VOLUME 69, NUMBER 2

JOC Articles

JANUARY 23, 2004

Estimation of pK_a Using Quantum Topological Molecular Similarity Descriptors: Application to Carboxylic Acids, Anilines and Phenols

U. A. Chaudry and P. L. A. Popelier*

Department of Chemistry, UMIST, Manchester M60 1QD, England

pla@umist.ac.uk

Received May 31, 2003

The current availability of cheap computer power enables the construction of QSARs from modern ab initio quantum chemical data. Multivariate models for three classes of compounds are developed by means of the quantum topological molecular similarity (QTMS) tool, which incorporates descriptors originating from the "Atoms in Molecules" (AIM) theory. Correlations obtained outperform the Hammett and other traditional parameters. The advantage of QTMS over semiempirical and empirical descriptors is demonstrated by the following r^2/q^2 values: 0.920/0.891 (acids), 0.974/0.953 (anilines), and 0.952/0.884 (phenols).

Introduction

Understanding the effect of substituents on the molecular property pK_a is of interest to the pharmaceutical industry in order to calculate pharmokinetic properties and more generally to the chemical industry in computating the environmental fate or hazard of compounds. Concerning the estimation of pharmokinetic properties, the pK_a can affect protonation states of weak acids and bases at a physiological pH level. This change in protonation state will thus influence the rate at which a compound diffuses across membranes and other physical barriers, such as the blood−brain barrier. The environmental perspective involves degradation of chemicals such as pesticides, which is again determined by their pK_a, as well as the hazard associated with the ability of carboxylic acids to cause skin corrosion in the workplace.

A number of studies reported a correlation between pK_a and various pharmaceutical parameters.[1,2] A number of methods[3] including titrimetry[4] target the calculation of pK_a's. There are clear benefits to a technique that predicts dissociation constants without the need for "wet" experiments. Efficient methods have been implemented in software packages such as ACD Labs,[5] pK_a predictor, and pK_{calc}.[6] However, due to their fragment based approach they are inadequate when fragments present in a molecule under study are absent in the database. In other words, pK_as can only be reliably predicted for compounds very similar to those in the training set. In this study we have chosen to model the pK_a of three well-known classes of compounds: carboxylic acids, anilines and phenols. Previous attempts at modeling the pK_a of carboxylic acids used semiempirical methods.[7] In their work[7] Tehan and co-workers rejected the use of ab initio methods as computationally expensive for drug-sized molecules in a large database of molecules of diverse structure and complexity. Their study incorporated about eight times more data than ours but, as made clear in Section 2, we estimate that a modest Linux PC cluster of about a dozen nodes would require no more than a few days of computing time. Although the processing time on modern day computers is not a significant barrier, the reliability and true predictive capability of ab initio methods off-sets these computational demands.

Tehan and co-workers produced a quantitative structure−activity relationship (QSAR) for a set of a set of 141 aliphatic carboxylic acids yielding an r^2 value of 0.80 and a q^2 value of 0.80. Gruber and Buss[8] developed a three-term equation, with an r^2 of 0.80, using HOMO energies. Citra[9] also reported a three-term equation with an r^2 of 0.84 for 56 acids. Quantum mechanical analyses assisting the prediction of the dissociation constant for

* To whom correspondence should be addressed. Tel: +44-161-2004511. Fax: +44-161-2004559.

(1) Jayasekhar, P.; Kasture, A. V. *Bull. Chim. Far.* **1999**, *138*, 489−492.
(2) Jones, T.; Taylor, G. *Proc.−Eur. Congr. Biopharm. Pharmacokinet.* **1987**, *2*, 181−190.
(3) Perrin, D. D.; Dempsey, B.; Serjean, E. P. *pK_a Prediction for Organic Acids and Bases*; Chapman and Hall: London, 1981.
(4) Albert, A.; Serjeant, E. P. *The Determination of Ionisation Constants*, 2nd ed.; 1971.
(5) *ACD/Labs version 3*; ACD Labs: Toronto, ON, Canada.
(6) *pKalc*; Comudrug International: San Francisco, CA.
(7) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Montana, J. G.; Manallack, D. T.; Gancia, E. *Quant. Struct.-Act. Relat.* **2002**, *21*, 457−471.
(8) Gruber, C.; Buss, V. *Chemosphere* **1989**, *19*, 1595−1609.
(9) Citra, M. J. *Chemosphere* **1999**, *38*, 191−206.

10.1021/jo0347415 CCC: $27.50
Published on Web 09/06/2003

aliphatic carboxylic acids used self-similarity measures producing a model with a r^2 of 0.915 for a set of 10 acids[10] but no measure of predictability (cross-validated r^2 or q^2) was reported.

Moving across to 3D QSAR, comparative molecular field analysis (CoMFA)[11] has also been applied to a small set of acids,[12] introducing issues of molecular alignment. Impressive results were obtained by Adam[13] who, as in our study, incorporated the theory of "Atoms in Molecules" (AIM),[14−19] using the energy of the dissociating proton in solution as the only descriptor and obtained an r^2 of 0.983 for a set of 19 acids (no q^2 was stated). As demonstrated at the end of this section, QTMS has a wider applicability than the prediction of pK_a's. Moreover, we show there is no need for AIM's rather compute intensive atomic properties.

The carboxylic acid data set of the present study was acquired from Eriksson et al.[20] Recently, Gross et al.[21] investigated the applicability of ab initio (quantum chemical) parameters as alternatives to the Hammett constant in modeling the pK_a of anilines and phenols. A set of 36 anilines was used in constructing possible one-parameter regression descriptors for pK_a that included the natural charge of amino nitrogen, relative proton-transfer enthalpy, minimum molecular surface local ionization energies, and molecular electrostatic potential minima; results were comparable with the Hammett constants. This ab inito approach was extended to model the pK_a for a set of 19 phenols via the examination of several quantum chemical parameters: the natural charges on the phenolic hydrogen and the phenoxide oxygen, the phenoxide HOMO energy, and the relative proton-transfer energy. This study showed that E_{HOMO} is superior to the Hammett constants in describing the substituent-induced pK_a effects. The aforementioned study[8] of Gruber and Buss also incorporated a set of 99 phenols yielding a commendable r^2 of 0.94 (without reporting q^2).

The rapid growth of QSAR studies illustrates the progress in this area of modern (bio)chemistry and demonstrates the abundance of data in an age of mass information. Coupled with increased computer processing power, the development of algorithms delivered the status quo where routine generation of a plethora of descriptors requires a matter of minutes. Such descriptors, as in the present study, are often of quantum chemical origin and provide more accurate descriptions of electronic effects than empirical methods. Moreover, in most cases, they do not necessarily suffer from the approximate nature of the method or neglect of solvation effects assuming relative values are used.[22] This leads to a situation where there are many "flavors" of descriptors. For example, bond rotation energy barrier, bond angle, and natural atomic charge coupled with empirical descriptors such as the Hammett σ and Taft σ^* constants attempt to model one property, such as electronic effects. Furthermore, such linear free energy relationships (LFER) have been criticized for lacking solid scientific basis in their empirical approach.[23] This state of affairs leads us to quantum topological molecular similarity (QTMS),[24] which is based on the increasingly popular[25,26] theory of AIM. This theory, which is deeply rooted in quantum mechanics,[27] can be used to extract chemical insight from modern ab initio wave functions.

AIM has stimulated the use of new descriptors in chemometric analyses, such as in StrucQT,[28,29] the prediction of hydrogen bond basicity[30] and hydrogen bond donor capacity[31] or of physicochemical properties of amino acids, polycyclic aromatic hydrocarbons and the opiates,[32] and Oripavine PEO, enkephalins, and morphine.[33] In our group, QTMS delivered excellent QSARs of environmental, biological, and industrial interest. Examples include the prediction of toxicity of polychlorinated dibenzo-*p*-dioxins (PCDDs),[34] of toxicity and biodegradability of para-substituted phenols, and ^{13}C NMR chemical shifts in para- and meta-substituted benzonitriles,[35] of antibacterial activity of nitrofuran derivatives, of antitumor activity of (*E*)-1-phenyl-but-3-en-ones,[36] of mutagenicity,[37] of furanones and triazenes, of the corticosteroid binding binding of the classical steroid dataset,[38] of hydrolysis rate constants of polar esters,[39] and of σ_p, σ_m, σ_I, and σ_p^0 parameters of mono-[40] and polysubstituted benzoic acids, phenylacetic acids, and bicyclocarboxylic acids.[34] In summary, QTMS is so reliable in predicting activities and properties dominated by electronic effects, such that when it fails for a given data set one can safely conclude

(10) Ponec, R.; Lluis, A.; Carbo-Dorca, R. *J. Phys. Org. Chem.* **1999**, *12*, 447−454.
(11) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(12) Kim, K. H.; Martin, Y. C. *J. Org. Chem.* **1991**, *56*, 2723−2729.
(13) Adam, K. R. *J. Phys. Chem. A* **2002**, *106*, 11963−11972.
(14) Bader, R. F. W. *Atoms in Molecules. A Quantum Theory*; Oxford University Press: Oxford, 1990.
(15) Bader, R. F. W. *Can. J. Chem.* **1998**, *76*, 973−988.
(16) Bader, R. F. W. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: Chichester, 1998; pp 64−86.
(17) Bader, R. F. W. *Acc. Chem. Res.* **1985**, *18*, 9−15.
(18) Bader, R. F. W. *Chem. Rev.* **1991**, *91*, 893−928.
(19) Popelier, P. L. A. *Atoms in Molecules. An Introduction.*; Pearson Education: London, 2000.
(20) Eriksson, L.; Berglind, R.; Sjostrom, M. *Chemom. Intell. Lab. Syst.* **1994**, *23*, 235−245.
(21) Gross, K. C.; Seybold, P. G. *J. Org. Chem.* **2001**, *66*, 6919−6925.
(22) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. *Chem. Rev.* **1996**, *96*, 1027−1043.
(23) Bodor, N.; Gabanyi, Z.; Wong, C. K. *J. Am. Chem. Soc.* **1989**, *111*, 3783−3786.
(24) O'Brien, S. E.; Popelier, P. L. A. *J. Chem. Inf. Comput.Sc.* **2001**, *41*, 764−775.
(25) Popelier, P. L. A.; Aicken, F. M.; O'Brien, S. E. In *Chemical Modelling: Applications and Theory*; Hinchliffe, A., Ed.; Royal Society of Chemistry Specialist Periodical Report; Royal Society of Chemistry: London, 2000; Vol. 1, Chapter 3, pp 143−198.
(26) Popelier, P. L. A.; Smith, P. J. In *Chemical Modelling: Applications and Theory*; Hinchliffe, A., Ed.; Royal Society of Chemistry Specialist Periodical Report; Royal Society of Chemistry: London, 2002; Vol. 2, Chapter 8, pp 391−448.
(27) Bader, R. F. W. *Pure Appl. Chem.* **1988**, *60*, 145−155.
(28) Alsberg, B. K.; Marchand-Geneste, N.; King, R. D. *Chemo. Intell. Lab. Syst.* **2000**, *54*, 75−91.
(29) Alsberg, B. K.; Marchand-Geneste, N.; King, R. D. *Anal. Chim. Acta* **2001**, *446*, 3−13.
(30) Platts, J. A. *Phys. Chem. Chem. Phys.* **2000**, *2*, 3115−3120.
(31) Platts, J. A. *Phys. Chem. Chem. Phys.* **2000**, *2*, 973−980.
(32) Matta, C. F. *J. Comput.Chem.* **2003**, *24*, 453−462.
(33) Matta, C. F. *J. Phys. Chem. A* **2001**, *105*, 11088−11101.
(34) Popelier, P. L. A.; Chaudry, U.; Smith, P. J. *J. Chem. Soc., Perkin Trans. 2* **2002**, 1231−1237.
(35) O'Brien, S. E.; Popelier, P. L. A. In *ECCOMAS*; Barcelona, Spain, 2000.
(36) O'Brien, S. E.; Popelier, P. L. A. *J. Chem. Soc., Perkin Trans. 2* **2002**, 478−483.
(37) Popelier, P. L. A.; Chaudry, U. A.; Smith, P. J. *Internet Electron. J. Mol. Des* **2003**, accepted.
(38) Smith, P. J.; Popelier, P. L. A. *submitted* **2003**.
(39) Chaudry, U. A.; Popelier, P. L. A. *J. Phys. Chem. A* **2003**, *107*, 4578−4582.
(40) Popelier, P. L. A. *J. Phys. Chem. A* **1999**, *103*, 2883−2890.

that electronic effects are not important. Note that we use the expression "electronic effects" in the strict QSAR sense, that is, to distinguish them from steric effects and log *P*. The term "electronic effects" should not be confused with the quantum mechanical electron density.

2. Method and Computational Details

Details on QTMS can be found in ref 41, but here we reiterate salient features. QTMS consists of three stages: the generation of geometry-optimized bond lengths and wave functions, computation of quantum topological properties, and a chemometric analysis.

Using the program GAUSSIAN98,[42] geometries were optimized and single-point energies were obtained at four levels of theory, denoted by A, B, C, and E, for consistency with our previous and future publications. Level A corresponds to the semiempirical model AM1,[43] which yields reasonable bond lengths for nonesoteric (hence already parametrized) molecules but fails to provide an electron density that can be analyzed topologically. All higher levels (B, C, and E) used in this work generate the "single point" wave function at the optimized geometry. Levels B and C correspond to HF/3-21G(d)[44] and HF/6-31G(d), respectively. The most expensive level, level E, corresponds to B3LYP/6-311+G(2d,p), where electron correlation is modeled by a well-known hybrid density functional[45] of applied density functional theory (DFT).[46]

A local version of the program MORPHY98[47] delivers the four topological properties that we use to describe the bonds in each molecule. Loosely speaking, AIM defines a bond critical point (BCP) as a point at which the gradient of the electron density vanishes, lying roughly between two bonded nuclei. Properties evaluated at the BCP characterize the corresponding bond and are selected as topological descriptors. In this study, the properties are the electron density, ρ, the Laplacian of the electron density, $\nabla^2\rho$, the ellipticity ϵ,[19] and a type of local kinetic energy density, *K*. Of course, when computed for a Kohn−Sham-based density functional (e.g., B3LYP), *K* refers only to a noninteracting reference system. Although not strictly a BCP property, we added the equilibrium bond length r_e to the set of four descriptors (for each bond). Thus, a QTMS descriptor matrix was constructed for each of the acid, aniline, and phenol data sets with 40[compounds] × 3[bonds], 36 × 14, and 19 × 13 entries, respectively. BCP properties and equilibrium bond lengths were used separately to produce models, as made clear in Tables 1−6 in the Results and Discussion. Although this is not a fundamental restriction of QTMS, we work here with a common skeleton. This is the largest fragment common to all molecules allowing for different atomic numbers. For example, the anilines are described by 14 bonds because they all contain the two bonds of NH_2, the six aromatic CC bonds, and six C−X bonds where X is either hydrogen, nitrogen, or a substituent Y. If Y is Cl in one substituted aniline and F in another, the C−Cl and C−F bonds correspond to each other in the descriptor matrixes of the respective molecules. Clearly, the atomic numbers of Cl and F are different, which is allowed by the common skeleton requirement. However, if the substituent were methoxy, OCH_3, then the bonds of the methyl group would be unmatched in a comparison with, say, fluoroaniline, and hence, the methyl group would not be included in the descriptor matrix.

Models were constructed using the partial least squares (PLS)[48] method, as implemented in the program SIMCA-P.[49] Note that for level *A* we can only consider the optimized bond lengths as descriptors. The PLS technique has been designed to handle thousands of descriptors (so-called *X* variables), which can be noisy and highly correlated (virtually collinear). Although we do not involve as many *X* variables as a typical CoMFA analysis, we benefit from the advantages PLS offers. This is so because the generally nonlinear dependence[50] of the descriptors ρ, $\nabla^2\rho$, ϵ, and *K* on r_e could become linear. Also, we could interpret the small deviations between the values of the five descriptors generated at the current levels of theory and of the exact wave function as "noise".

The quality of the PLS regression is assessed by the correlation coefficient r^2 and the cross-validated correlation coefficient,[51] q^2, based on leaving out one-seventh of the data. A data randomization test guards against "correlation by chance" by monitoring the deterioration of the model (measured by r^2 and q^2) as the *Y* variables are randomly permuted. We adopted the default SIMCA-P cutoff values beyond which the model ceases to be valid. We used the so-called variable importance in the projection (VIP)[52] to detect the "active center" of the compound. Descriptors (or *X* variables) with a VIP value smaller than one can be rejected as unimportant, whereas those with the highest VIP values constitute the "active center". An optional step is to compress the number of descriptor variables for each bond by principal component analysis (PCA) using the program SPSS.[53] PLS is carried out again, this time on the extracted PCs rather than on the "raw" variables. The introduction of PCs allows us to isolate descriptors corresponding to a specific bond in the common-skeleton across the data set. In doing so we are not only correlating descriptors across an entire molecule to the pK_a, but we also illustrate the usefulness of QTMS in isolating the fragment(s) of a molecule that are important to the physical property. Gaining "added value" from the model is considered an important outcome in building QSAR.[54] However, we prefer not to provide explicit QSAR equations, which customarily link the observed activity to the descriptors in 2D QSAR studies. Indeed, we do not generate global but rather local descriptors (such as in CoMFA) proportional to the number of bonds in the common skeleton. This means for example that such an equation for the anilines would contain 56 terms.

Finally, it is important to give an idea of the CPU times required for a typical QTMS analysis, realizing that the generation of optimized wave function is the rate-limiting step.

(41) O'Brien, S. E.; Popelier, P. L. A. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 764−775.

(42) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*, revision A.7; Gaussian, Inc.: Pittsburgh, PA, 1998.

(43) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(44) Foresman, J. B.; Frisch, A. *Exploring Chemistry with Electronic Structure Methods*, 2nd ed.; Gaussian Inc.: Pittsburgh, PA, 1996.

(45) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(46) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: Weinheim, 2000.

(47) MORPHY98, a program written by P. L. A. Popelier with a contribution from R. G. A. Bone, UMIST, Manchester, England, EU 1998. http://morphy.ch.umist.ac.uk/.

(48) Wold, S.; Sjostrom, M.; Eriksson, L. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: Chichester, 1998; pp 2006−2021.

(49) UMETRICS; info@umetrics.com: www.umetrics.com, 1998.

(50) O'Brien, S. E.; Popelier, P. L. A. *Can. J. Chem.* **1999**, *77*, 28−36.

(51) Livingstone, L. *Data Analysis for Chemists*, 1st ed.; Oxford University Press: Oxford, 1995.

(52) UMETRICS, A. SIMCA-P 8.0 User Guide and Tutorial Umea, 1999.

(53) SPSS Inc., version 10.0.7, Chicago, IL, 2000; http://www.spss.com.

(54) Cronin, M. T. D.; Schultz, T. W. *THEOCHEM* **2003**, *622*, 39−51.

TABLE 1. Substituted Carboxylic Acids and Their Observed and Predicted pK_a Values Obtained at Level B Using BCP Properties for the Highest Ranked Bond C_1-O_3 According to the VIP Plot

N	compd	exptl pK_a	calcd pK_a	ρ_{C-O}	$\nabla^2\rho_{C-O}$	ϵ_{C-O}	K_{C-O}
1	acetic acid	4.76	4.18	0.4023	−0.3520	0.0300	0.6013
2	bromoacetic acid	2.90	3.04	0.4074	−0.3125	0.0368	0.6148
3	chloroacetic acid	2.82	2.80	0.4080	−0.3090	0.0367	0.6164
4	dichloroacetic acid	1.26	1.57	0.4054	−0.3403	0.0441	0.6104
5	trichloroacetic acid	0.63	0.45	0.4102	−0.3013	0.0488	0.6238
6	trifluoroacetic acid	0.23	0.61	0.4103	−0.2966	0.0505	0.6209
7	acrylic acid	4.25	4.65	0.4000	−0.4090	0.0186	0.5937
8	formic acid	3.55	3.18	0.4029	−0.2609	0.0296	0.6049
9	mercaptoacetic acid	3.67	3.43	0.4040	−0.3450	0.0319	0.6054
10	propionic acid	4.87	4.50	0.4011	−0.3418	0.0206	0.5978
11	2-chloropropionic acid	2.88	2.92	0.4017	−0.3487	0.0263	0.6000
12	3-chloropropionic acid	4.00	4.03	0.4011	−0.3628	0.0231	0.5976
13	methacrylic acid	4.66	4.79	0.3985	−0.4209	0.0185	0.5903
14	butyric acid	4.82	4.56	0.4008	−0.3471	0.0208	0.5969
15	vinylacetic acid	4.34	4.36	0.4011	−0.3495	0.0188	0.5975
16	crotonic acid	4.70	5.08	0.3988	−0.4302	0.0154	0.5902
17	isocrotonic acid	4.41	4.36	0.3994	−0.4018	0.0297	0.5930
18	isobutyric acid	4.86	4.83	0.4000	−0.3456	0.0190	0.5949
19	valeric acid	4.86	4.56	0.4009	−0.3473	0.0208	0.5969
20	isovaleric acid	4.78	4.76	0.3999	−0.3612	0.0179	0.5941
21	pivalic acid	5.05	5.02	0.3991	−0.3515	0.0182	0.5928
22	cyanoacetic acid	2.45	2.48	0.4077	−0.3280	0.0404	0.6160
23	2-bromobutyric acid	2.55	3.14	0.4007	−0.3563	0.0242	0.5971
24	glycolic acid	3.83	3.34	0.4077	−0.2921	0.0345	0.6158
25	lactic acid	3.86	4.83	0.3971	−0.3649	0.0026	0.5868
26	2-hydroxybutyric acid	3.68	3.38	0.3984	−0.3855	0.0090	0.5891
27	oxalic acid	1.27	0.65	0.4076	−0.2969	0.0425	0.6148
28	malonic acid	2.83	3.08	0.4075	−0.3317	0.0419	0.6146
29	succinic acid	4.20	4.35	0.4005	−0.3624	0.0206	0.5957
30	maleic acid	1.94	1.65	0.3970	−0.4553	0.0170	0.5878
31	glutaric acid	4.35	4.30	0.4017	−0.3455	0.0233	0.5991
32	2-chlorobutyric acid	2.84	2.98	0.4012	−0.3556	0.0255	0.5987
33	3-chlorobutyric acid	4.06	3.36	0.4004	−0.3711	0.0283	0.5966
34	4-chlorobutyric acid	4.52	4.13	0.4018	−0.3481	0.0239	0.5994
35	nitroacetic acid	1.68	1.95	0.4116	−0.3206	0.0527	0.6256
36	difluoroacetic acid	1.24	1.89	0.4054	−0.3179	0.0372	0.6072
37	fluoroacetic acid	2.59	3.12	0.4075	−0.3055	0.0333	0.6148
38	2-bromopropionic acid	2.97	3.09	0.4011	−0.3496	0.0248	0.5984
39	3-bromopropionic acid	3.99	4.14	0.4009	−0.3615	0.0222	0.5970
40	4-bromobutyric acid	4.58	4.18	0.4017	−0.3477	0.0237	0.5992

On a moderately priced PC (dual AMD Athlon MP1900+, 1 GB DDR RAM) the maximum time to compute a wave function at HF/3-21G(d) level (level B) for compounds of the acid data set was 15 CPU minutes, this level of theory producing the best results. Our dataset of 40 carboxylic acids thus required less than 10 CPU hours, or 1 h on a small Linux cluster of 10 PCs. The most expensive set however, that of the 36 anilines at level E, cost about 15 CPU hours on a cluster on 10 PCs.

3. Results and Discussion

3.1. Carboxylic Acids. Table 1 shows the measured pK_a values for the 40 carboxylic acids obtained from a set devised by Eriksson.[55] Iodoacetic acid was not included since basis sets for iodine were not readily available. The deprotonation of an aliphatic carboxylic acid (R is an alkyl group) can be summarized by the following scheme:

$$R-C(=O)-O-H \leftrightarrow R-C^{+}(-O^{-})-O-H \leftrightarrow R-C(-O^{-})=O^{+}-H \rightleftharpoons R-C(=O)-O^{-} + H^{+} \leftrightarrow R-C(-O^{-})=O + H^{+}$$

(55) Eriksson, L. A.; Berglind, R.; Sjostrom, M. *Chemom. Intell. Lab. Syst.* **1994**, *23*, 235−245.

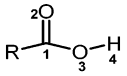


FIGURE 1. Numbering scheme of the common skeleton of the carboxylic acids.

The numbering scheme used to identify the location of descriptors important to the PLS model is shown in Figure 1. We are now in a position to map information obtained from one set of bond descriptors to the corresponding set in another molecule. Since only the (O═C)−O−H fragment is common to all molecules we restrict the descriptors to the three bonds it contains.

In the PLS data matrix we now have 40 observations (i.e., measured pK_a values) and 12 descriptors, four descriptors obtained for each of the three bonds in the carboxyl group at levels B, C, and E. A summary of the PLS analyses at the four levels of theory is shown in Table 2. Two QSAR models were developed at each level, using BCP descriptors in one and equilibrium bond lengths in the other. The BCP models always outperform the bond length model in terms of r^2 but not always in terms of predictivity or q^2. The analysis at the more computationally expensive level *E* does not provide significantly better results, but it is encouraging that descriptors obtained at level *B* perform better than those

TABLE 2. Summary of PLS Analyses for the Carboxylic Acids

level	descriptors	LV[a]	r^2	q^2
A	bond lengths	2	0.787	0.598
B	bond lengths	2	0.885	0.883
	BCP properties	2	0.920	0.891
C	bond lengths	2	0.879	0.871
	BCP properties	2	0.918	0.819
E	bond lengths	2	0.853	0.839
	BCP properties	2	0.891	0.839

[a] Number of latent variables.

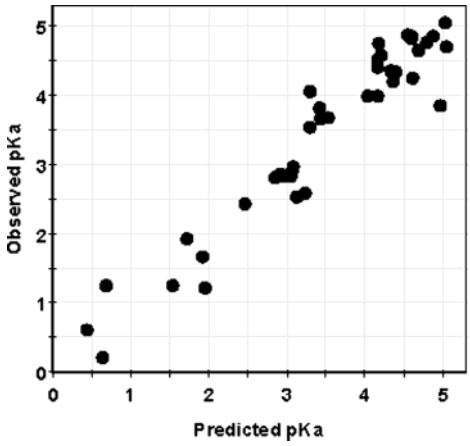


FIGURE 2. Observed versus predicted pK_a values for the set of 40 carboxylic acids at level B with BCP properties, using the carboxylic acid group $(O_2{=}C_1){-}O_3{-}H_4$ as the common skeleton.

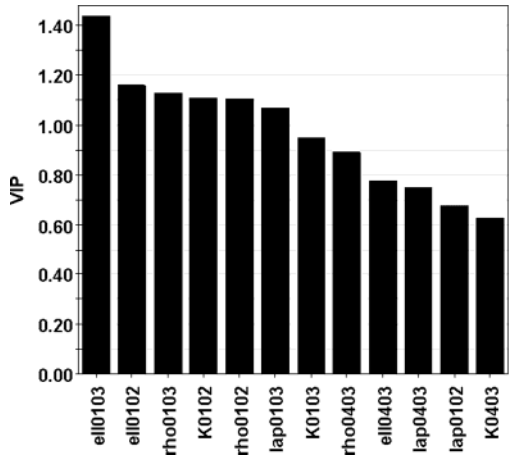


FIGURE 3. VIP plot for the complete set of 40 carboxylic acids at level B with BCP properties using the carboxyl group $(O_2{=}C_1){-}O_3{-}H_4$ as the common skeleton.

of AM1 or other semiempirical analyses.[9] In that work, Citra obtained a three-term equation for aliphatic carboxylic acids yielding an r^2 value of 0.84 and one for aromatic carboxylic acids yielding an r^2 value of 0.89. The strongest model was obtained at level B with the use of BCP properties as descriptors yielding two components with an r^2 value of 0.920 and a q^2 value of 0.891. Figure 2 illustrates the quality of this prediction.

We clearly see from the VIP plot in Figure 3 that the ellipticity of the $C_1{-}O_3$ bond is the most important descriptor in building the PLS model followed by the ellipticity of $C_1{=}O_2$. The ellipticity describes the ovality of a bond and reflects its π character,[19] which means our top two descriptors are consistent with the chemistry that occurs during the dissociation process. The resonance structures in the scheme above show that the fluctuations of π character at $C_1{=}O_2$ and $C_1{-}O_3$ would have a significant presence during the deprotonation process and as such are reflected in the PLS model.

The results indicate that QTMS can be employed to estimate the pK_a of aliphatic carboxylic acids with good accuracy compared to experimental values or values predicted by LFER and semiempirical methods. The value of our descriptors is apparent when QTMS results are compared with other more established electronic descriptors such as E_{HOMO}, E_{LUMO}, electronegativity, and partial charges.

In their study of the same carboxylic acid set, Eriksson et al.[55] reported a principal component analysis (PCA) that showed a 2D plot of the first and second loading vector for a set of nine classical descriptors, such as melting point, molecular weight, density, etc. In that loading plot, the descriptors log P and pK_a lay on the extremities of the two orthogonal loading vectors (or axes). From this, one can assume that the two descriptors explain a completely dissimilar aspect of the chemistry behind the observed activity. Second, as QTMS succeeds in modeling pK_a on the basis of its reliability in capturing electronic effects, it inevitably cannot explain observed activities that involve a log P effect. This latter assertion is consistent with the many other studies our group has carried out over the years and indicates the position of QTMS within the larger arena that is QSAR. Given the success[41] of QTMS in predicting the Hammett σ of polysubstituted benzoic acids, future work on the prediction of pK_a values of polysubstituted aliphatic carboxylic acids may prove equally successful.

3.2. Anilines. Table 3 shows the measured pK_a values for the 36 anilines obtained from a recent study by Gross et al.[21] The pK_a here refers to the conjugate acids, but it is used as a measure of the amine's basicity since pK_a + pK_b = pK_w, where K_w is the ionization constant of water. The protonation of aniline is outlined by the following equation:

All atoms provided of a numerical label were included in the common skeleton, which encompasses 14 bonds. A summary of the PLS analyses at the four levels of theory are shown in Table 4. Again, two QSAR models were constructed at each level of theory, using BCP descriptors in one and equilibrium bond lengths in the other. We obtain progressively better correlation statistics with increasing level, and ultimately the best model is obtained when electron correlation is incorporated, at level E. Figure 4 shows predicted versus observed pK_a values for this best model, maintaining quality over almost six pK_a units. The benefit of using descriptors at DFT level more than outweighs the computationally expense required. Also, for anilines, BCP models always outperform the bond length models, both in terms of r^2

TABLE 3. Aniline Substituents and Their Observed and Predicted pK_a Values Obtained at Level E Using BCP Properties for the Highest Ranked Bond C_4-N_1 According to the PC VIP Plot

N	substituent	exptl pK_a	calcd pK_a	ρ_{C-N}	$\nabla^2\rho_{C-N}$	ϵ_{C-N}	K_{C-N}
1	H	4.58	4.52	0.2987	−0.8987	0.0983	0.3577
2	*m*-amino	4.88	5.02	0.2985	−0.8995	0.0946	0.3578
3	*m*-bromo	3.51	3.38	0.3017	−0.9155	0.1031	0.3643
4	*m*-chloro	3.34	3.39	0.3017	−0.9153	0.1031	0.3642
5	*m*-cyano	2.76	3.11	0.3037	−0.9262	0.1061	0.3690
6	*m*-fluoro	3.59	3.37	0.3017	−0.9166	0.1033	0.3645
7	*m*-hydroxy	4.17	3.97	0.3004	−0.9106	0.0983	0.3621
8	*m*-methoxy	4.2	4.17	0.2994	−0.9042	0.0974	0.3592
9	*m*-methyl	4.69	4.74	0.2983	−0.8977	0.0968	0.3573
10	*m*-nitro	2.5	2.37	0.3043	−0.9290	0.1099	0.3697
11	p-amino	6.08	5.66	0.2922	−0.8520	0.1051	0.3377
12	*p*-bromo	3.91	3.93	0.3006	−0.9079	0.1035	0.3613
13	*p*-chloro	3.98	3.96	0.3001	−0.9045	0.1039	0.3598
14	*p*-cyano	1.74	2.02	0.3076	−0.9565	0.1046	0.3857
15	*p*-fluoro	4.65	4.47	0.2972	−0.8834	0.1049	0.3501
16	*p*-hydroxy	5.5	5.32	0.2939	−0.8620	0.1053	0.3414
17	*p*-methoxy	5.29	5.52	0.2939	−0.8630	0.1042	0.3420
18	*p*-methyl	5.12	4.99	0.2968	−0.8858	0.0994	0.3521
19	*p*-nitro	1.02	1.21	0.3106	−0.9767	0.1047	0.3974
20	3,4-dimethyl	5.17	5.19	0.2966	−0.8855	0.0976	0.3520
21	3-amino-4-hyroxy	5.7	5.91	0.2938	−0.8638	0.1008	0.3416
22	3-bromo-4-methoxy	4.08	4.55	0.2962	−0.8749	0.1078	0.3463
23	3-bromo-4-methyl	3.98	3.81	0.2997	−0.9016	0.1037	0.3581
24	3-chloro-4-methyl	4.05	3.87	0.2997	−0.9013	0.1037	0.3579
25	3-methyl-4-nitro	1.5	1.51	0.3101	−0.9743	0.1011	0.3965
26	4-chloro-3-nitro	1.9	2.11	0.3053	−0.9334	0.1136	0.3718
27	4-methyl-3-nitro	2.96	2.87	0.3026	−0.9176	0.1095	0.3646
28	3,5-dibromo	2.34	2.08	0.3044	−0.9298	0.1073	0.3703
29	3,5-dimethoxy	3.82	4.08	0.2991	−0.9021	0.0965	0.3577
30	3,5-dimethyl	4.91	5.01	0.2979	−0.8954	0.0950	0.3563
31	3-chloro-5-methoxy	3.1	3.32	0.3013	−0.9127	0.1018	0.3625
32	3-methoxy-5-nitro	2.11	1.92	0.3045	−0.9302	0.1086	0.3694
33	3,5-dibromo-4-hydroxy	3.2	3.27	0.2985	−0.8847	0.1125	0.3498
34	3,5-dibromo-4-methoxy	2.98	2.68	0.3017	−0.9095	0.1090	0.3608
35	3,5-dibromo-4-methyl	2.87	2.71	0.3023	−0.9155	0.1069	0.3638
36	3,5-dimethyl-4-nitro	2.59	2.74	0.3069	−0.9537	0.0995	0.3846

TABLE 4. Summary of PLS Analyses for the Anilines

level	descriptor	LV	r^2	q^2
A	bond lengths	2	0.857	0.758
B	bond lengths	2	0.917	0.862
	BCP properties	1	0.940	0.915
C	bond lengths	2	0.916	0.882
	BCP properties	3	0.968	0.925
E	bond lengths	2	0.954	0.921
	BCP properties	2	0.974	0.953

and q^2, without exceptions. Again, we improve on the results of Gross et al.[21] In their study, Hammett constants yielded an r^2 value of 0.940 and an r^2 of 0.949 when using the minimum ionization energy. No q^2 was reported.

From the VIP plot shown in Figure 5, we see that the most important variables in explaining the activity are those that belong to the NH_2 group, that is N_1-H_2 and N_1-H_3. This is expected since this functional group is the center of activity. Although the descriptor "Lap0401" (Figure 5) appears with a lower VIP value than "Lap0301" and "Lap0201", we believe that a plausible link with aniline basicity can be suggested.

The Laplacian, $\nabla^2\rho$, gauges charge concentration and depletion and operates as a simple measure for covalency versus ionicity. If $\nabla^2\rho < 0$ at a BCP, the bond is said to represent a shared interaction, and if $\nabla^2\rho > 0$ it is a closed-shell interaction. Ionic bonds reside under the latter category while covalent ones under the former. In our QSAR we cannot use the Laplacian to distinguish both extremes of bond types since Table 3 lists only

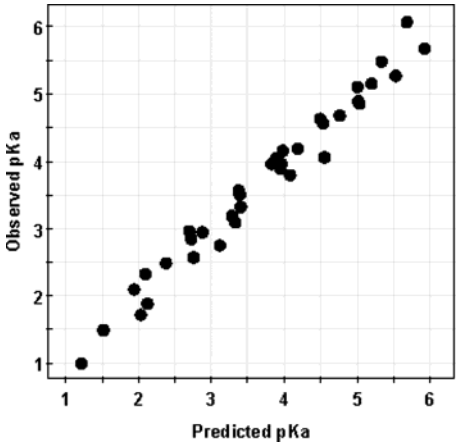


FIGURE 4. Observed versus predicted pK_a for the set of 36 anilines at level E with BCP properties including all 14 bonds of the common skeleton.

negative values of $\nabla^2\rho$ at the BCP of the relevant C_4-N_1 bond in all anilines. However, in the current work, we can say that the more positive the Laplacian (i.e., the smaller its absolute value) the more the covalent bond has a tendency to segregate electron density toward the bonded nuclei, a feature pointing toward ionic bonds in the limit. How can we use this interpretation of the Laplacian in the context of basicity?

It is well-known that the partially pyramidal amino group can conjugate with the phenyl π-system. This delocalization renders the lone-pair less available for protonation. Hence, conjugation of nitrogen with the

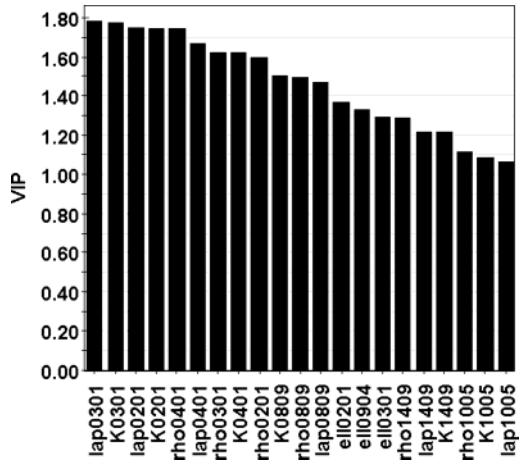


FIGURE 5. VIP plot for the complete set of 36 anilines at level E with BCP properties.

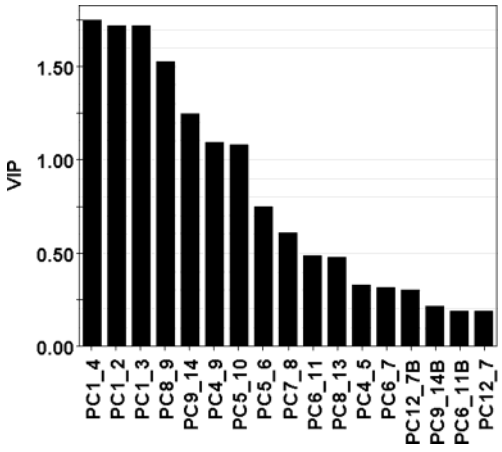


FIGURE 6. VIP plot of Principal Components (composed of BCPs and bond lengths) for each bond for the complete set of 36 anilines at level E (numbers refer to bonds in common skeleton).

TABLE 5. Phenol Substituents and Their Observed and Predicted pK_a Values Obtained at Level E Using BCP Properties and QCT Descriptors for the Highest Ranked Bond O_2-H_3 According to the PC VIP Plot

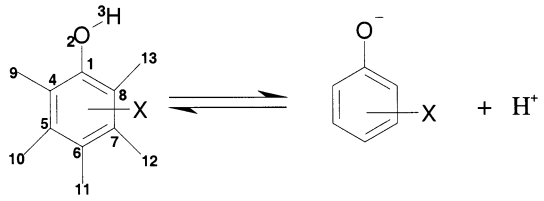
N	compd	pK_a	ρ_{O-H}	$\nabla^2\rho_{O-H}$	ϵ_{O-H}	K_{O-H}
1	H	9.98	0.3623	−2.4006	0.0225	0.6741
2	*m*-amino	9.87	0.3628	−2.4003	0.0226	0.6744
3	*m*-bromo	9.03	0.3618	−2.4078	0.0220	0.6750
4	*m*-chloro	9.02	0.3618	−2.4070	0.0220	0.6749
5	*m*-cyano	8.61	0.3613	−2.4115	0.0217	0.6754
6	*m*-fluoro	9.28	0.3619	−2.4068	0.0221	0.6749
7	*m*-hydroxy	9.44	0.3625	−2.4003	0.0225	0.6742
8	*m*-methoxy	9.65	0.3625	−2.3974	0.0226	0.6736
9	*m*-methyl	10.08	0.3623	−2.3986	0.0226	0.6737
10	*m*-nitro	8.4	0.3612	−2.4161	0.0215	0.6762
11	*p*-amino	10.3	0.3631	−2.3980	0.0235	0.6744
12	*p*-bromo	9.36	0.3620	−2.4061	0.0222	0.6749
13	*p*-chloro	9.38	0.3620	−2.4053	0.0223	0.6748
14	*p*-cyano	7.95	0.3613	−2.4137	0.0211	0.6757
15	*p*-fluoro	9.95	0.3624	−2.4051	0.0227	0.6751
16	*p*-hydroxy	9.96	0.3626	−2.3976	0.0232	0.6739
17	*p*-methoxy	10.21	0.3625	−2.3958	0.0232	0.6734
18	*p*-methyl	10.14	0.3624	−2.3983	0.0228	0.6737
19	*p*-nitro	7.15	0.3612	−2.4182	0.0207	0.6764

phenyl ring decreases the basicity of the amino group in anilines. The C_4-N_1 bond, connecting the amino group with the phenyl group, is an appropriate monitor of the degree of delocalization between the amino group and the phenyl ring. Now we take an extreme entry of Table 3 that illustrates how the Laplacian can be invoked to measure this delocalization. *p*-Aminoaniline has a high pK_a value (6.08) and, hence, a lower pK_b value compared with aniline, indicating higher basic character of the NH_2 group. Therefore, we deduce the lone pair to be more "free" to accept the proton, since it conjugates less with the phenyl ring. This would infer that density is more segregated onto the carbon (C_4) and the nitrogen (N_1), which in turn infers a more positive Laplacian. Indeed, the value of −0.8520 listed in Table 3 is one of the most positive (i.e., least negative or smallest absolute value) values found. The same reasoning can be applied to *p*-nitroaniline, at the other extreme, having a lower pK_a (1.02). Here, the very negative Laplacian value indicates substantial sharing in the bond, hence conjugation from phenyl across to the amino group, which reduces the basicity (since the lone pair is less available to accept the proton).

When we invoke the second chemometric step within QTMS, which involves a reduction in dimensionality of descriptor space via PCA, we find that the active region of the aniline molecules is indeed recovered, as shown in Figure 6. The decision to introduce the PCA here, and later for the phenol set, is to retrieve information on important bonds within the molecule that are responsible for the observed property. Unlike with the acids where the common skeleton was small and centered on the functional group (COOH), we now have a larger common skeleton. This illustrates the usefulness of QTMS in locating key regions in a larger molecule. In other words, even if the common skeleton contains 14 bonds, only three bonds predominantly explain the activity.

Principal components in Figure 6 are ranked according to their importance. The three highest ranked PCs correspond to the "active-region" within the anilines (i.e., the $-NH_2$ fragment) and exhibit a clear lead over those PCs that represent the remainder of the molecule. As with the work carried out by Gross et al., the set of anilines was split into para and meta-substituted compounds. Such a division is not necessary in QTMS but if carried out we obtain equally robust results, that is, an r^2 value of 0.970 and q^2 value of 0.838 for monosubstituted anilines and $r^2 = 0.983$ and $q^2 = 0.927$ for the para-substituted anilines.

3.3. Phenols. Table 5 shows the measured pK_a values for the 19 phenols taken from ref 21.[21] The numbering scheme used to identify the location of descriptors in the deprotonation of phenol follows:



A summary of the PLS analyses at the four levels is shown in Table 6, three levels (B, C, and E) being treated with and without BCP properties, as before. Only for levels C and E does the addition of BCP properties

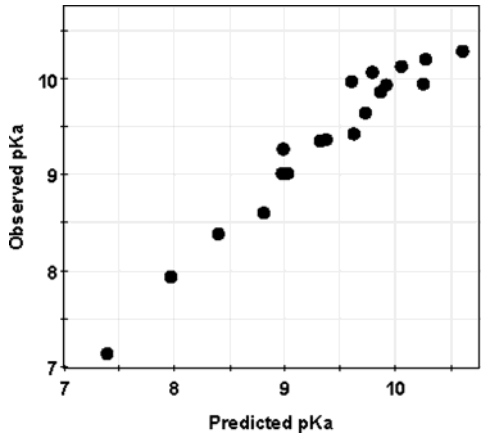


FIGURE 7. Observed versus predicted rate constant (pK_a) for the set of 19 phenols at level E with BCP properties of all 13 bonds of the common skeleton.

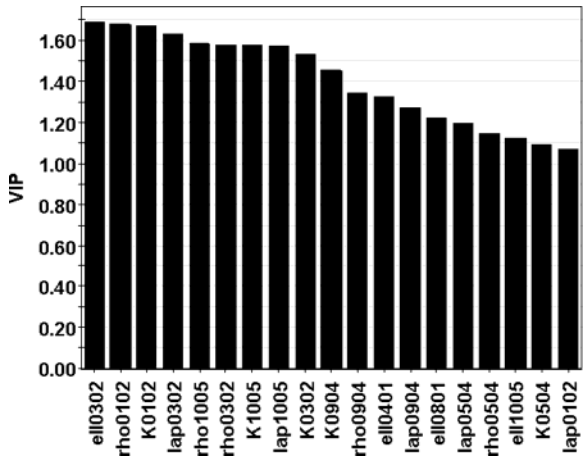


FIGURE 8. VIP plot for the set of 19 phenols at level E with BCP properties.

improve the correlation statistics. The best model is obtained when electron correlation is incorporated, i.e., level E with BCP properties. This again justifies the use of a higher level of theory, which posesas the rate-limiting step in the QTMS analysis. Figure 7 compares the observed and predicted pK_a over more than 3 pK_a units.

Comparing these results with those obtained by Gross et al.[21] when using the Hammett constants and other quantum chemical parameters we find QTMS producing stronger QSARs. They found that the Hammett constants produced a correlation coefficient r^2 of 0.816 for the pK_a values of these phenols, which was outperformed by all their quantum chemical descriptors, Their best model yielded an r^2 of 0.911 when using the phenoxide oxygen charge.

Figure 8 shows the VIP plot, for the best model, of the original (i.e., uncompressed) variables. As with the anilines and carboxylic acids, we see that the VIP plot corresponds to the chemistry one would expect in the deprotonation of a phenol. The O_2-H_3 bond is readily broken upon deprotonation, and the resulting anion is stabilized through resonance. The top four descriptors in the VIP plot illustrate this, thus highlighting the active center in this reaction, the added benefit of QTMS over other 2-D QSAR methods whereby we can "magnify" activities of important bonds.

TABLE 6. Summary of PLS Analyses for the Phenols

level	descriptor	LV	r^2	q^2
A	bond lengths	2	0.890	0.825
B	bond lengths	2	0.938	0.831
	BCP properties	2	0.909	0.826
C	bond lengths	1	0.911	0.856
	BCP properties	1	0.930	0.863
E	bond lengths	2	0.894	0.832
	BCP properties	2	0.952	0.884

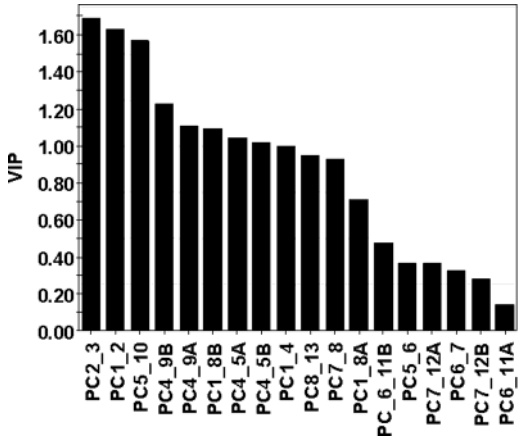


FIGURE 9. VIP plot of principal components (composed of BCPs and bond lengths) for each bond of the set of 19 phenols at level E (numbers refer to bond in common skeleton).

Figure 9 shows the VIP plot of the compressed descriptors or PCs. Reassuringly, the PCs of the C_1-O_2 and O_2-H_3 bonds tower above the others, with the exception of the C_5-H_{10}.

A split of the phenol set into para and meta-substituted compounds further illustrated the strength of QTMS descriptors by yielding models of $r^2 = 0.970$ and $q^2 = 0.838$ (meta) and $r^2 = 0.983$ and $q^2 = 0.927$ (para).

Conclusion

We showed that QTMS delivers strong QSAR models for a set of aliphatic carboxylic acids, anilines, and phenols in estimating their pK_a values. These results improve on previous attempts at modeling pK_a by others and highlight the increasing use of quantum chemically derived descriptors over empirical parameters such as the Hammett and semiempirically obtained descriptors. In analyzing the common skeleton, we take advantage of QTMS's capability to "highlight" the important bonds that are responsible for the observed property. We do not bias the analysis in only picking certain bonds to build our QSAR but allow the data to provide us with a chemical insight. This can be significant for systems where modes of action are not known. QTMS may be applied to compounds that are not amenable to analysis via Hammett parameters such as when values are not obtainable for the fragments under observation. Even though electron correlation was necessary in some cases to produce good models, such as with anilines and phenols, a QTMS analysis of large (industrially relevant) molecules is perfectly feasible from a computational point of view. It is gratifying to realize that this computational investment injects genuine quantum chemical information into a QSAR and hence increases the probability that

the activity is explained for reasons more closely linked to exact solutions of the Schrödinger equation.

Acknowledgment. Gratitude is expressed to EPSRC and ICI for sponsoring this work. We express our gratitude to Dr. S. Liem for his invaluable help in designing the cover of this issue.

Supporting Information Available: Lists of total energies (au) and atomic coordinates (bohr) of substituted carboxylic acids, anilines, and phenols at the level of theory with the best predictive power. This material is available free of charge via the Internet at http://pubs.acs.org.

JO0347415