression analysis was made for first-, second- and third-order polynomials.

Table 2 summarises the results for p$K_a$ estimations for the series of carbon acids CHWYZ studied in this work, together with Rezende's data, experimental values in water and absolute average deviations for both methods: HF/3-21G//HF/3-21G (Method 1) and B3LYP/6-31G(*d*)//HF/3-21G (Method 2), respectively. The W, Y and/or Z groups comprised in addition to alkyl groups, the following substituents: CN, $NO_2$, COMe, $CO_2Me$, $CO_2Et$, COSEt, $CONH_2$, $SO_2Me$, F, Cl and Br.

Statistical results for regression equations are displayed in Table 3.

Before analysing the estimations of the property under study, we must take into account that some of the experimental data are estimated values, the quoted experimental value refers to the gross acid constant, uncorrected for enol content, and significant discrepancies in p$K_a$ values are found in the literature.[21] These features showing a lack of desirable accuracy and uniformity on available experimental data bestow that the average deviations obtained in the present regressions are quite reasonable, considering the uncertainty of some of the data taken from the standard literature.

The fitting equations are better than previous results derived for Rezende's work,[21] so that atom and bond corrections are suitable parameters for this sort of correlations. Besides, second-order regression equations give better predictions than linear ones. However, third-order equations (unpublished results) do not improve significantly predictions. In order to judge properly the merits of the present approximation, one have to consider the relatively wide range of experimental values {–6.2, 25.6} and the large variety of W, X, and Y substituents in the chosen molecular set. Particularly noticeable are the relatively low average absolute deviations obtained from equations (3) and (4) (compare these deviations with similar data for previous results, last row in Table 2). Besides, results derived from the DFT method, which takes into account correlation energy contributions and utilises a larger basis set, is superior to the rather modest estimations obtained from the rather modest HF/3-21G procedure (*i.e.* compare results reported for calculations 3 and 4 with respect to calculations 5 and 6 in Table 2).

**Table 2** Experimental and theoretical estimations of p$K_a$ for carbonic acids.

| Molecule | Exp. | Calc.1[a] | Calc.2[b] | Calc.3[c] | Calc.4[d] | Calc.5[e] | Calc.6[f] |
|---|---|---|---|---|---|---|---|
| 1 | 25.6 | 21.8 | 21.7 | 22.5 | 23.7 | 21.8 | 23.9 |
| 2 | 25.0 | 21.4 | 22.1 | 23.9 | 24.4 | 22.3 | 23.4 |
| 3 | 23.0 | 20.6 | 19.7 | 21.8 | 23.9 | 21.0 | 23.9 |
| 4 | 21.0 | 20.1 | 21.0 | 21.0 | 21.0 | 21.0 | 21.0 |
| 5 | 19.3 | 20.3 | 19.8 | 21.6 | 19.7 | 19.8 | 18.7 |
| 6 | 18.3 | 19.7 | 19.0 | 18.3 | 18.3 | 18.3 | 18.3 |
| 7 | 16.5 | 17.3 | 16.6 | 16.5 | 15.2 | 16.5 | 15.8 |
| 8 | 15.0 | 16.9 | 15.8 | 14.5 | 14.6 | 12.0 | 14.5 |
| 9 | 14.0 | 11.7 | 9.7 | 13.7 | 14.0 | 13.5 | 14.0 |
| 10 | 13.3 | 13.3 | 12.8 | 12.7 | 13.3 | 12.6 | 13.3 |
| 11 | 12.4 | 9.8 | 13.3 | 12.5 | 12.7 | 12.3 | 12.8 |
| 12 | 11.2 | 8.4 | 9.3 | 7.6 | 10.6 | 6.6 | 11.6 |
| 13 | 11.0 | 14.3 | 13.7 | 12.3 | 11.7 | 11.3 | 11.7 |
| 14 | 10.7 | 12.3 | 11.4 | 11.5 | 11.4 | 11.2 | 11.2 |
| 15 | 10.2 | 11.6 | 12.5 | 13.0 | 10.5 | 15.5 | 10.3 |
| 16 | 10.1 | 9.4 | 10.9 | 10.0 | 10.2 | 10.4 | 10.2 |
| 17 | 10.0 | 11.3 | 10.9 | 11.9 | 9.1 | 12.9 | 9.2 |
| 18 | 9.0 | 11.4 | 10.1 | 9.0 | 9.9 | 8.0 | 10.4 |
| 19 | 8.6 | 11.8 | 12.4 | 12.2 | 11.7 | 14.2 | 12.1 |
| 20 | 7.2 | 9.2 | 10.2 | 6.0 | 6.2 | 10.5 | 8.2 |
| 21 | 6.0 | 9.2 | 8.9 | 7.9 | 8.0 | 8.6 | 7.5 |
| 22 | 5.9 | 4.1 | 5.4 | 3.6 | 5.1 | 6.2 | 5.0 |
| 23 | 5.2 | 5.6 | 5.6 | 4.2 | 5.2 | 4.6 | 5.2 |
| 24 | 5.1 | 5.7 | 5.0 | 4.9 | 5.7 | 4.4 | 5.7 |
| 25 | 3.8 | 2.3 | 1.1 | 2.2 | 2.1 | 1.0 | 0.6 |
| 26 | 3.6 | 0.6 | –0.6 | –0.1 | 0.0 | –1.0 | –0.2 |
| 27 | 2.3 | 1.6 | –0.5 | 1.6 | 0.4 | 1.6 | –0.2 |
| 28 | 1.3 | –0.6 | –0.1 | 2.3 | 1.3 | 1.8 | 1.3 |
| 29 | 0.1 | –3.8 | –3.2 | 1.0 | 1.8 | –0.8 | 2.4 |
| 30 | –2.8 | 1.5 | 2.7 | 0.7 | –1.6 | 2.0 | –1.6 |
| 31 | –5.1 | –2.3 | –0.7 | –4.6 | –5.3 | –4.4 | –5.6 |
| 32 | –6.2 | –5.8 | –6.0 | –5.7 | –4.3 | –5.2 | –3.8 |
| Average absolute deviation | — | 1.99 | 1.95 | 1.32 | 0.93 | 1.78 | 1.05 |

[a]Equation (3), Method 1.[21] [b]Equation (4), Method 2.[21] [c]Equation (4), Method 1, this work. [d]Equation (5), Method 1, this work. [e]Equation (4), Method 2, this work. [f]Equation (5), Method 2, this work.

The improvement of pKa estimations when corrections involving atom and bond parameters are in line with previous results.[24–29] This feature is significant, since atom and bond parameters are the simplest ones among the host of molecular and topological descriptors, but they are not widely applied.

Since the training set is rather small, it is necessary to confirm the predictive quality of our models *via* a cross-validation (also known as 'jack-knifing') method.[34] This involves leaving

**Table 3** Statistical parameters corresponding to equations (4) and (5).[a]

| Equation | Correlation coefficient | Standard error | Durbin–Watson Statistic | Average absolute deviation |
|---|---|---|---|---|
| Equation (4) Method 1 | 0.9754 | 2.5888 | 1.7295 | 1.32 |
| Equation (5) Method 1 | 0.9871 | 2.7580 | 1.4791 | 0.93 |
| Equation (4) Method 2 | 0.9510 | 3.6346 | 1.3389 | 1.78 |
| Equation (5) Method 2 | 0.9827 | 3.1891 | 1.0278 | 1.05 |
| Equation (3)[21] | 0.9580 | — | — | 1.99 |
| Equation (4)[21] | 0.9500 | — | — | 1.95 |

[a]Complete results including the coefficients in the fitting equations are available upon request to one of us (E.A.C.).

**Table 4** Cross-validation results.

| Method | *r* (original) | *r* (cross-validated) |
|---|---|---|
| Linear HF | 0.9754 | 0.9754 |
| Linear DFT | 0.9510 | 0.9510 |
| Quadratic HF | 0.9871 | 0.9604 |
| Quadratic DFT | 0.9827 | 0.9189 |

out a number of samples from the data set, calculating the regression model and then predicting values for the samples which were left out. One obvious way to choose these samples is to leave one out at a time (LOO) and this is probably the most commonly used form of cross-validation. Using the LOO method it is possible to calculate a cross-validate *r*, by comparison of predicted values (when the samples were not used to calculate the model) with the measured dependent variable values. Such correlation coefficients will normally be lower than a 'regular' correlation coefficient and are said to be more representative of the performance (in terms of prediction) than can be expected from a regression. Cross-validation can give a measure of the likely performance of a regression model; it can also be used to asses how 'robust' or stable the model is. If the model is generally well fitted to a set of data, then omission of one or more points should not greatly disturb the regression coefficients. The results for the present cross-validation calculations are given in Table 4.

The analysis of this validation data shows that the present model is robust, since both sets of correlation coefficients are similar. In fact, for linear equations they are the same and just minor differences are noted for quadratic fitting polynomials.

## References


1 H. R. Christen and F. Vögtle, *Organische Chemie-Von den Grundlagen zur Forschung*, ed. O. Salle, Verlag, Frankfurt, 1988, vol. 1, p. 419.
2 J. March, *Advanced Organic Chemistry – Reactions, Mechanisms and Structure*, 4th edn., Wiley-Interscience, New York, 1992, ch. 8.
3 R. F. Cookson, *Chem. Rev.*, 1974, **74**, 5.
4 G. Schüürmann, *Quant. Struct-Act. Relat.*, 1996, **15**, 121.
5 R. Stewart, *The Proton: Applications to Organic Chemistry*, ed. H. H. Wasserman, vol. 46 of *Organic Chemistry, A Series of Monographs*, Academic Press, New York, 1985.
6 C. O. Silva, E. C. da Silva and M. A. C. Nascimento, *J. Phys. Chem. A*, 2000, **104**, 2402.
7 W. L. Jorgensen and J. M. Briggs, *J. Am. Chem. Soc.*, 1989, **111**, 4190.
8 G. Schüürmann, *Quantitative Structure-Activity in Environmental Sciences – VII*, eds. F. Chen and G. Schüürmann, SETAC Press, Pensacola, FL, 1997, p. 225.
9 C. J. Cramer and D. G. Truhlar, *Science*, 1992, **256**, 213.
10 C. Lim, D. Bashford and M. Karplus, *J. Phys. Chem.*, 1991, **95**, 5610.
11 W. H. Ricahrdson, C. Peng, D. Bashford, L. Noodleman and D. A. Case, *Int. J. Quantum Chem.*, 1997, **61**, 207.
12 J. Andzelm, C. Kölmel and A. Klamt, *J. Chem. Phys.*, 1995, **103**, 9312.
13 G. Schüürmann, M. Cossi and V. Barone, *J. Phys. Chem. A*, 1998, **102**, 6706.
14 C. O. Silva, E. A. da Silva and M. A. C. Nascimento, *J. Phys. Chem.*, 2000, **104**, 2409.
15 C. L. Russom, S. P. Bradbury, S. J. Brtoderius, D. E. Hammermeister and R. A. Drummond, *Environ. Toxicol. Chem.*, 1997, **16**, 948.
16 H. Könemann and A. Musch, *Toxicology*, 1981, **19**, 223.
17 J. R. Seward and T. W. Schultz, *SAR-QSAR Environ. Res.*, 1999, **10**, 557.
18 M. C. Rezende, *J. Braz. Chem. Soc.*, 2001, **12**, 73.
19 R. R. Contreras, P. Fuentealba, M. Galván and P. Pérez, *Chem. Phys. Lett.*, 1999, **304**, 405.
20 F. D'Souza, M. E. Zandler, G. R. Deviprasad and W. Kutner, *J. Phys. Chem.*, 2000, **104**, 6887.
21 M. C. Rezende, *Tetrahedron*, 2001, **57**, 5923.
22 O. Mekenyan and S. S. Basak, in *Graph Theoretical Approaches to Chemical Reactivity*, eds. D. Bonchev and O. Mekenyan, Kluwer Academic Publishers, Dordrecht, 1994, p. 224.
23 J. Devillers and A. T. Balaban, *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach Science Publishers, The Netherlands, 1999.
24 G. Krenkel, E. A. Castro and A. A. Toropov, *Int. J. Mol. Sci.*, 2001, **2**, 57.
25 P. Duchowicz and E. A. Castro, *J. Indian Chem. Soc.*, 2001, **78**, 192.
26 P. Duchowicz and E. A. Castro, *J. Korean Chem. Soc.*, 2000, **44**, 281.
27 P. Duchowicz and E. A. Castro, *Acta Chem. Slov.*, 2000, **47**, 281.
28 M. Firpo, L. Gavernet and E. A. Castro, *Rom. J. Phys.*, 1999, **44**, 181.
29 M. Firpo, L. Gavernet and E. A. Castro, *Polish J. Chem.*, 1999, **73**, 1041.
30 P. Duchowicz and E. A. Castro, *J. Korean Chem. Soc.*, 1999, **43**, 621.
31 E. A. Castro and M. J. Tueros, *Russ. J. Phys. Chem.*, in press.
32 M. Randic and S. C. Basak, in *Some Aspects of Mathematical Chemistry*, eds. D. K. Sinha, S. C. Basak, R. K. Mohany and J. N. Busa Mallic, Visva-Bharati University Press, Santiniketan, India, 2000, p. 24.
33 S. Wolfram, The MATHEMATICA® Book, 4th edn., Wolfram Media, Cambridge University Press, Cambridge, 1999.
34 D. Livingstone, *Data Analysis for Chemists*, Oxford Science Publications, Oxford, Oxford University Press, Oxford, 1995, p. 134.


*Received: 4th April 2002; Com. 02/1914*