pK_a Prediction from "Quantum Chemical Topology" Descriptors

A. P. Harding, D. C. Wedge, and P. L. A. Popelier*

Manchester Interdisciplinary Biocentre (MIB), 131 Princess Street, Manchester M1 7DN, Great Britain, and School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, Great Britain

Received May 12, 2009

Knowing the pK_a of a compound gives insight into many properties relevant to many industries, in particular the pharmaceutical industry during drug development processes. In light of this, we have used the theory of Quantum Chemical Topology (QCT), to provide *ab initio* descriptors that are able to accurately predict pK_a values for 228 carboxylic acids. This Quantum Topological Molecular Similarity (QTMS) study involved the comparison of 5 increasingly more expensive levels of theory to conclude that HF/6-31G(d) and B3LYP/6-311+G(2d,p) provided an accurate representation of the compounds studies. We created global and subset models for the carboxylic acids using Partial Least Square (PLS), Support Vector Machines (SVM), and Radial Basis Function Neural Networks (RBFNN). The models were extensively validated using 4-, 7-, and 10-fold cross-validation, with the validation sets selected based on systematic and random sampling. HF/6-31G(d) in conjunction with SVM provided the best statistics when taking into account the large increase in CPU time required to optimize the geometries at the B3LYP/6-311+G(2d,p) level. The SVM models provided an average q^2 value of 0.886 and an RMSE value of 0.293 for all the carboxylic acids, a q^2 of 0.825 and RMSE of 0.378 for the ortho-substituted acids, a q^2 of 0.923 and RMSE of 0.112 for the para- and meta-substituted acids, and a q^2 of 0.906 and RMSE of 0.268 for the aliphatic acids. Our method compares favorably to ACD/Laboratories, VCCLAB, SPARC, and ChemAxon's pK_a prediction software based of the RMSE calculated by the leave-one-out method.

1. INTRODUCTION

The pK_a of a compound is an important property in both life sciences and chemistry since the propensity of a compound to donate or accept a proton is fundamental to understanding chemical and biological processes. As the pK_a value of a molecule also determines the amount of protonated and deprotonated species at a specific pH, for example at physiological pH, knowing the pK_a of a molecule gives insight into pharmacokinetic properties. The latter includes the rate at which a molecule will diffuse across membranes and other physiological barriers, such as the blood brain barrier. More often than not, phospholipid membranes easily absorb neutral molecules, while ionized molecules tend to remain in the plasma or the gut[1] before being excreted. Many biological systems also use proton-transfer reactions to communicate between the intra- and extracellular media, and the rate of the proton-transfer reaction depends, in-part, on the pK_a values of the species involved.[2]

There are a number of well established experimental techniques,[3] such as spectroscopy, potentiometry, conductimetry, competitive reactions, and titrimetry that can accurately determine pK_a values for a molecule. However, experimental determination of acidity of a specific part of a large biological molecule such as a protein is not a straightforward task[4] and is often associated with large uncertainties in the results. The benefits of a technique that accurately predicts the dissociation constant without the need for "wet" experiments are clear. The chemical industry, in particular the pharmaceutical and agrochemical sectors, screen thousands of compounds during the discovery process for many properties simultaneously, including the dissociation constant. More efficient techniques are required because of the logistics of measuring the pK_a values of these compounds are coupled with known problems associated with certain techniques. For example, high-throughput UV absorption measurements can often miss groups not in close proximity to a UV-chromophore.[5] It is clear that *in silico* techniques can increase efficiency and reduce costs at a challenging time for the pharmaceutical industry as they can be consulted before expensive synthetic work is undertaken.

The need for accurate pK_a prediction has led to a large number of publications tackling the issue. While the precise implementation of methodology varies, the publications generally fall into three categories: (i) predictive models, using a range of descriptors and statistical methods, (ii) *ab initio* quantum chemical methods based on different thermodynamic cycles, and (iii) fragment-based approaches.

The first category of models relies on choosing the right descriptors to model the pK_a of a particular data set. Structural, physiochemical, topological, geometrical, constitutional, electrostatic, quantum-chemical, and thermodynamic descriptors have all been used to predict pK_a with varying success. Gruber and Buss[6] performed semiempirical calculations on some 190 phenols and carboxylic acids. They used multilinear regression, with descriptors such as heats of formation, molecular orbital energies, and charge densities, to produce a three-term equation for the benzoic acid derivatives and a four-term equation for the aliphatic acids. Citra[7] criticized this linear free energy relationship based approach for lacking scientific basis and vast use of correc-

* Corresponding author e-mail: pla@manchester.ac.uk.

10.1021/ci900172h CCC: $40.75
Published on Web 07/24/2009

tion factors favoring quantum mechanical methods. A three-term equation for benzoic acids was reported using a method similar to that of Gruber and Buss. Gross and Seybold[8] rejected the use of semiempirical methods, instead using density functional theory, after a set of survey calculations demonstrated it performed significantly better for the descriptors they employed. After studying phenols they found that atomic charge, and the HOMO and LUMO energies, correlated with pK_a. After rejecting the use of *ab initio* methods as too computationally demanding, Tehan[1] and co-workers produced QSARs for aliphatic acids and substituted benzoic acids. Xing and Glen[9] fashioned a novel structure tree representation of atoms to align molecules. Atom types and group types that were of biological interest were used in conjunction with PLS to produce a QSAR for a large set of acids and bases. Following this procedure, Xing[10] and co-workers reduced the number of atom types and increased the group types used, also noting that splitting the data set improved their results. The approach introduced by Xing has been taken up by numerous authors.[11,12] For example, Jelfs et al.[5] utilized the tree fingerprint method to develop a prediction method using semiempirical chemical properties, such as partial charge and electrophilic superdelocalizability of atoms undergoing protonation or deprotonation, to produce an online prediction Web-tool for pK_a prediction at Novartis. In the second category of methods, *ab initio* quantum chemical methods based on different thermodynamic cycles[13,14] have started to receive more attention.[15,16]

The method involves calculating the standard change in Gibbs energy related to the dissociation of a proton from the compound under study in water. The method utilizes gas phase and aqueous phase *ab initio* calculations but, depending on the thermodynamic cycle used, involves at least four separate geometry optimizations for each prediction. Out of all the commercial packages available for pK_a prediction, Schrödinger's Jaguar[17] application is the only tool that employs this method. Although these methods do not rely upon a predictive model, the results can be associated with large uncertainties because different thermodynamic cycles, levels of theory and models of solvation produce different pK_a values. The Jaguar package, for example, uses empirical correction terms, where calculated values are fitted to experimental values stored in a database, to repair deficiencies in both the *ab initio* calculations and solvation models. Namazian[4,18−21] and co-workers used an equation[2,22] that relates the standard change of Gibbs energy to the pK_a of various carboxylic acids.

In the third and final category of methods, the fragment-based approach supported by a back end database is implemented in the LogD suite of ACD/Laboratories.[23] It uses an internal database of around 16,000 structures with over 31,000 experimental values. Problems associated with such methods occur when fragments present in a molecule are absent from the database.

Quantum Topological molecular similarity (QTMS), pioneered in our group,[24,25] is a novel approach to solving QSAR/QSPR problems using descriptors defined by Quantum Chemical Topology (QCT).[26,27] QTMS uses properties evaluated at special points in space as a means of summarizing the electronic structure of a set of molecules and correlating them with some given activity. These special points are called bond critical points (BCPs) and are briefly reviewed in the next Section. Over the last several years QTMS methods have produced excellent results of relevance to biological,[28,29] medicinal,[30−34] environmental,[35] industrial, and physical organic chemistry.[36−39] In all cases the models had excellent validation statistics and also provide information about the active center or region of the compounds thought to be important to the activity. Here we evaluate QTMS descriptors in an extension to previous QTMS studies, which have shown good predictive ability for pK_a.[40,41]

Regarding pK_a, Adam[42] obtained impressive results by incorporating QCT into his study. Using transferability between similar molecules, an idea at the origin of QCT,[43] he obtained an r^2 for aliphatic and benzoic acids greater than 0.84, in most cases, using the energy of the dissociating proton in solution as the only descriptor. On the other hand, QTMS descriptors have been successfully employed for carboxylic acids and anilines[41] and phenols in aqueous[44] and polar solvents.[40]

From the aforementioned QTMS publications it has become clear that QTMS descriptors are effective at capturing electronic effects. Therefore we deduce that when QTMS fails electronic effects are not as important to the predicted property or activity as, for example, solubility or steric effects. This was the case for the predictive QSAR models of phenols[32] where logP was included to describe the importance of hydrophobicity for hepatocyte toxicity prediction in conjunction with QTMS descriptors capturing the important electronic effects. Another example is the QTMS study[45] on a remarkable and unusual set of *ortho* alkyl substituted phenols, known for their cytotoxicity and previously investigated by the Hansch group.[46] The QTMS results do not support their proposal that a steric factor is important in the determination of the cytotoxicity. In fact, the QTMS results suggest no steric contribution whatsoever.

In the past, we tended to utilize the interpretative ability of PLS at the chemometric stage. The program SIMCA-P[47] contains a module that calculates the relative importance of each descriptor to the model; these are known as the Variable Importance in the Projection (VIP). The most important descriptors for prediction related to individual bonds can be examined and the so-called active center of the compounds under study detected. While this is a compelling feature of QTMS, we concentrate here on highly predictive models for pK_a estimation. For this reason we have used other statistical methods, such as support vector machines (SVM) and radial basis function neural networks (RBFNN), which yield models that are not so interpretable but possibly more accurate. We have also extensively cross-validated our models and moved away from relying on SIMCA-P for validation. The 228 carboxylic acid compounds included in this study constitute the largest set of compounds investigated with QTMS. This large set of diverse carboxylic acids facilitates the aim of extending the domain of applicability and producing more predictive models.

2. METHODS AND COMPUTATIONAL DETAILS

2.1. Data Set. We seek to predict the pK_a of molecules or fragments with pharmaceutical relevance. We therefore used the data set of Tehan et al.,[1] who had previously applied a variety of filters in order to remove non-druglike molecules from their data set. We selected carboxylic acids as com-

pounds of interest because we want to apply the QTMS methodology to a set of large and diverse compounds further developing previous applications of QTMS. After iodine containing molecules were removed, since the basis sets were not readily available, our data set contained 228 carboxylic acids with a pK_a range of 0.51−6.20. This included 44 meta- and para-substituted benzoic acids, 50 ortho-substituted benzoic acids, and 134 aliphatic carboxylic acids. The observed pK_a values for all 228 carboxylic acids are listed in Table S1 of the Supporting Information.

2.2. Bond Critical Points. Broadly speaking, a BCP forms when the gradient of the electron density, ρ, vanishes ($\nabla\rho=\mathbf{0}$) at a point in space between two bonded nuclei. The electron density at the BCP, ρ_b, is the first QTMS descriptor, from which bond orders can be derived.[48] It also displays strong correlations with bond energy.[49] At the BCP, the Hessian of the electron density has two negative eigenvalues ($\lambda_1 < \lambda_2 < 0$) and one positive one ($\lambda_3 > 0$). The eigenvector associated with λ_3 is tangential to the bond, and so λ_3 describes the curvature along the bond. The eigenvectors corresponding to λ_1 and λ_2 are orthogonal to the bond, and so λ_1 and λ_2 describe the curvatures perpendicular to the bond. The sum of the three eigenvalues is the Laplacian of the electron density, denoted by $\nabla^2\rho$, which gives a measure of the local charge concentration or depletion at the BCP. If the negative eigenvalues λ_1 and λ_2 dominate, then an accumulation of charge takes place in the plane perpendicular to the bond. This is common for shared interactions, such as covalent bonds. This results in a negative value for the Laplacian. If the positive eigenvalue λ_3 dominates, then the electron density accumulates along the bond toward the nuclei. This is common for closed-shell interactions such as ionic, hydrogen, and van der Waals bonds. This results in a positive value for the Laplacian. The ellipticity of ρ, denoted by ε and defined as $\varepsilon = \lambda_1/\lambda_2 - 1$, provides a further useful property associated with BCPs. If ρ protrudes more in one of two directions perpendicular to a bond, then an oval pattern appears, such as in a pure double bond. This ovality is measured by the ellipticity. Single bonds are characterized by λ_1 and λ_2 being nearly identical, and hence ε is near zero. Since $\nabla^2\rho$ and ε are calculated using the three eigenvalues, we have chosen to exclude $\nabla^2\rho$ and ε in this study and instead work with λ_1, λ_2, and λ_3. Two types of kinetic energy, denoted by K(**r**) and the more 'classical' kinetic energy G(**r**)[50] can also be used as QTMS descriptors. We also include the equilibrium bond length, R_e, as a QTMS descriptor. In this QTMS study we therefore have seven QTMS descriptors to represent each bond used, namely ρ_b, λ_1, λ_2, λ_3, K(**r**), G(**r**), and R_e.

2.3. Data Generation. The data generation process for QTMS can be found in previous publications.[24] In short, an approximation of the geometry of each molecule is provided by MOLDEN.[51] Using the program GAUSSIAN03[52] geometries were optimized successively at five different levels of theory: AM1, HF/3-21G(d), HF/6-31G(d), B3LYP/6-31+G(d,p), and B3LYP/6-311+G(2d,p). The levels are denoted by letters *A, B, C, D*, and *E,* respectively, for consistency with previous publications. The optimization steps are by far the most computationally expensive stage in the QTMS process. The electronic wave function calculated by GAUSSIAN is then passed on to a local version of the program MORPHY98,[53] which locates the BCPs using a robust algorithm.[54] This yields a property vector for each BCP, providing a discrete "quantum fingerprint" of molecules, if all BCPs appearing in a molecule are combined. In this study the electron density (ρ), the three eigenvalues of the Hessian of the electron density (λ_1, λ_2, and λ_3), the two types of kinetic energy (K(**r**) and G(**r**)), and the equilibrium bond lengths (R_e) have been used to describe each BCP. The numbering scheme given to atoms in the common skeleton of each molecule is shown in Figure 1. This allows the location of descriptors important to the statistical analysis to be identified in each molecule. In addition, the scheme allows BCPs in one molecule to be mapped onto corresponding BCPs in other molecules. This is not a fundamental requirement of the method, and the constraint of a common skeleton can be relaxed.[30] We end up with five data matrices (one for each level of theory) consisting of 228 observations (i.e., measured pK_a values) and 21 descriptors for each observation, that is, 7 descriptors obtained for each of the three bonds in the common skeleton.

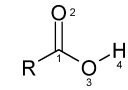

Figure 1. Numbering scheme of the common skeleton of the carboxylic acids.

2.4. Machine Learning and Chemometric Analysis. *2.4.1. Partial Least Squares.* Partial least-squares (PLS)[55] analysis was carried out to fit the BCP descriptor variables to the experimental pK_a values. The program SIMCA-P[47] was used along with its predefined criterion for determining the significant number of Latent Variables (LVs) to appear in the PLS equation. If the value of q^2 of the newly constructed LV is less than 0.097, then that LV is considered not significant, and no more LVs are computed; the PLS regression is then deemed complete.

Two statistics provided by SIMCA-P are the squared correlation coefficient, r^2, and the cross-validated r^2, denoted by q^2. The generated q^2 is based on 'leave-one-seventh' of the data out rather than the popular 'leave-one-out', which is not recommended because of its known pitfalls.[55,56] First, the initial model is constructed, involving all descriptors at each level of theory. Then the VIP plots are examined because they offer a concise summary of the importance of each of the descriptors. Descriptor variables with a VIP value less than unity are considered unimportant to the model and hence discarded.[55] The models are then reconstructed with the reduced set of variables. We also built models using just R_e to demonstrate that using BCP properties as descriptors provides more information than R_e alone. As well as producing global models for the carboxylic acids, we repeated the PLS analysis after splitting the data set into aliphatic and benzoic acids, which were further split into meta/para-substituted and ortho-substituted sets. Altogether there are three subsets.

2.4.2. Support Vector Machines. Support vector machines, originally proposed by Vapnik[57] to solve pattern recognition problems,[58] were extended in 1996 for linear and nonlinear support vector regression (SVR).[59] Due to their remarkable generalization performance, SVM have found numerous applications in chemistry including drug design, QSAR/QSPR, chemometrics, sensors, chemical engineering, and text mining.[60]

Table 1. Summary of the Initial PLS Analysis To Determine the Level of Theory To Use for the Comparison of Learning Methods[b]

		all acids			ortho subset			para/meta subset			aliphatic subset		
level	descriptors	LV[a]	r^2	q^2	LV[a]	r^2	q^2	LV[a]	r^2	q^2	LV[a]	r^2	q^2
A	bond lengths	2	0.554	0.537	1	0.733	0.717	1	0.768	0.756	2	0.795	0.767
B	bond lengths	2	0.506	0.484	1	0.741	0.717	1	0.664	0.573	2	0.728	0.693
	BCP properties	1	0.600	0.595	1	0.761	0.716	1	0.683	0.601	2	0.771	0.731
	BCP properties (VIP > 1)	3	0.638	0.616	1	0.757	0.718	1	0.708	0.651	2	0.726	0.713
C	bond lengths	2	0.593	0.581	1	0.770	0.754	1	**0.783**	**0.758**	2	0.729	0.696
	BCP properties	2	0.660	0.637	1	0.770	0.747	1	0.779	0.750	3	0.805	0.744
	BCP properties (VIP > 1)	2	0.661	0.648	1	0.783	0.763	1	0.779	0.752	4	**0.823**	**0.791**
D	bond lengths	2	0.448	0.431	1	0.745	0.728	1	0.720	0.672	2	0.704	0.674
	BCP properties	6	**0.783**	**0.742**	1	0.769	0.752	1	0.759	0.724	5	0.815	0.768
	BCP properties (VIP > 1)	7	0.696	0.646	1	0.794	0.775	1	0.772	0.742	2	0.788	0.758
E	bond lengths	2	0.462	0.446	1	0.767	0.757	1	0.737	0.700	2	0.700	0.672
	BCP properties	6	0.766	0.731	1	0.767	0.754	1	0.749	0.710	5	0.813	0.763
	BCP properties (VIP > 1)	7	0.728	0.693	1	**0.792**	**0.782**	1	0.750	0.712	4	0.808	0.778

[a] Number of latent variables. [b] The bold text highlights the best models for each set (i.e all, ortho, para/meta and aliphatic) based on the highest q^2.

As with other multivariant statistical methods, the performance of SVM for regression depends on the combination of several parameters. We employed a Gaussian Radial Basis Function kernel for SVR because of its effectiveness and speed in the training process.[61] This function contains an extra parameter γ (a constant) that controls the amplitude of the Gaussian function, thereby controlling the generalization ability of the SVM to some extent. The user-prescribed parameters (i.e γ, ε of the ε-insensitive loss function and capacity parameter C) were chosen based on the lowest root mean squared error (RMSE) of the training data. The SVR programs were written by former group member C. X. Xue in an R-file, based on a script written in the R language for SVM, which utilized the e1071 package.[62] The scripts were compiled using the R 2.5.1 compiler[63] and run on a Pentium D PC with 1GB RAM.

2.4.3. Radial Basis Function Neural Networks (RBFNN). The subject of neural networks is covered in depth in the work by Haykin[64] and Gurney.[65] The theory of RBFNNs as applied to QSARs has been extensively described in the paper of Yao et al.[66] The training procedure involved the forward subset selection routine, which selected the centers for the RBF one at a time and adjusted the weights between the hidden layer and the output layer after the addition of each center, using a least-squares[67] solution. One third of the training set was randomly selected and 'held back' as test data, and training was terminated when the error on the test data showed no further improvement. RBFNN training was carried out using a range of RBF widths between 0.2 and 5.0, and the width yielding the lowest error on the test set was selected.

2.4.4. Comparison of the Methods. To compare these three machine learning methods we used r-fold cross-validation (CV), where the data sets were divided in 4, 7 (as implemented in SIMCA-P), and 10 CV groups. The division of the data sets was carried out using systematic sampling where the compounds were ordered according to their pK_a values and assigned to a group accordingly. For example, for 4-fold CV the first compound was grouped with the fifth, ninth, thirteenth, etc. In the random sampling method, the compounds were ordered by random numbers and divided into groups of different sizes depending on the r-fold CV being used (e.g., for 10-fold CV for the 50 ortho-substituted benzoic acids the first five compounds were group one, the next five group 2, etc.). Each CV group was excluded in turn so that each compound was excluded from the training data exactly once, and the RMSE of prediction and q^2 were calculated for the CV group according to the following equations

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_{obs,i} - \hat{y}_{pred,i})^2}{n}} \quad (1)$$

$$\mathrm{q2} = 1 - \frac{\sum_{i=1}^{n} (y_{obs,i} - \hat{y}_{pred,i})^2}{\sum_{i=1}^{n} (y_{obs,i} - \bar{y})^2} \quad (2)$$

where n is the number of observations in the CV set, $y_{obs,i}$ is the observed pK_a value for the molecule i in the CV set, $\hat{y}_{pred,i}$ is the associated predicted pK_a, and $\bar{y}$ is the mean pK_a value of the entire data set. The RMSE and q^2 values obtained for each CV group were then averaged with the values obtained from the corresponding groups to give final values for the global and subset models, for each CV group size, selected based on systematic and random sampling (see Table S2 in the Supporting Information for a complete table of the averaged RMSE and q^2 values). The method that produced the lowest RMSE in conjunction with the highest q^2 was considered to be the most accurate.

3. RESULTS AND DISCUSSION

3.1. Choice of Level of Theory. Table 1 shows a summary of the initial PLS analysis at the five different levels of theory for the data set and subsets. At each level, three different models were generated: a bond length only model, a model with bond length involving all BCP descriptors (amounting to 21, which derives from 3 bonds and 7 descriptors per bond), and a model including only those descriptors from the bond length and BCP descriptor model having a VIP score greater than one. At level *A*, only a bond length model can be generated. This is because semiempirical (AM1) wave functions do not contain core densities, which corrupts the topology by affecting the position or even appearance of BCPs.[25,68] The results in Table 1 are rather

disappointing compared to the previous QTMS study of carboxylic acids,[41] where an r^2 and q^2 of 0.920 and 0.891 were obtained, respectively, and compared to the results in the literature discussed in the introduction earlier. Outlier detection was undertaken using the subset models as they were more easily distinguishable from correct predictions than in the models containing all the carboxylic acids. In all the para- and meta-substituted models, compound 37 (3,4-diaminobenzoic acid) was always an outlier. This compound was one of three zwitterions in the subset including compound 15 (3-aminobenzoic acid) and 21 (4-aminobenzoic acid). These compounds were predicted reasonably well (observed pK_a values of 4.74 and 4.85 and predicted pK_a values of 4.14 and 4.82, respectively) according to the best para- and meta-substituted model marked in bold in Table 1. However, compound 37 was predicted consistently poorly (observed pK_a of 3.49 and a predicted pK_a of 4.62 according to the same model). The zwitterions were all modeled in their neutral form, which sufficed for the monoamino-benzoic acids but was not appropriate for diaminobenzoic acid, which was predicted to have a larger pK_a value due to the fact that the increased stability in its zwitterionic form is not encapsulated in the BCP descriptors. As in this work, Tehan[1] and co-workers struggled to model this effect and had few problems with the monoaminobenzoic acids. In line with Tehan we omitted 37 as an outlier. In the ortho-substituted models, two compounds were identified as outliers and removed from the models, namely 72 (2,6-dihydroxybenzoic acid) and 79 (2-hydroxy-3,5-dinitrobenzoic acid) whose pK_a values were predicted to be 2.60 and 1.86 compared to observed values of 1.05 and 0.70, respectively. The hydroxyl group at the ortho position(s) in these compounds could be held responsible for their overprediction because internal hydrogen bonding in the anionic form could increase the stability of the ion therefore decreasing their pK_a values. This effect is not encapsulated in the BCP descriptors and therefore absent from the models. The issue with this reasoning is that there are a further 11 compounds in the subset that are hydroxyl-substituted at the 2 position and they are predicted well.

Four further compounds were identified as outliers from the aliphatic carboxylic acid models and removed. They were 124 (4-[(4-chloro-2-methylphenyl)oxy]butanoic acid), 150 (cyanoacetic acid), 155 (9-hydroxy-9H-fluorene-9-carboxylic acid/flurenol), and 228 (4-(cyclopropylcarbonyl)-3,5-dioxo-cyclohexanecarboxylic acid)). Tehan[1] and co-workers brought into question the reliability of the observed pK_a value of compound 124, with which we concur. Compound 155 (Figure 2a) was excluded from their model on the basis that the proximity of the carboxyl group to the two aromatic rings and the presence of an α-hydroxy group make its pK_a difficult to predict. Alternatively, the observed pK_a is incorrect. We can confirm that it is the *observed* pK_a value that is the most likely cause for discrepancy. Our best model predicts the pK_a to be 2.87, while the experimental value given by Tehan (our source data set) is 1.09. A different source[69] gives the observed pK_a value of 2.96, which is close to our predicted value and the value predicted by ACD/Laboratories[23] as 3.04. Furthermore, there is a similar structure (104, hydroxy-(diphenyl)acetic acid) in the data set that has an observed pK_a of 3.05 (Figure 2b). A literature value for the observed pK_a values of these compounds (155 and 104) with their

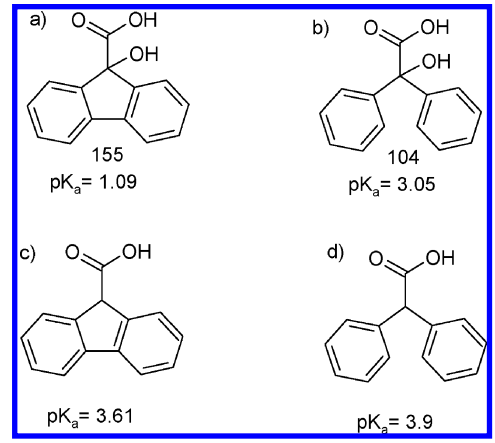

Figure 2. Structures and observed pK_a values for compounds 155 and 104 and their analogues fluorene-9-carboylic acid and diphenylacetic acid.

respective hydroxyl group removed was found to be 3.61[70] for 155 (Figure 2c) and 3.9[71] for 104 (Figure 2d), a difference of only 0.3 log units. The difference of 1.96 log units (=3.05−1.09) between compounds 155 and 104 generated by the addition of a hydroxyl group at the same position in each is unlikely when considering the difference is 0.3 log units between the analogous compounds, thus further supporting a wrong observed pK_a. No reason for compounds 150 and 228 being outliers can be offered, but they were both excluded.

Table 2 shows the results of the PLS analysis after the outliers were removed. Here, one of the BCP property models always outperforms the bond length model at each level of theory in terms of both r^2 and q^2. Generally, the models improved when the VIP < 1 "cut-off" was used. Level *E* gave the best results out of the all acid models. After observing the r^2 and q^2 for all of the models we found that level *C* gave the best results for the subsets. However, level *E* gave the best results for a model involving *all* carboxylic acids, although this model is constructed from 6 LVs. Figure 3 shows a comparison of the CPU time needed for optimization for nine of the compounds (three from each subset) that converged at each level of theory without any restarts versus the highest q^2 from each level of theory. This demonstrates that the less computationally expensive level *C* provides better results for the subsets and only worse results for the all acid models. Level *A* and *B* are inferior in all cases. At this stage we are in a position to explore different statistical learning methods to predict pK_a values based of BCP properties. To test the suitability of Level *C* in pK_a prediction and to examine whether the initial results still hold we carried out the analysis with level *C* and the more computationally expensive level *E*.

3.2. The Data for all the Comparisons in Section 3.2. to 3.5. can be Found in Table S2 in the Supporting Information. Comparison of the Statistical Learning Methods. For each learning method, there are four groups of compounds to test: all carboxylic acids, aliphatic acids, meta/para- and ortho-substituted acids. There are three different sizes of CV sets for both the systematic and random sampling method. This gave 24 values (= 4 groups of compounds × 3 CV set sizes × 2 sampling methods) of q^2 and RMSE to compare for each learning method employed. At level *C*, SVM gave the lowest RMSE value 16 times out

Table 2. Summary of the Initial PLS Analysis after the Outliers Had Been Removed[b]

		all acids			ortho subset			para/meta subset			aliphatic subset		
level	descriptors	LV[a]	r^2	q^2	LV[a]	r^2	q^2	LV[a]	r^2	q^2	LV[a]	r^2	q^2
A	bond lengths	2	0.630	0.612	2	0.797	0.787	1	0.900	0.896	2	0.795	0.748
B	bond lengths	2	0.601	0.581	2	0.838	0.801	1	0.905	0.902	2	0.837	0.828
	BCP properties	4	0.769	0.730	1	0.837	0.810	1	0.912	0.911	2	0.863	0.852
	BCP properties (VIP > 1)	4	0.764	0.749	1	0.835	0.815	1	0.909	0.908	6	0.895	0.861
C	bond lengths	2	0.696	0.688	1	0.841	0.832	1	0.931	0.929	1	0.875	0.875
	BCP properties	3	0.802	0.779	1	0.837	0.819	1	0.931	0.928	3	0.897	0.873
	BCP properties (VIP > 1)	2	0.787	0.782	1	**0.848**	**0.836**	1	**0.933**	**0.932**	4	**0.911**	**0.901**
D	bond lengths	2	0.486	0.455	1	0.777	0.762	1	0.912	0.906	3	0.807	0.784
	BCP properties	6	0.860	0.839	2	0.851	0.813	1	0.923	0.919	3	0.882	0.869
	BCP properties (VIP > 1)	4	0.738	0.689	1	0.842	0.833	1	0.927	0.925	2	0.890	0.881
E	bond lengths	2	0.535	0.498	1	0.803	0.794	2	0.917	0.907	2	0.810	0.788
	BCP properties	6	**0.879**	**0.844**	2	0.866	0.818	1	0.917	0.913	2	0.875	0.858
	BCP properties (VIP > 1)	7	0.845	0.818	1	0.844	0.835	1	0.918	0.913	3	0.897	0.886

[a] Number of latent variables. [b] The bold text highlights the best models for each group based on having the highest q^2.

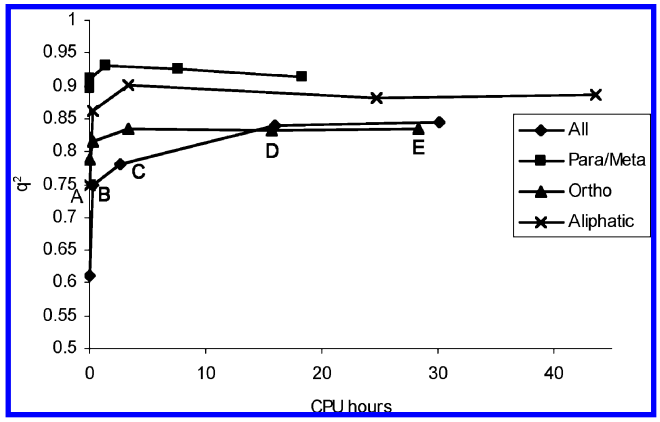

Figure 3. Comparison of the CPU times needed to optimize the compounds versus the highest q^2 value at each level of theory (*A, B, C, D*, and *E*).

of 24 comparisons, PLS 7 out of 24 comparisons, and RBFNN 2 times out of 24 comparisons (the RMSE for the random sampling para/meta 10-fold CV models were identical for PLS and SVM). Again out of 24 comparisons but at level *E* this time, SVM gave the lowest RMSE 21 times, PLS 3 times, and RBFNN 2 times (the RMSE for the systematic sampling para/meta 4-fold CV models and the random sampling 7-fold CV models were identical for PLS and SVM). This clearly demonstrates that SVM is superior to both PLS and RBFNN, at both levels of theory.

3.3. Comparison of the Level of Theory. Comparing the results of each q^2 and RMSE values obtained at both levels of theory, level *C* produces the highest q^2 46 times out of 72 ($=3 \times 24$) comparisons and the lowest RMSE 45 times out of the 72 comparisons, with one value being the same for level *C* and *E*. This clearly demonstrates that level *C* is superior to level *E*.

3.4. Comparison of the Validation Set Selection Methods. At each level of theory, there are 36 RMSE and q^2 values to compare because there are 3 machine learning methods, 3 values for *r* in r-fold CV, and 4 compound groups ($36=3 \times 3 \times 4$). At level *C*, there are 24 higher q^2 values for systematic sampling compared to 11 for random sampling and 1 q^2 value that is the same ($36=24 + 11 + 1$). There are 22 RMSE values lower for systematic sampling compared to 13 lower RMSE values for random sampling and 1 RMSE value that is the same ($36=22 + 13 + 1$). At level *E*, there are 19 higher q^2 values for systematic sampling compared to 16 for random sampling and 1 q^2 value that is the same ($36=19 + 16 + 1$). There are 21 RMSE values that are lower for systematic sampling compared to 15 values for random sampling ($36=21 + 15$). The better results gained from systematic sampling is not surprising because this method ensures the maximum value for the denominator in the q^2 equation. Using systematic sampling means that the most dissimilar compounds in relation to their pK_a values are excluded, and so it is more likely that the training set will contain similar compounds to the CV set therefore leading to better predictions.[72] When random sampling is used, one cannot ascertain that the CV sets contain all similar compounds in terms of their pK_a values. Hence, the models are possibly trained using compounds dissimilar to the CV set, which can lead to poor predictions. If the CV sets contain similar compounds in terms of their pK_a values, then this can lead to a small denominator in eq 1, thus reducing the q^2 value.

3.5. Comparison of Validation Set Size. With regards to the r-fold validation, 10-fold validation generally provided the highest q^2 and the lowest RMSE values, although this was not always the case. When the difference is calculated between the highest and lowest q^2 values for the r-fold CV set size for each subset and each method, the mean difference is 0.057 (standard deviation (SD) of 0.127) and 0.039 (SD=0.042) for levels *C* and *E*, respectively. When the same is carried out for the RMSE values, then the mean difference is 0.041 (SD=0.040) and 0.042 (SD=0.042) for levels *C* and *E*, respectively. When the poor result for 4-fold-para/meta-random sampling using RBFNN at level *C* is omitted, then the mean difference for q^2 is 0.032 (SD=0.025), and the mean value for RMSE is 0.035 (SD=0.024). The r-fold CV does give different results, but the difference is small. It is not surprising that 10-fold CV generally gives the best validation statistics because the smaller the CV groups, the more compounds there are to train the models. What these results do suggest is that the models still provide good predictions even when 25% of the compounds are omitted in training when using 4-fold CV.

3.6. Confirmation of Findings Based on Averages. Table 3 and Figure 4 show the results of all the subsets and learning methods when the r-fold CV results have been averaged and the results of the subsets and learning methods when the sampling methods have been averaged (from Table S2 in

Table 3. Average Values of the Results from Table S2

		PLS		SVM		RBFNN	
	set	q^2	RMSE	q^2	RMSE	q^2	RMSE
Level *C*							
systematic sampling	all	0.780	0.422	0.893	0.291	0.897	0.293
	ortho	0.834	0.375	0.821	0.380	0.784	0.427
	para/meta	0.930	0.109	0.924	0.113	0.896	0.128
	aliphatic	0.906	0.296	0.903	0.271	0.891	0.294
random sampling	all	0.777	0.430	0.879	0.295	0.875	0.322
	ortho	0.821	0.377	0.829	0.375	0.774	0.426
	para/meta	0.928	0.110	0.922	0.110	0.634	0.212
	aliphatic	0.904	0.270	0.909	0.265	0.885	0.303
average of systematic and random sampling	all	0.778	0.426	0.886	0.293	0.886	0.307
	ortho	0.828	0.376	0.825	0.378	0.779	0.427
	para/meta	0.929	0.109	0.923	0.112	0.765	0.170
	aliphatic	0.905	0.283	0.906	0.268	0.888	0.298
Level *E*							
systematic sampling	all	0.817	0.381	0.880	0.291	0.872	0.327
	ortho	0.825	0.386	0.851	0.356	0.763	0.431
	para/meta	0.913	0.129	0.916	0.125	0.893	0.144
	aliphatic	0.884	0.299	0.905	0.267	0.852	0.336
random sampling	all	0.805	0.392	0.887	0.295	0.881	0.316
	ortho	0.861	0.389	0.822	0.380	0.686	0.508
	para/meta	0.896	0.136	0.912	0.127	0.820	0.176
	aliphatic	0.881	0.300	0.903	0.273	0.886	0.301
average of systematic and random sampling	all	0.811	0.387	0.883	0.293	0.876	0.322
	ortho	0.843	0.388	0.837	0.368	0.724	0.469
	para/meta	0.905	0.132	0.914	0.126	0.857	0.160
	aliphatic	0.883	0.300	0.904	0.270	0.869	0.319

the Supporting Information). These results confirm what had previously been suggested. There is little difference between the CV statistics for the random and systematic sampling methods. Excluding the RBFNN para/meta results at level *C*, the largest difference in q^2 and RMSE values is 0.021 and 0.029, respectively, for the RBFNN all acid models. At level *E*, the largest difference in q^2 and RMSE is 0.078 and 0.077, respectively, for the RBFNN ortho models. Based on

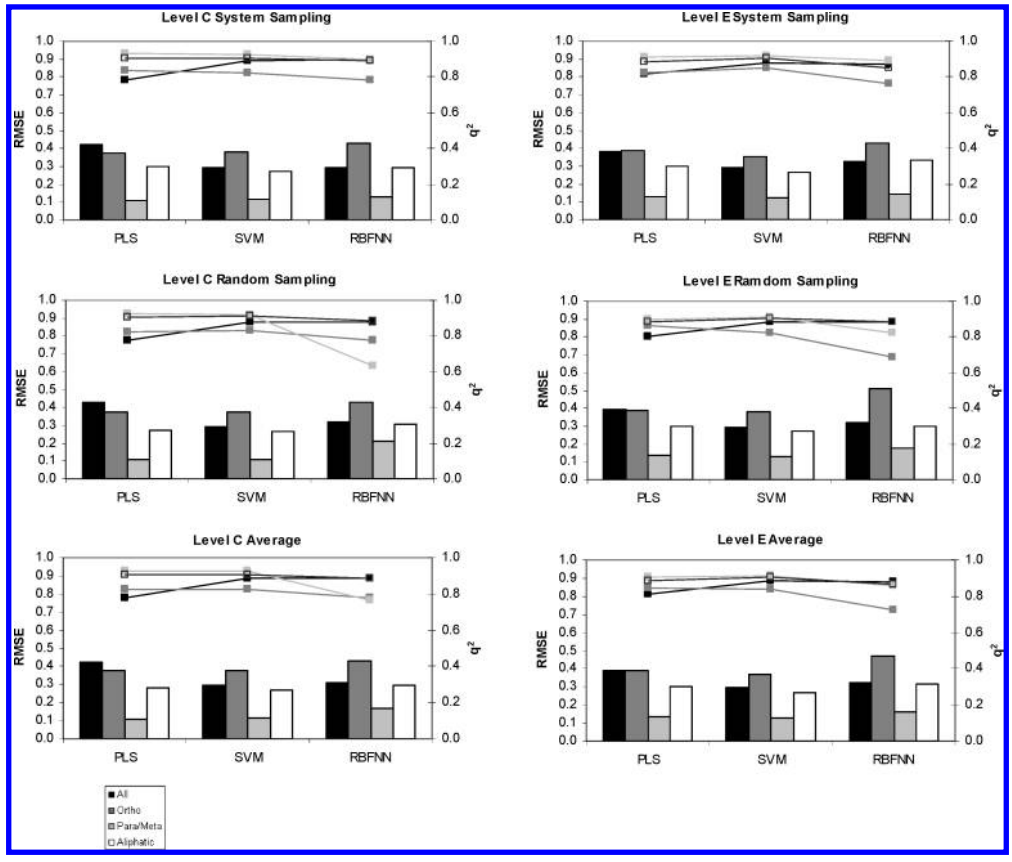

Figure 4. A graphical representation of Table 3. The bar charts represent the RMSE, and the lines represent the q^2 values obtained.

Table 4. RMSE for the Three Learning Methods Based on Leave-One-out[a]

	RMSE		
method	PLS	SVM	RBFNN
Level *C*			
C QTMS (all)	0.427	0.293	0.363
C QTMS (subset avg.)	0.285	0.276	0.321
C QTMS ortho subset	0.388	0.407	0.477
C QTMS para/meta subset	0.121	0.123	0.138
C QTMS aliphatic subset	0.278	0.252	0.291
Level *E*			
E QTMS (all)	0.396	0.301	0.323
E QTMS (subset avg.)	0.311	0.278	0.323
E QTMS ortho subset	0.407	0.367	0.486
E QTMS para/meta subset	0.136	0.131	0.168
E QTMS aliphatic subset	0.313	0.276	0.285
compared software/tools	RMSE		
ACD/Laboratories	0.263		
VCCLAB	0.279		
SPARC	0.356		
ChemAxon	0.398		

[a] The RMSE for the commercial computer programs are given at the bottom of the table.

the final average results of the r-fold and sampling methods, it is clear that SVM generally gives the best CV statistics. However, at level *C* the PLS statistics for the para/meta and ortho models provide the highest q^2 value and lowest RMSE. However, these are close to the q^2 value and RMSE value provided by SVM. Comparing the "all acid" models and the aliphatic models the q^2 value and RMSE for SVM are much better than the q^2 and RMSE for PLS. At level *E*, SVM provides the highest q^2 and lowest RMSE in all cases. Although PLS was better than SVM, in two cases at level *C* we chose SVM as the best learning method. This decision is based on the fact that, when PLS was superior, the difference between the statistics was small and, when SVM was better, then the difference between the statistics was large. Comparing the averages of systematic and random sampling for SVM level *E* only provides better CV statistics for the ortho data set. The difference between the statistics is small, but this may suggest that the more expensive level *E* accounts more for the steric effects than level *C* in some way.

Improved models were created when the data set was split into aliphatic and aromatic subsets, which were further split into meta/para and ortho-substituted carboxylic acids. The subset models provided significant improvements on the "all acid" model. The para/meta model was the most accurate in prediction, followed by the aliphatic and then the ortho model. The excellent CV results of the para/meta models are not surprising because the pK_a difference caused by substituent changes can be accurately predicted by the Hammett equation and BCP properties display a strong correlation with Hammett's sigma parameter.[25] The lower q^2 value and higher RMSE for the ortho model can be explained in terms of steric effects.[73] Whereas the pK_a of meta- and para-substituted carboxylic acids is affected mainly by inductive and resonance contributions, the pK_a of ortho-substituted carboxylic acids is highly sensitive to steric contributions. Primary steric hindrance to deprotonation is important where there are bulky groups around the acidic center. Secondary steric effects may either be acid-weakening (if there is steric hindrance to solvation) or acid-strengthening (if there is steric inhibition of resonance in the neutral molecule). We have already stated that when QTMS fails one can be certain that steric effects are very important. Since the BCP properties do not account for steric effects, we can be confident that this is the reason for the poorer results. Since the results of the aliphatic subset are an improvement on the ortho subset, the steric effects must have less importance than electronic contributions.

3.7. Comparison to pK_a Prediction Software. To compare pK_a prediction by the QTMS method with available (commercial) software, we have used the methods provided by a number of organizations, namely ACD/Laboratories' pK_a DB,[23] the SPARC[74] online calculator (SPARC Performs Automated Reasoning in Chemistry), VCCLAB's[75] Web-based ALOGPS 2.1 program, and ChemAxon's pK_a plugin[76] for their Marvin Beans applications. Recently, it was stated that the Web version of the SPARC performs 50,000−100,000 calculations per month.[77] Each software package enables the user to input the structures in SMILES[78] format.

We removed each of the compounds in turn from the global and subset carboxylic acid models built using PLS, SVM, and RBFNN, for both levels *C* and *E* and rebuilt the models using the new model to predict the pK_a of the compound omitted.[79] Using this method we acknowledge that the compounds are not an external test set (e.g. they have been used for initial variable and parameter selection in some cases), but nor can we be sure that they have not been used to train the packages we compare to. Table 4 gives the RMSE for the methods based on leave-one-out. The RMSE obtained from testing the alternative computer programs are also given in Table 4. These results confirm that SVM provides the best models to predict pK_a. The SVM models have the lowest RMSE in all the LOO cases apart from the level *C* ortho and para/meta carboxylic acid models, where the PLS models have the lowest RMSE of 0.388 for the ortho model and 0.121 for the para/meta model, compared to 0.407 and 0.123 for the SVM models, respectively. Comparing levels of theory confirms that level *C* is the best as it generally provides the lowest RMSE for the models. There are some exceptions to this. For example, the ortho

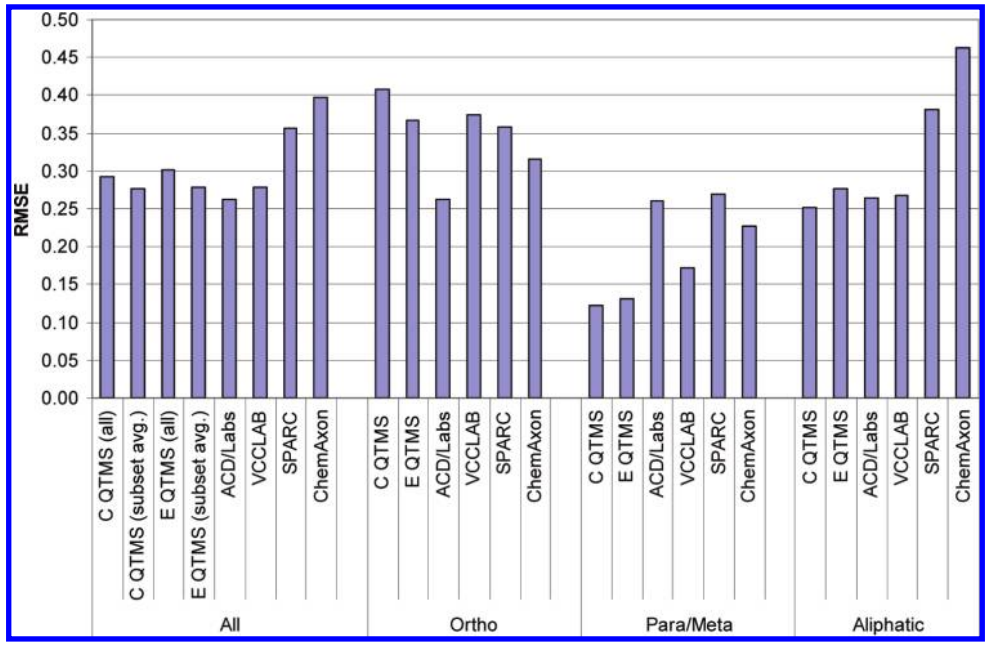

Figure 5. Comparison between QTMS and other pK_a prediction software, based on the RMSE.

model at level *E* using SVM has an RMSE of 0.367 compared to 0.407 for level *C*. Where level *E* provides a lower RMSE the largest difference observed in RMSE between levels *C* and *E* is 0.04 for the SVM ortho models and the PLS all carboxylic acid models. In fact, the difference between the models at the different levels judged by RMSE (based on LOO) is negligible when considering the large increase in CPU time needed to optimize the compounds at level *E* (see Figure 3).

Figure 5 graphically compares our QTMS SVM models to the results obtained from the commercial predictions. Recently, Meloun and Bordovská[77] have rigorously compared the same packages using 64 drugs and other organic molecules with complex and diverse structural patterns. Although we only base our ranking on the RMSE, we too found ACD/pK_a to be the most accurate method. This conclusion contradicts the findings of Dearden et al.[80] who compared ten prediction tools using a test set of 653 compounds. They found that ACD/pK_a was the *least* accurate out of the four programs we compared with, even when tautomeric compounds were excluded. The other three programs were ordered in the same way as our results: VCCLAB being the most accurate, followed by SPARC, and then ChemAxon. Apart from ACD/Laboratories, which is consistent across all the subsets, the different methods vary in their prediction ability. Out of all the methods, QTMS has the lowest RMSE for the para/meta substituted benzoic acids and aliphatic carboxylic acids but has the highest RMSE for the ortho-substituted benzoic acids. As has been previously pointed out,[33,45] QTMS fails when steric effects are important, which is the case for the ortho substituted benzoic acids.

4. CONCLUSION

The results presented in this systematic study indicate that BCP descriptors are effective in predicting the pK_a of small to large sized carboxylic acids of pharmaceutical relevance. Furthermore, extensive cross-validation shows that there is no need to use the computationally more expensive level *E* when level *C* provides similar, if not superior, CV statistics. More predictive models were gained from splitting the data set. Generally, SVM provides the best learning method although the lack of interpretability may mean it is not necessarily the most suitable method when mechanistic understanding is important. Finally, we have also demonstrated that predictions from our QTMS method compete with contemporary and frequently used pK_a prediction tools.

ACKNOWLEDGMENT

The work was funded by the EPSRC with sponsoring from GlaxoSmithKline (GSK), Stevenage, Great Britain. Thanks to Dr. M. Kranz at GSK for helpful discussions.

Supporting Information Available: Experimental pK_a values for the carboxylic acids (Table S1) and results for the three machine learning methods (PLS, SVM, RBFNN), using r-fold CV ($r = 4$, 7, and 10), and the systematic and random sampling method, at levels *C* and *E* (Table S2). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Montana, J. G.; Manallack, D. T.; Gancia, E. Estimation of pKa Using Semiemperical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids. *Quant. Struct.-Act. Relat.* **2002**, *21*, 457–471.
(2) da Silva, C. O.; da Silva, E. C.; Nascimento, M. A. C. Ab Initio Calculations of Absolute pK_a Values in Aqueous Solution I. Carboxylic Acids. *J. Phys. Chem. A* **1999**, *103*, 11194–11199.
(3) Cookson, R. F. The determination of acidity constants. *Chem. Rev.* **1974**, *74*, 5–28.
(4) Namazian, M.; Halvani, S. Calculations of pK_a values of carboxylic acids in aqueous solution using density function theory. *J. Chem. Thermodyn.* **2006**, *38*, 1495–1502.
(5) Jelfs, S.; Ertl, P.; Selzer, P. Estimation of pK_a for Druglike Compounds Using Semiempirical and Information-Based Descriptors. *J. Chem. Inf. Model.* **2007**, *47*, 450–459.
(6) Gruber, C.; Buss, V. Quantum-Mechanically Calculated Properties for the Development of Quantitative-Structure Activity Relationships. *Chemosphere* **1989**, *19*, 1595–1609.
(7) Citra, M. J. Estimating the pKa of Phenols, Carboxylic Acids and Alcohols from Semi-Empirical Quantum Chemical Methods. *Chemosphere* **1999**, *38*, 191–206.

(8) Gross, K. C.; Seybold, P. G. Substituent Effects on the Physical Properties and pKa of Phenols. *Int. J. Quantum Chem.* **2001**, *85*, 569–579.

(9) Xing, L.; Glen, R. C. Novel Methods for the Prediction of logP, pK_a and logD. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796–805.

(10) Xing, L.; Glen, R. C.; Clark, R. D. Predicting pK_a by Molecular Tree Structure Fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 870–879.

(11) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and original pK_a prediction method using grid molecular interaction firlds. *J. Chem. Inf. Model.* **2007**, *47*, 2172–2181.

(12) Kogej, T.; Muresan, S. Database Mining for pK_a Prediction. *Curr. Drug Discovery Technol.* **2005**, *4*, 221–229.

(13) Liptak, M. D.; Shields, G. C. Accurate pK_a Calculations for Carboxylic Acids Using Complete Basis Set and Gaussian-n Models Combined with CPCM Continuum Solvation Methods. *J. Am. Chem. Soc.* **2001**, *123*, 7314–7319.

(14) Liptak, M. D.; Shields, G. C. Experimentation with Different Thermodynamic Cycles Used for pK_a Calculations on Carboxylic Acids Using Complete Basis Set and Gaussian-n Models Combined with CPCM Continuum Solvation Methods. *Intern. J. Quant. Chem.* **2001**, *85*, 727–741.

(15) Lu, H.; Chen, X.; Zhan, C. G. First-principles calculation of pK_a for cocaine, nicotine, neurotransmitters, and anilines in aqueous solution. *J. Phys. Chem. B* **2007**, *111*, 10599–10605.

(16) Soriano, E.; Cerdan, S.; Ballesteros, P. Computational determination of pK_a values. A comparison of different theoretical approaches and a novel procedure. *J. Mol. Struct. (Theochem)* **2004**, *684*, 121–128.

(17) *Jaguar version 6.0*; Schrodinger, LLC: New York, USA, 2005.

(18) Namazian, M.; Halvani, S. Calculations of pK_a Values of Carboxylic Acids in Aqueous Solution Using Moller-Plesset Perturbation Theory. *J. Iran. Chem. Soc.* **2005**, *2* (1), 65–70.

(19) Namazian, M.; Halvani, S.; Noorbala, M. R. Density Functional Theory Response to the Calculations of pK_a Values of some Carboxylic Acids in Aqueous Solution. *J. Mol. Struct.* **2004**, *711*, 13–18.

(20) Namazian, M.; Heidary, H. Ab Initio Calculations of pK_a Values of some Organic Acids in Aqueous Solution. *J. Mol. Struct.* **2003**, *620*, 257–263.

(21) Namazian, M.; Kalantary-Fotooh, F.; Noorbala, M. R.; Searles, D. J.; Coote, M. L. Moller-Plesset Perturbation Theory Calculations of the pK_a Values for a Range of Carboxylic Acids. *J. Mol. Struct. THEOCHEM* **2006**, *758*, 275–278.

(22) da Silva, C. O.; da Silva, E. C.; M, A. C. Nascimento, Ab Initio Calculations of Absolute pK_a Values in Aqueous Solution II. Aliphatic Alcohols, Thiols, and Halogenated Carboxylic Acids. *J. Phys. Chem. A* **2000**, *104*, 2402–2409.

(23) ACD/Labs, version 3;ACD Labs, Toronto, ON, Canada.

(24) O'Brien, S. E.; Popelier, P. L. A. Quantum Molecular Similarity. Part 3: QTMS descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 764–775.

(25) Popelier, P. L. A. Quantum molecular similarity. 1. BCP space. *J. Phys. Chem. A* **1999**, *103* (15), 2883–2890.

(26) Bader, R. F. W. *Atom in Molecules. A Quantum Theory*; Oxford Univ. Press: 1990.

(27) Popelier, P. L. A. *Atoms in Molecules. An Introduction*; Pearson Education: London, Great Britain, 2000.

(28) O'Brien, S. E.; Popelier, P. L. A. Quantum Topological Molecular Similarity. Part 4: A QSAR study of Cell Growth Inhibitory Properties of Substituted (E)-1-Phenylbut-1-en-3-ones. *J. Chem. Soc., Perkin Trans. 2* **2002**, 478–483.

(29) Popelier, P. L. A.; Chaudry, U. A.; Smith, P. J. Quantitative Structure-Activity Relationships of Mutagenic Acitvity from Quantum Topological Descriptors: Triazenes and Halogenated Hydroxfuranones (mutagen-X) Derivatives. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 709–718.

(30) Popelier, P. L. A.; Chaudry, U. A.; Smith, P. J. Quantum Topological Molecular Similarity. Part 5: Further Development with an Application to the toxicity of Polychlorinated dibenzo-p-dioxins (PCDDs). *J. Chem. Soc., Perkin II* **2002**, 1231–1237.

(31) Mohajeri, A.; Hemmateenejad, B.; Mehdipour, A.; Miri, R. Modeling calcium channel antagonistic activity of dihydropyridine derivatives using QTMS indices analyzed by GA-PLS and PC-GA-PLS. *J. Mol. Graphics Modell.* **2008**, *2008*, 1057–1065.

(32) Roy, K.; Popelier, P. L. A. Exploring predictive QSAR models for hepatocyte toxicity of phenols using QTMS descriptors. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 2604–2609.

(33) Popelier, P. L. A.; Smith, P. J. QSAR models based on Quantum Topological Molecular Similarity. *Eur. J. Med. Chem.* **2006**, *41*, 862–873.

(34) Hemmateenejad, B.; Mehdipour, A. R.; Popelier, P. L. A. Quantum Topological QSAR Models based on the MOLMAP Approach. *Chem. Biol. Drug Des.* **2008**, *72*, 551–563.

(35) Roy, K.; Popelier, P. L. A. Exploring Predictive QSAR Models Using Quantum Topological Molecular Similarity (QTMS) Descriptors for Toxicity of Nitroaromatics to Saccharomyces cerevisiae. *QSAR Comb. Sci.* **2008**, *27*, 1006–1012.

(36) Alsberg, B. K.; Marchand-Geneste, N.; King, R. D. A New 3D Molecular Structure Representation using Quantum Topology with application to structure-property relationships. *Chemom. Intell. Lab. Syst.* **2000**, *54*, 75–91.

(37) Chaudry, U. A.; Popelier, P. L. A. Ester hydrolysis rate constant prediction from quantum topological molecular similarity (QTMS) descriptors. *J. Phys. Chem. A* **2003**, *107*, 4578–4582.

(38) Alsberg, B. K.; Marchand-Geneste, N.; King, R. D. Modeling quantitative structure-property relationships in calculated reaction pathways using a new 3D quantum topological representation. *Anal. Chim. Acta* **2001**, *446*, 3–13.

(39) Hemmateenejad, B.; Mohajeri, A. Application of quantum topological molecular similarity descriptors in QSPR study of the O-methylation of substituted phenols. *J. Comput. Chem.* **2007**, *29*, 266–274.

(40) Roy, K.; Popelier, P. L. A. Predictive QSPR modeling of the acidic dissociation constant (pKa) of phenols in different solvents. *J. Phys. Org. Chem.* **2009**, *22*, 186–196.

(41) Chaudry, U. A.; Popelier, P. L. A. Estimation of pKa using Quantum Topological Molecular Similarity (QTMS) descriptors: Application to Carboxylic Acids, Anilines and Phenols. *J. Org. Chem.* **2004**, *69*, 233–241.

(42) Adam, K. R. New Density Functional and Atoms in Molecules Method of Computing Relative pKa Values in Solution. *J. Phys. Chem. A* **2002**, *106*, 11963–11972.

(43) Bader, R. F. W.; Beddall, P. M. The Spatial Partitioning and Transferability of Molecular Energies. *Chem. Phys. Lett.* **1971**, *8*, 29–36.

(44) Esteki, M.; Hemmateenejad, B.; Khayamian, T.; Mohajeri, A. Multi-way analysis of Quantum Topological Molecular Similarity descriptors for modelling acidity constants of some phenolic compounds. *Chem. Biol. Drug Des.* **2007**, *70*, 413–423.

(45) Loader, R. J.; Singh, N. K.; O'Malley, P. J.; Popelier, P. L. A. The Cytotoxicity of ortho alkyl substituted 4-X-phenols: A QSAR based on theoretical bond lengths and electron densities. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1249–1254.

(46) Selassie, C. D.; Verma, R. P.; Kapur, S.; Shusterman, A. J.; Hansch, C. QSAR for the cytotoxicity of the radical reaction. *J. Chem. Soc., Perkin Trans. II* **2002**, 1112–1117.

(47) UMETRICS. *SIMCA-P 10.0*; Umeå, Sweden, 2002. info@umetrics.com, www.umetrics.com (accessed month day, year).

(48) Howard, S. T.; Lamarche, O. Description of covalent bond orders using the charge density topology. *J. Phys. Org. Chem.* **2003**, *16*, 133–141.

(49) Bader, R. F. W.; Slee, T. S.; Cremer, D.; Kraka, E. Description of Conjugation and Hyperconjugation in Terms of Electron Distributions. *J. Am. Chem. Soc.* **1983**, *105* (15), 5061–5068.

(50) Bader, R. F. W.; Preston, H. J. T. The Kinetic Energy of Molecular Charge Distributions and Molecular Stability. *Int. J. Quantum Chem.* **1969**, *3*, 327–347.

(51) Schaftenaar, G.; Noordik, J. H. Molden: a pre- and post-processing program for molecular and electronic structures. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 123–134.

(52) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, J., T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. Gaussian; 2003.

(53) Popelier, P. L. A. *MORPHY98*; MORPHY98 - a program written by P. L. A. Popelier with a contribution from R. G. A. Bone. UMIST: Manchester, England, 1998.

(54) Popelier, P. L. A. A Robust Algorithm to Locate Automatically All Types of Critical- Points in the Charge-Density and Its Laplacian. *Chem. Phys. Lett.* **1994**, *228* (1−3), 160–164.

(55) Wold, S.; Sjostrom, M.; Eriksson, L. Partial Least Squares Projections to Latent Structures (PLS) in Chemistry. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: Chichester, GB, 1998; Vol. 3, pp 2006−2021.

(56) Livingstone, D. J. *Data Analysis for Chemists*, 1st ed; Oxford University Press: Oxford, Great Britain, 1995.

(57) Vapnik, V. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer-Verlag: New York, USA, 1995.

(58) Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining Knowl. Disc.* **1998**, *2*, 121–167.
(59) Smola, A. J.; Schoelkopf, B. A tutorial on support vector regression. *Stat. Comp.* **2004**, *14* (3), 199–222.
(60) Ivanciuc, O. Applications of Support Vector Machines in Chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Cundari, T. R., Eds.; Wiley-VCH: 2007; Vol. *23*, pp 291−400.
(61) Jover, J.; Basque, R.; Sales, J. QSPR prediction of pK_a for benzoic acids in different solvents. *QSAR Comb. Sci.* **2008**, *27* (5), 563–581.
(62) Dimitriadou, E.; Hornik, K.; Leisch, D. M. F.; Weingessel, A. *e1071: Misc Functions of the Department of Statistics (e1071)*; R package version 1.5−16; TU Wien: 2006.
(63) Team, R. D. C. R: A language and environment for statistical computing; R Foundation for Statistical Computing: Vienna, Austria, 2007. ISBN 3-900051-07-0. URL: http://www.R-project.org (accessed month day, year).
(64) Haykin, S. *Neural Networks: A Comprehensive Foundation*, 2nd ed.; Prentice-Hall: Upper Saddle River, New Jersey, USA, 1999.
(65) Gurney, K. *An Introduction to Neural Networks*; Routledge: London, Great Britain, 1997.
(66) Yao, X. J.; Wang, Y. W.; Zhang, X. Y.; Zhang, R. S.; Lui, M. C.; Hu, Z. D.; Fan, B. T. Radial basis function neural network-based QSPR for the prediction of critical temperature. *Chemom. Intell. Lab. Syst.* **2002**, *62*, 217–225.
(67) Chen, S.; Cowan, C.; Grant, P. Orthogonal least squares learning for radial basis function networks. *IEEE Trans. Neural Networks* **1991**, *2*, 302–309.
(68) Ho, M.; Schmider, H.; Edgecombe, K. E.; Smith, V. H. J. Topological Analysis of Valence Electron Charge Distributions from Semi-empirical and ab initio methods. *Int. J. Quant. Chem., Quant. Chem. Symp.* **1994**, *28*, 215–226.
(69) http://www.chemicaldictionary.org/dic/F/Flurenol_2074.html (accessed month day, year).
(70) Kresge, A. J.; Pojarlieff, I. G.; Rubinstein, E. M. The acidity constant of fluorene-9-carboylic acid in aqueous solution. Determination of the pKa, of a sparingly soluble substance. *Can. J. Chem.* **1993**, *71*, 227.
(71) Jumppanen, J. H.; Siren, H.; Riekkola, M. L. Correlation of Resolution with Frictional Coefficients and pK, Values in Capillary Electrophoresis of Four Diuretics: Determination of Electric Field Strength and Electro-Osmotic Velocity. *J. Microcolumn Separation* **1993**, *5* (5), 451.
(72) Leonard, J. T.; Roy, K. On selection of training and test sets for the development of predictive QSAR models. *QSAR Comb. Sci.* **2006**, *25*, 235–251.
(73) Perrin, D. D.; Dempsey, B.; Serjean, E. P. *pKa Prediction for Organic Acids and Bases*; Chapman and Hall: London, GB, 1981.
(74) Hilal, S. H.; Karickoff, S. W.; Carreira, L. A. A rigorous test for SPARC's chemical reactivity models: Estimation of more than 4300 ionization pK_a's. *Quant. Struc. Act. Relat.* **1995**, *14* (348), .
(75) VCCLAB. *Virtual Computational Chemistry Laboratory*; 2005. http://www.vcclab.org (accessed month day, year).
(76) ChemAxon. *Calculator Plugins were used for structure property prediction and calculation, Marvin 2.0.4, 2006*; ChemAxon. http://www.chemaxon.com (accessed month day, year).
(77) Meloun, M.; Bordovska, S. Benchmarking and Validating algorithms that estimate pK_a values of drugs based on their molecular structure. *Anal. Bioanal. Chem.* **2007**, *389*, 1267–1281.
(78) Weininger, D. SMILES, a Chemical Language and Information System, 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.
(79) Zhang, J.; Kleinoder, T.; Gasteiger, J. Prediction of pK_a Values for Aliphatic Carboxylic Acids and Alcohols with Empirical Atomic Charge Descriptors. *J. Chem. Inf. Model.* **2006**, *46*, 2256–2266.
(80) Dearden, J. C.; Cronin, T. D.; Lappin, D. C. A comparison of commercially avaliable software for the prediction of pK_a values (poster). At UK-QSAR, 24 April 2007, Astrazeneca, Alderley Park, UK, 2007.

CI900172H