Comparative Analysis of QSAR Models for Predicting pK_a of Organic Oxygen Acids and Nitrogen Bases from Molecular Structure

Haiying Yu,[†,‡] Ralph Kühne,[†] Ralf-Uwe Ebert,[†] and Gerrit Schüürmann*[,†,‡]

UFZ Department of Ecological Chemistry, Helmholtz Centre for Environmental Research, Permoserstrasse 15, D-04318 Leipzig, Germany and Institute for Organic Chemistry, Technical University Bergakademie Freiberg, Leipziger Strasse 29, D-09596 Freiberg, Germany

Received August 9, 2010

For 1143 organic compounds comprising 580 oxygen acids and 563 nitrogen bases that cover more than 17 orders of experimental pK_a (from −5.00 to 12.23), the pK_a prediction performances of ACD, SPARC, and two calibrations of a semiempirical quantum chemical (QC) AM1 approach have been analyzed. The overall root-mean-square errors (rms) for the acids are 0.41, 0.58 (0.42 without *ortho*-substituted phenols with intramolecular H-bonding), and 0.55 and for the bases are 0.65, 0.70, 1.17, and 1.27 for ACD, SPARC, and both QC methods, respectively. Method-specific performances are discussed in detail for six acid subsets (phenols and aromatic and aliphatic carboxylic acids with different substitution patterns) and nine base subsets (anilines, primary, secondary and tertiary amines, *meta*/*para*-substituted and *ortho*-substituted pyridines, pyrimidines, imidazoles, and quinolines). The results demonstrate an overall better performance for acids than for bases but also a substantial variation across subsets. For the overall best-performing ACD, rms ranges from 0.12 to 1.11 and 0.40 to 1.21 pK_a units for the acid and base subsets, respectively. With regard to the squared correlation coefficient r^2, the results are 0.86 to 0.96 (acids) and 0.79 to 0.95 (bases) for ACD, 0.77 to 0.95 (acids) and 0.85 to 0.97 (bases) for SPARC, and 0.64 to 0.87 (acids) and 0.43 to 0.83 (bases) for the QC methods, respectively. Attention is paid to structural and method-specific causes for observed pitfalls. The significant subset dependence of the prediction performances suggests a consensus modeling approach.

INTRODUCTION

In aqueous solution, proton transfer between ionizing organic compounds and water plays an important role for their speciation, which in turn affects the extent of sorption into organic matter, the volatilization into air, and the uptake into organisms. Taking pharmacology as an example, the degree of dissociation of a Brönstedt acid AH governs both its overall solubility as well as the processes related to absorption, distribution, metabolism, and excretion (ADME). As is well-known, quantification of the degree of dissociation proceeds through evaluation of the equilibrium constant of the proton-transfer reaction[1]

$$AH + H_2O \rightleftarrows A^- + H_3O^+ \quad (1)$$

In dilute aqueous solutions (assuming that activities can be replaced by concentrations), the concentration of water, $[H_2O]$, is constant, and thus can be built in the acid constant K_a:

$$K_a = \frac{[A^-]}{[AH]} \cdot [H_3O^+] \quad (2)$$

which leads to $pK_a = -\log K_a$:

$$pK_a = pH + \log \frac{[HA]}{[A^-]} \quad (3)$$

For organic bases B, eq 1 can be applied to their conjugated acids, $BH^+ + H_2O \rightleftarrows B + H_3O^+$. Accordingly, p$K_a$ of a base often (and in this study always) refers to the pK_a of its conjugated acid.[1]

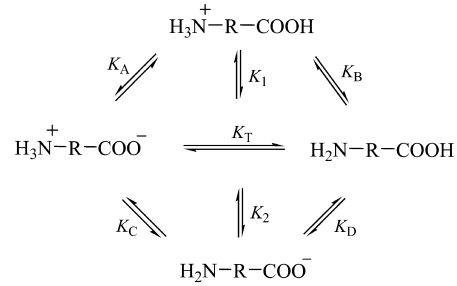
Methods to predict pK_a from molecular structure have been reviewed recently, covering both ionizing groups of proteins as well as small organic acids and bases.[2] A prominent group of prediction methods is based on Hammett-type linear free-energy relationships (LFERs).[1] ACD[3] as a commercial software package is a respective example, and SPARC (SPARC performs automated reasoning in chemistry)[4−6] employs a similar methodology but uses functional groups rather than compound classes for defining reference pK_a values as a starting point of the calculation. According to recent comparative analyses of the performance of several prediction methods with drugs,[7,8] both ACD and SPARC outperformed most other methods. For a larger set of 644 organics, however, mean absolute errors of 0.78 (SPARC) and 1.08 (ACD) pK_a units were reported,[1] reflecting an only moderate performance.

Quantum chemical approaches addressing aqueous solvation have been typically applied to smaller data sets.[9−15] Here, different levels of computation can be used for the intrinsic gas-phase energetics and the solvation contribution,[15] and for the sophisticated treatment of both long- and short-range electrostatics as well as solute−solvent H-

* Address correspondence to gerrit.schuurmann@ufz.de.
† UFZ Helmholtz Centre for Environmental Research.
‡ Technical University Bergakademie Freiberg.

10.1021/ci100306k
Published on Web 10/29/2010

Scheme 1. Organic Acid with Two Ionogenic Sites[a]



[a] Shows the relationship between four microdissociation constants K_A, K_B, K_C, and K_D, two macrodissociation constants K_1 and K_2, and one tautomerism constant K_T.

bonding enables first-principle calculations of pK_a.[13,14] Moreover, gas-phase quantum chemistry has been employed to empirically calibrate calculated parameters with pK_a,[16,17] and two-dimensional (2D) structural fingerprints have been combined with quantum chemistry[18] or used alone[19−21] for deriving pK_a prediction methods with larger data sets.

When comparing experimental and predicted pK_a for compounds with more than one ionogenic group, a fundamental issue is the difference between the individual site-specific pK_a values, the micro-pK_a's, and the macro-pK_a's as actually measured values that result from the additional interconversion of associated tautomers. The situation is illustrated in Scheme 1. As can be seen from this scheme, the fully protonated form $H_3N^+-R-COOH$ (with R representing some organic moiety) may deprotonate in two ways with associated acid constants K_A (instead of K_a (pathway A) we use the simplified notation K_A) and K_B (right pathway). The resultant tautomers are related to each other through the equilibrium constant K_T and may deprotonate further to $H_2N-R-COO^-$, governed by associated dissociation constants K_C and K_D (left and right pathways). Because $H_3N^+-R-COOH$ is in equilibrium with both $H_3N^+-R-COO^-$ and $H_2N-R-COOH$ (here called equilibrium 1), an apparent or macrodissociation constant K_1:

$$K_1 = \frac{([H_3N^+-R-COO^-] + [H_2N-R-COOH]) \cdot [H_3O^+]}{[H_3N^+-R-COOH]} \quad (4)$$

evolves. Taking *p*-amino benzoic acid ($H_2N-Ph-COOH$) as example, pK_1 = 2.42 (= pK_a of equilibrium 1). Correspondingly, the second macrodissociation constant K_2 reads

$$K_2 = \frac{[H_2N-R-COO^-] \cdot [H_3O^+]}{[H_3N^+-R-COO^-] + [H_2N-R-COOH]} \quad (5)$$

with pK_2 = 4.88 for our example. The four microconstants $K_A = ([H_3N^+-R-COO^-] \cdot [H_3O^+])/[H_3N^+-R-COOH]$, $K_B = ([H_2N-R-COOH] \cdot [H_3O^+])/[H_3N^+-R-COOH]$, $K_C = ([H_2N-R-COO^-] \cdot [H_3O^+])/[H_3N^+-R-COO^-]$, and $K_D = ([H_2N-R-COO^-] \cdot [H_3O^+])/[H_2N-R-COOH]$ and the two macroconstants K_1 and K_2 are related to each other as follows (see eqs 4 and 5 and Scheme 1):

$$K_A + K_B = K_1 \quad (6)$$

$$\frac{1}{K_C} + \frac{1}{K_D} = \frac{1}{K_2} \quad (7)$$

Moreover, the relationship between the tautomeric constant $K_T = [H_2N-R-COOH]/[H_3N^+-Ph-COO^-]$ and the microconstants is $K_T = K_B/K_A = K_C/K_D$. The more K_T approaches 1, the more K_1 is affected by both K_A and K_B (and K_2 by both K_C and K_D). Conversely, large K_T ($\gg$1) results in $K_1 \approx K_B$, and small K_T ($\ll$ 1) results in $K_1 \approx K_A$, with corresponding situations for K_2 with regard to K_C and K_D. For *p*-amino benzoic acid, pK_A = 3.57, pK_B = 2.41 ($\approx$pK_1, see above), pK_C = 3.72, pK_D = 4.75 ($\approx$pK_2), and K_T = 13. Prediction of pK_a of polyprotic species from molecular structure yields microscopic constants, which can then be converted to macroscopic constants through application of eqs 4−7 (or in more complex situations through corresponding extensions).

In the present study, ACD, SPARC and the Tehan quantum chemical method in both the original form (QC) and a recalibrated version (r-QC, see below) have been comparatively analyzed for their pK_a prediction performance using a set of 1143 organic compounds that comprise 580 oxygen acids (phenols and aliphatic and aromatic carboxylic acids) and 563 nitrogen bases (anilines, aliphatic amines, and N-heterocycles). The results show significant variations in prediction quality across compound classes as well as method-specific pitfalls related to specific structural patterns. ACD shows up as the overall best-performing method with rms values from 0.12 to 1.21 pK_a units and mean absolute errors ranging from 0.07 to 0.97 pK_a units. It follows that when applying ACD, SPARC, or the Tehan models for predicting pK_a, the performance may differ significantly, dependent on the structure type of interest.

MATERIALS AND METHODS

Data Set and Chemical Domain. For the comparative performance analysis of ACD,[3] SPARC,[4−6] the Tehan QC,[16,17] and r-QC (see below), 582 organic oxygen acids and 571 organic nitrogen bases with experimental pK_a data have been collected from literature,[14−17,22−30] including the 417 acids and 282 bases used for calibrating QC by Tehan et al.[16,17] The measurement temperature ranged from 10 to 30 °C and in most cases (93%) was confined to about 20 to 25 °C, and thus was not considered further because of its only moderate (and in this case probably negligible) impact on pK_a.

Initial calculations showed that four compounds which were already identified as outliers by Tehan et al.[16,17] also have large prediction errors with SPARC and ACD: 3,4-diamino-benzoic acid: pK_a 3.49 (experimental) vs 4.50 (Tehan), 5.05 (SPARC), and 5.02 (ACD); 4-(4-chloro-*o*-tolyloxy)butyric acid (MCPB): 6.20 vs 3.90, 4.46, 4.58; flurenol: 1.09 vs 4.18, 3.43, 3.25; and fenclorim: 4.23 vs 1.02, −3.54, −4.69. The unusually large prediction errors across all methods suggest that the experimental data may be questionable.

Moreover, the original base set[17] contained also 2-hydroxypyridine (0.75) and 8-quinolinol (4.90), which however are subject to lactam−lactim tautomerism and in fact are largely prevalent in the pyridone form in aqueous solutions. Accordingly, these two compounds appear to be better

handled as nitrogen acids (except that protonation would preferably occur at the exocyclic carbonyl oxygen), and thus do not fit the present set of oxygen acids and nitrogen bases.

Finally, substantial pK_a overestimations by SPARC and ACD (by 2.9−4.0 and 2.7−4.7 units, respectively) were obtained for the following four compounds: dinicotinic acid (exp. pK_a = 1.10), nicotinic acid (2.07), isonicotinic acid (1.70), picolinic acid (1.06). All of these compounds contain more than one ionizing group with similar pK_a values such that more than one micro-pK_a contributes significantly to the resultant macro-pK_a (COOH-substituted pyridine-N conjugated acid pK_a: 4.6 and aromatically bound COOH pK_a: 4.2). At the same time, the original reference suggests these compounds to be classified as oxygen acids rather than as conjugated acids of pyridine-N.[31] (Because both SPARC and ACD explicitly address macro-pK_a values, the large discrepancies suggest again that the experimental data may be questionable.)

At present, the reason for the unusually large discrepancies between experimental and predicted pK_a across all four methods is not clear. Note further that in the original literature reporting these data, there is no indication about any particular problems related to the experimental procedures. To avoid picking best-fit experimental values for evaluating the prediction performances, the pK_a data used preferably in our study were taken — as far as possible — from the EpiSuite PhysProp collection. However, for MCPB and 3,4-amino benzoid acid, as two of the ten outliers, there are individual literature pK_a data closer to the predicted values. A possible way forward would be to compare the pK_a data of the outlying compounds with results from first-principle benchmark calculations, which however is not the subject of the present investigation.

Omission of the 10 compounds just discussed resulted in the final test set of 1143 organic compounds consisting of 580 oxygen acids and 563 nitrogen bases. Note, however, that the degree of overlap between this test set and the training sets used for calibrating ACD[3] and SPARC[4,5] is unknown, because — to our best knowledge — these models did not publish their training set compounds and data. Concerning the quantum chemical Tehan model[16,17] QC, 417 of the 580 acids and 282 of the 563 bases belong to its training set (see also below), leaving 163 acids and 281 bases as truly external compounds. As discussed below in more detail, all currently collected 1143 organic acids and bases have also been used to explore the scope of the QC approach through recalibration (r-QC). In this way, comparative analysis of the performances of QC and r-QC provides information about limitations due to the previous smaller training set vs limitations resulting from the calculation methodology.

The molecular structures of our test set cover the atom types C, H, F, Cl, Br, I, N, O, and S, with dissociation and protonation being confined to −OH, >N−, and =N−, respectively. For the comparative analysis of the performance statistics of the different prediction methods, the compounds were subdivided into six acid subsets (phenols and aromatic as well as aliphatic carboxylic acids subdivided further according to substitution patterns) and nine base subsets (anilines, primary, secondary and tertiary amines, *meta/para*- and *ortho*-substituted pyridines, pyrimidines, imidazoles, and quinolines). In Tables 1 and 2, the respective compound subsets are listed together with the associated pK_a value ranges.

Table 1. Structural Characteristics and Subset-Specific pK_a Value Ranges of the Test Set of 580 Organic Oxygen Acids[a]

atom composition and compound class	n	pK_a min	pK_a median	pK_a max
CHO(X)	413	0.51	4.37	12.19
CHO(X) + N	561	0.38	4.41	12.23
CHON(X) + PS	580	0.38	4.37	12.23
m/*p*-phenols	79	5.43	9.38	11.47
o-phenols with intramolecular H-bond	29	3.03	7.10	9.87
o-phenols w/o intramolecular H-bond	116	0.38	7.85	12.23
m/*p*-aromatic carboxylic acids	70	2.82	3.99	5.03
o-aromatic carboxylic acids	90	0.65	2.95	5.09
aliphatic carboxylic acids	196	0.51	3.96	5.75
All				
phenols and aromatic + aliphatic carboxylic acids	580	0.38	4.37	12.23

[a] CHO(X) indicates that the compounds are built from carbon, hydrogen, oxygen, and possibly halogen; n = number of compounds; min and max = minimum and maximum experimental pK_a, respectively.

Table 2. Structural Characteristics and Subset-Specific pK_a Value Ranges of the Test Set of 563 Organic Nitrogen Bases[a]

atomic composition and compound class	n	pK_a min	pK_a median	pK_a max
CHN(X)	310	−2.86	5.90	11.72
CHN(X) + O	512	−5.00	5.49	11.72
CHON(X) + PS	563	−5.00	5.89	11.72
anilines	137	−5.00	3.61	6.57
primary amines	106	2.03	9.62	11.13
secondary amines	43	5.09	10.16	11.72
tertiary amines	85	4.64	8.72	11.25
m/*p*-pyridines	55	0.67	4.67	10.14
o-pyridines	38	−2.86	3.59	7.90
pyrimidines	16	−1.63	3.08	7.34
imidazoles	48	−0.81	5.85	10.30
quinolines	35	2.69	5.12	9.13
All				
anilines, amines, pyridines, pyrimidines, imidazoles, and quinolines	563	−5.00	5.89	11.72

[a] CHN(X) indicates that the compounds are built from carbon, hydrogen, nitrogen, and possibly halogen; n = number of compounds; min and max = minimum and maximum experimental pK_a of the (conjugated acids of the) bases, respectively.

A table with the experimental pK_a values used for the present study is given in the Supporting Information.

Computational Details. The quantum chemical model suite developed by Tehan et al.[16,17] employs molecular parameters derived from perturbational molecular orbital (MO) theory. For most of the compound classes listed above, linear regression models based on the electrophilic delocalizability (that had been called superdelocalizability,[16] although this latter term would in principle be confined to π-electron systems):[32]

$$D^{E}(A) = 2\sum_{i}^{occ}\sum_{\mu(A)}\frac{c_{\mu i}^{2}}{\varepsilon_i - \alpha} \qquad (8)$$

yielded the best statistics. In eq 8, ε_i denotes the energy of the i-th molecular orbital (MO), α is a reference energy[32]

that had been set to zero, $c_{\mu i}$ specifies the contribution of the μ-th atomic orbital (AO) to the i-th MO (the LCAO−MO coefficient, with LCAO = linear combination of AOs), A is the atomic site of interest where the relevant AOs μ are located, and the sum includes all (in the electronic ground state) doubly occupied MOs i.

For predicting pK_a, D^E was evaluated at the atoms involved in ionizing groups (see below and Tables 4 and 5), employing the semiempirical quantum chemical AM1[33] model including respective geometry optimization. The only additional variables were the Coulson net atomic charge Q_A for carboxylic acids and *ortho*-substituted pyridines, and for quinolines the energy of the lowest unoccupied molecular orbital (LUMO), E_{LUMO}, was used as the only descriptor.[16] For atom A, Q_A can be calculated as the difference between the core charge Z_A (which is the charge obtained at atom A when all its valence electrons are removed) and the actual amount of electronic charge at A as provided by the associated AOs according to the LCAO−MO expansion:[32]

$$Q_A = Z_A - 2\sum_{i}^{occ}\sum_{\mu(A)} c_{\mu i}^2 \tag{9}$$

For the present investigation, we have used both the original Tehan models (calibrated with 417 acids and 282 bases) as well as the recalibrated versions, employing the present test set of 1143 compounds as the training set.

SPARC[4,5] is based on parametrization of intra- and intermolecular interactions following respective concepts of physical organic chemistry, and includes a module for predicting pK_a. Here, the algorithm (though apparently not published in full detail) addresses resonance and electrostatic and steric effects as well as H-bonding and aqueous solvation:

$$pK_a = pK_a^0 + \sum_p \delta_p(pK_a) \tag{10}$$

In eq 10, pK_a denotes the pK_a of a functional group, pK_a^0 is its intrinsic pK_a value (if the functional group would stand alone), and δ_p(pK_a) quantifies the change from p$K_a{}^0$ to pK_a through intramolecular electrostatic (σ-inductive and through space), mesomeric, steric, and H-bond interactions as well as through aqueous solvation, all of which SPARC calls perturbations (hence the index p used for δ_p).[4,5] For our comparative analysis, the Web-based SPARC version has been used.[34]

ACD is a model suite based on Hammett-type LFERs, and is implemented in the commercial software ACD/pK_a DB (without providing the algorithm) included in ACD/Laboratories (version 12.0).[3] Here, pK_a prediction for a given compound proceeds through allocation to an automatically selected parent compound and subsequent quantification of substituent effects and further structural changes on the resultant pK_a:

$$pK_a = pK_a^0 + \sum_i \Delta_i(pK_a) \tag{11}$$

In eq 11, pK_a refers to the compound of interest, pK_a^0 quantifies the pK_a value of a suitable parent compound, and Δ_i(pK_a) represents the effect of substitutions and further structural changes on pK_a. ACD/pK_a is most likely a considerable extension of the LFER approach for predicting pK_a as summarized by Perrin et al.,[1] and thus includes electrostatic (inductive and field), mesomeric, and steric effects as well as intramolecular H-bonding. However, it seems that even the number of ACD model parameters (covering a range of parent compounds and associated LFERs) has not been published yet.

In the case of molecular species with two different protic sites and similar associated pK_a values, both ACD and SPARC address the resultant additional tautomeric equilibria through providing macroscopic pK_a's calculated from the initially predicted site-specific microscopic pK_a's (see also above),[1] and these macroscopic constants have then been used for comparison with experimental values.

Statistical Performance. The calibration performance was quantified through the squared correlation coefficient r^2:

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i^{fit} - y_i^{obs})^2}{\sum_{i=1}^{n}(y_i^{obs} - y^{mean})^2} \tag{12}$$

Here, y_i^{fit}, y_i^{obs}, and y^{mean} denote the regression-fitted and observed target values (in our case: pK_a) and the observed mean, and n is the number of respective data. In addition, the prediction performance of methods is assessed using the predictive squared correlation coefficient q^2:[35]

$$q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i^{pred} - y_i^{obs})^2}{\sum_{i=1}^{n}(y_i^{obs} - y^{mean})^2} \tag{13}$$

In eq 13, y_i^{fit} has been replaced by y_i^{pred}, the value predicted by (not fitted through) the regression model under investigation, and y^{mean} is the mean of observed values of the test set. As opposed to r^2, q^2 does not involve postprocessing in terms of least-squares rescaling of the predicted values but compares the untuned output of the original model to the experimental data. Note further that q^2 ranges from 1 (perfect agreement) to $-\infty$, with $q^2 = 0$ representing the case where the model prediction is as good as taking the average target value as the predicted value for all compounds. In addition, the following parameters have been used to further characterize the statistical performance: rms, mean error (me), systematic error (bias), maximum negative error (mne, largest underestimation), and maximum positive error (mpe, largest overestimation).

RESULTS AND DISCUSSION

Data Set Characteristics. As outlined above, the original data set of 699 compounds with 417 organic acids and 282 organic bases of the Tehan study[16,17] has been augmented to a total of 1143 compounds, now covering 580 organic acids and 563 organic bases with associated pK_a ranges from 0.38 to 12.23 and −5.00 to 11.72, respectively (see Tables 1 and 2). This extended data set provided opportunity for recalibrating the QC model suite of Tehan et al.[16,17] Accordingly, our comparative analysis includes both the

Table 3. Performance of the Tehan QC Model[16,17] for the Original Training Set and the Newly Added and Combined Data Sets[a]

	Tehan data set			additional data set				combined data set			
	n	r^2	rms	*n*	r^2	q^2	rms	*n*	r^2	q^2	rms
Acids											
m/*p*-phenols	58	0.92	0.32	21	0.78	0.77	0.61	79	0.87	0.87	0.42
o-phenols w H-bond	26	0.83	0.69	3	0.67	−8.91	2.32	29	0.76	0.75	0.90
o-phenols w/o H-bond	91	0.92	0.75	25	0.88	0.80	0.92	116	0.90	0.90	0.78
m/*p*-aromatic carboxylic acids	45	0.86	0.17	25	0.67	0.63	0.27	70	0.79	0.79	0.21
o-aromatic carboxylic acids	53	0.79	0.46	37	0.45	0.35	0.72	90	0.64	0.63	0.58
aliphatic carboxylic acids	141	0.80	0.39	55	0.73	0.72	0.53	196	0.78	0.78	0.43
Bases											
anilines	55	0.77	0.96	82	0.66	0.62	0.90	137	0.75	0.72	1.03
primary amines	23	0.86	0.40	83	0.71	0.50	1.06	106	0.73	0.53	0.99
secondary amines	23	0.42	0.95	20	0.85	0.42	1.31	43	0.78	0.42	1.12
tertiary amines	31	0.79	0.56	54	0.31	0.10	1.52	85	0.43	0.29	1.37
m/*p*-pyridines	45[b]	0.74	0.79	10	0.92	0.85	1.03	55	0.76	0.76	0.85
o-pyridines	32[b]	0.87	1.10	6	0.64	−0.02	2.67	38	0.79	0.76	1.40
pyrimidines	13	0.79	1.09	3	0.97	0.42	1.85	16	0.80	0.77	1.19
imidazoles and benzimidazoles	26	0.87	0.81	22	0.82	0.80	1.21	48	0.83	0.82	1.00
quinolines	27[b]	0.63	0.56	8	0.23	0.12	1.83	35	0.58	0.56	0.91

[a] Statistical parameters are: *n* = number of compounds, r^2 = squared correlation coefficient of calibration (eq 12), q^2 = predictive squared correlation coefficient (eq 13),[35] rms = root-mean-square error. [b] As outlined in the Materials and Methods Section, three *m*/*p*-pyridines (dinicotinic acid, nicotinic acid, and isonicotinic acid), two *o*-pyridines (picolinic acid and 2-hydroxy-pyridine), and one quinoline (8-quinolinol) were deleted from the original Tehan data set.

original Tehan (QC) models and their recalibrated updates (r-QC), which will be discussed below in more detail.

Table 1 shows that 413 of the 580 organic acids and 310 of the 563 organic bases are built from C, H, O, and possibly halogen (X), thus making up ca. 71% of the current acid test set. The N atom as a further atomic constituent applies for 148 acids (26%), and there are 19 acids (3%) containing P or S (or both). With regard to the 563 organic bases, 310 (55%) are built from C, H, N, and possibly halogen, 202 (36%) contain O, and 51 (9%) include P or S or both (see Table 2). At the same time, all acids are oxygen acids involving the dissociation from −OH attached to a phenyl ring (phenols) or as part of a carboxylic group attached to aliphatic or aromatic carbon (aliphatic and aromatic carboxylic acids), and the protonation site of the bases is >N− (aliphatic and aromatic amines) or =N− (N-heterocycles).

Performance of QC Model Suite for Extended Data Set. In Table 3, the performance of the original QC models is compared for the original set of 699 compounds, for the newly added compounds ($n = 444$), and for the extended current test set of 1143 compounds. With regard to the rms, the extended data set yields a reduced performance for all 15 class-specific models. The slightly increased r^2 for the extended set of *meta/para*-substituted pyridines ($n = 55$) as compared to the smaller original subset ($n = 48$) is probably caused by the associated substantial increase of the pK_a range covered (9.47 vs 5.80), the latter of which may have a substantial influence on the squared correlation coefficients.[35]

Note further that because the extended data set contains the compounds used for the original training set,[16,17] the observed similarity between r^2 and q^2 for most (but not all) extended subsets is not surprising. Restriction to the newly added 444 compounds (163 acids and 281 bases), however, yields q^2 values that now differ more substantially from the calibration r^2 values. In this case, q^2 reflects the actual (external) prediction performance,[35] keeping in mind the confounding factor of small data sets with small target value ranges as mentioned above. Taking *ortho*-substituted phenols without intramolecular H-bonding as an example, QC yields $r^2 = 0.92$ and rms = 0.75 for the original training set ($n = 91$),[16] $r^2 = 0.88$, $q^2 = 0.80$, and rms = 0.92 for the newly added 25 compounds, and $r^2 = q^2 = 0.90$ as well as rms = 0.78 for the combined extended set ($n = 116$; see Table 3).

Among the three newly added *ortho*-substituted phenols with intramolecular H-bonding, *o*-hydroxybenzaldehyde, and 4-amino-2-(5-chloro-1,3-benzoxazol-2-yl)phenol yield particularly large prediction errors of more than 1.7 pK_a units (*o*-hydroxybenzaldehyde: exp. pK_a = 8.34 vs QC-predicted pK_a = 5.61, 4-amino-2-(5-chloro-1,3-benzoxazol-2-yl)phenol: 9.81 vs 8.07), and the other compound shows still significant deviations between experimental and QC-predicted pK_a (*o*-hydroxyacetophenone: 9.19 vs 8.63). In this case, q^2 becomes even negative (−8.91), indicating that for these compounds the QC predictions are even inferior to taking just the data set average of pK_a in each case.[35] Because, however, ACD yielded reasonable results for these three compounds (8.18, 10.17, and 8.68), we decided to keep them in the extended data used for the overall comparative analysis as discussed below.

For bases, the QC performance is generally inferior to the ones for acids, as can be seen when comparing the upper and lower part of Table 3. Moreover, the decrease in QC performance when comparing the original training subsets and the newly added compounds is more pronounced for bases than for acids, which becomes particularly apparent when focusing on rms that is not confounded by the target value range.

The small subset of six newly added *ortho*-substituted pyridines yields the largest rms (2.67) among all subsets and also a slightly negative q^2 (−0.02), showing that the QC approach cannot handle these additional compounds properly despite reasonable statistics ($r^2 = 0.87$ and rms = 1.10) for the training set. Much better results, however, are obtained for anilines with rms values similar for the original training set (0.96 and $n = 55$), the additional 82 compounds (0.90), and the combined set (1.03 and $n = 137$; see Table 3).

Table 4. Regression Equations of the Recalibrated Quantum Chemical Model r-QC for Phenols and Carboxylic Acids Employing the Semiempirical AM1 Level of Calculation[a]

data set	n	regression equation
m/*p*-phenols	79	$pK_a = -5.30*D^E(O) - 40.72$
o-phenols with intra-H-bond	29	$pK_a = -11.08*D^E(O) - 96.74$
o-phenols w/o intra-H-bond	116	$pK_a = -7.69*D^E(O) - 63.74$
m/*p*-aromatic carboxylic acids	70	$pK_a = -2.09*D^E(=O) - 17.77$
o-aromatic carboxylic acids	90	$pK_a = -3.20*D^E(=O) - 30.18$
aliphatic carboxylic acids	196	$pK_a = -4.40*D^E(O) - 17.48*Q_{=O} - 43.48$
	141[b]	$pK_a = -4.14*D^E(O) - 19.34*Q_{=O} - 44.28$ **(adjusted original model)**

[a] D^E denotes the electrophilic or donor delocalizability (eq 8, with MO energies expressed in eV), evaluated either at the oxygen carrying the acidic H atom (O) or at the carbonyl oxygen of the acidic carboxylic group (=O), $Q_{=O}$ is the Coulson net atomic charge (eq 9, atomic units) of this carbonyl oxygen, and n = number of compounds. For the performance statistics, see Table 6. [b] Calibration to the original Tehan training set[16] of 141 carboxylic acids (see text).

Recalibration of Quantum Chemical Model Suite. Table 4 lists the regression coefficients of the r-QC models derived for the six subsets of the 580 organic acids, and Table 5 provides the respective information for the nine subsets of 563 organic bases. The associated statistics are presented and discussed below for all subsets as well as for the total test set of 1143 compounds when comparing the prediction performances of the four methods QC (Tehan et al.),[16,17] r-QC, SPARC, and ACD.

With regard to the subset of 196 carboxylic acids, the original QC model[16] could not be reproduced in terms of predicted pK_a values and associated performance statistics. Therefore, Table 4 contains both the r-QC model and a respective second new calibration, this time confined to the original subset[16] of 141 carboxylic acids.

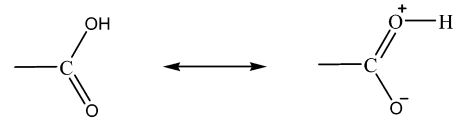
Inspection of Table 4 reveals further that D^E (eq 8) is evaluated in most cases at the O atom carrying the acidic H. For the two subsets of aromatic carboxylic acids, D^E at the carbonyl oxygen of the carboxylic group turned out to provide superior statistics, and for the aliphatic carboxylic acids, both $D^E(O)$ and $Q_{=O}$ (net atomic charge at the carbonyl O of the carboxylic group; see eq 9) are used. Because D^E (a kinetic property) has been designed to quantify the susceptibility of a molecule for a covalent attack by an electrophile (and thus characterizes the nucleophilicity or donor delocalizability of the molecule of interest),[32] its suitability to model pK_a (a thermodynamic property) might be surprising at first sight. However, the negative sign of its regression coefficient combined with its negative value (because of the negative sign of the occupied MO energies ε_i, see eq 8) indicates that increasing donor delocalizability D^E correlates with increasing pK_a and thus decreasing acidity (or increasing bond strength A−H).

For the organic bases, D^E is evaluated at the relevant nitrogen in eight of the nine subsets (see Table 5). Again, decreasing the negative value of D^E and thus decreasing donor delocalizability decreases pK_a and thus increases acidity, which in this case concerns the conjugated acid BH^+ of base B. Accordingly, increasing (the negative value of) donor delocalizability (or nucleophilicity) D^E now correlates with increasing basicity.

For the subset of 35 quinolines, E_{LUMO} turned out to be the best single descriptor (see Table 5). E_{LUMO} has a positive sign and is inversely related to electron affinity.[32] The observed correlation between E_{LUMO} and pK_a thus indicates that increasing electron affinity of base B increases the acidity of BH^+ (note that the dissociation of BH^+ increases the electron density at B) and correspondingly decreases the basicity of B.

In two cases, the net atomic charge Q turned out to be a suitable second descriptor. Interestingly, however, the sign of its regression coefficients differs between carboxylic acids and *ortho*-substituted pyridines (Tables 4 and 5, respectively), although the net atomic charge is negative in both cases. For the carboxylic acids, increasingly negative $Q_{=O}$ increases pK_a and thus decreases acidity. This might be explained by considering the following mesomeric stabilization:

Scheme 2. Mesomeric Stabilization of Carboxylic Acids



According to Scheme 2, a larger negative charge at the carbonyl oxygen corresponds to a smaller weight of the acidic −OH form in the electronic structure and thus to a smaller

Table 5. Regression Equations of the Recalibrated Quantum Chemical Model r-QC for Anilines, Amines, and Heterocycles Employing the Semiempirical AM1 Level of Calculation[a]

data set	n	regression equation
anilines	137	$pK_a = -7.56*D^E(N) - 64.14$
primary amines	106	$pK_a = -5.68*D^E(N) - 43.96$
secondary amines	43	$pK_a = -7.66*D^E(N) - 62.10$
tertiary amines	85	$pK_a = -5.11*D^E(N) - 38.89$
m/*p*-pyridines	55	$pK_a = -8.34*D^E(N) - 68.78$
o-pyridines	38	$pK_a = -12.25*D^E(N) + 137.85*Q(N) - 99.29$
pyrimidines	16	$pK_a = -4.81*D^E(N) - 40.58$
imidazoles and benzimidazoles	48	$pK_a = -7.51*D^E(N) - 62.65$
quinolines	35	$pK_a = 6.38*E_{LUMO} + 8.34$

[a] D^E denotes the electrophilic or donor delocalizability (eq 8 with MO energies expressed in eV), evaluated at the basic nitrogen (N), Q_N is the Coulson net atomic charge (eq 9, atomic units) of this nitrogen atom, E_{LUMO} the energy (in eV) of the LUMO of the electronic ground state, and n = number of compounds. For the performance statistics, see Table 6.

acidity of the carboxylic group. For the *ortho*-substituted pyridines, increasingly negative Q_N now decreases pK_a and thus increases acidity, which would be in accord with the conventional electrostatic interpretation that charge separation facilitates dissociation.

Overall Performance Statistics. Table 6 provides performance statistics in terms of r^2, q^2, rms, bias, me, mne, and mpe for all four methods (QC, r-QC, SPARC, and ACD), as applied to the total sets of 580 acids and 563 bases as well as to the respective 6 and 9 subsets, respectively.

Before proceeding with the results, a note of caution with regard to the overall statistics appears appropriate. As can be seen from r^2 and q^2 listed for QC and r-QC (and to a lesser degree also for SPARC and ACD), the overall statistics (e.g., r-QC: acids: $r^2 = q^2 = 0.96$; bases: $r^2 = q^2 = 0.92$) are significantly superior to each of the method-specific subset statistics, which holds for both acids and bases (r-QC: acids: r^2 range = 0.64 − 0.90; bases: r^2 range =0.42 − 0.82). The reason is that when combining the subsets to a total set (of acids or bases), the local subset errors become less important with regard to the enlarged total pK_a range and data distribution. Thus the total acid or base set statistics, though formally calculated correctly, do not reflect properly the actual model performance. The latter differs for the different subsets and can be appreciated only in a subset-specific manner. An attempt to address this issue is summarized as r^2 and q^2 with the specification "subset-weighted performance" in Table 6, now calculated as averages of the corresponding subset-specific results and weighted according to the respective subset sizes (numbers of compounds). These latter (modified) global r^2 and q^2 values appear more informative, although they do not conform to the calculation procedure foreseen for a single (in our case combined) data set.

Considering the accordingly modified global q^2 values, ACD shows up as the overall best-performing method that is significantly superior to SPARC (acids: modified q^2 0.91 vs 0.79; bases: 0.88 vs 0.86), which in turn is again significantly superior to both r-QC (acids: 0.80; bases: 0.71) and QC (acids: 0.79; bases: 0.56). Indeed, this order of method performance results also from inspection of the global rms values (acids: 0.41 vs 0.42 (without the 29 *ortho*-substituted phenols with intramolecular H-bonding, see legend of Table 6) vs 0.55 vs 0.55; bases: 0.65 vs 0.71 vs 0.92 vs 1.07), and it reflects also the overall result when comparing all subset-specific performances.

For an unknown reason, ACD could not handle methamphetamine. Accordingly, the ACD statistics for bases refers to 562 (instead 563) compounds. Figures 1 and 2 show the calculated vs experimental pK_a data distributions for 580 oxygen acids and 562 nitrogen bases for ACD, and the associated prediction error vs experimental data plots are given in Figures 3 and 4, respectively. The latter two reveal more clearly outlying data and demonstrate that the overall error distributions show no systematic trend with regard to pK_a.

Table 6. Statistical Performance for Predicting pK_a of the Two Quantum Chemical Models QC (Tehan et al.)[16,17] and r-QC and of SPARC and ACD for the Test Set of 580 Oxygen Acids and 563 Nitrogen Bases[a]

data set	method	n	r^2	q^2	rms	bias	me	mne	mpe
acids	QC	580	0.96	0.96	0.55	−0.01	0.39	−2.73	2.59
	r-QC	580	0.96	0.96	0.55	−0.01	0.39	−2.53	2.54
	SPARC	580 (551)[b]	0.96	0.95	0.58 (0.42)[b]	0.14	0.36	−1.22	2.88
	ACD	580	0.98	0.98	0.41	0.05	0.22	−3.32	1.97
acids (subset-weighted performance)[c]	QC	580	0.79	0.79					
	r-QC	580	0.80	0.80					
	SPARC	580	0.88	0.79					
	ACD	580	0.92	0.91					
m/*p*-phenols	QC	79	0.87	0.87	0.42	0.00	0.30	−1.01	1.75
	r-QC	79	0.87	0.87	0.42	0.00	0.30	−1.01	1.72
	SPARC	79	0.93	0.93	0.31	−0.05	0.21	−1.06	0.74
	ACD	79	0.95	0.95	0.25	−0.04	0.17	−0.91	0.61
o-phenols with intra-H-bond	QC	29	0.76	0.75	0.90	−0.17	0.62	−2.73	1.00
	r-QC	29	0.77	0.77	0.87	0.00	0.64	−2.53	1.22
	SPARC	29	0.84	−0.10	1.88	1.66	1.69	−0.33	2.88
	ACD	29	0.86	0.62	1.11	0.82	0.97	−1.13	1.92
o-phenols w/o intra-H-bond	QC	116	0.90	0.90	0.78	0.05	0.62	−1.58	2.59
	r-QC	116	0.90	0.90	0.78	0.00	0.63	−1.61	2.54
	SPARC	116	0.95	0.95	0.55	0.10	0.32	−1.17	2.42
	ACD	116	0.96	0.96	0.52	0.11	0.28	−3.32	1.97
m/*p*-aromatic carboxylic acids	QC	70	0.79	0.79	0.21	0.01	0.15	−0.46	0.78
	r-QC	70	0.80	0.80	0.20	−0.00	0.15	−0.44	0.75
	SPARC	70	0.89	0.68	0.26	−0.20	0.23	−0.47	0.48
	ACD	70	0.92	0.92	0.12	0.00	0.07	−0.33	0.69
o-aromatic carboxylic acids	QC	90	0.64	0.63	0.58	−0.09	0.42	−2.01	1.45
	r-QC	90	0.64	0.64	0.57	0.00	0.40	−1.90	1.56
	SPARC	90	0.77	0.74	0.49	0.12	0.37	−1.22	1.66
	ACD	90	0.87	0.87	0.35	−0.05	0.23	−1.12	1.02
aliphatic carboxylic acids	QC	196	0.78	0.78	0.43	0.00	0.32	−1.53	1.96
	r-QC	196	0.78	0.78	0.44	0.00	0.33	−1.48	1.99
	SPARC	196	0.86	0.83	0.38	0.13	0.29	−1.04	1.26
	ACD	196	0.92	0.92	0.26	0.00	0.15	−1.69	1.24
bases	QC	563	0.89	0.89	1.07	−0.12	0.79	−4.81	3.43
	r-QC	563	0.92	0.92	0.92	0.00	0.70	−3.45	3.52
	SPARC	563	0.96	0.95	0.71	−0.14	0.43	−6.32	2.53
	ACD	562	0.96	0.96	0.65	−0.05	0.36	−5.23	2.91
bases (subset- weighted performance)[c]	QC	563	0.70	0.56					
	r-QC	563	0.71	0.71					
	SPARC	563	0.89	0.86					
	ACD	562	0.89	0.88					
anilines	QC	137	0.75	0.72	1.03	0.00	0.79	−2.42	3.02
	r-QC	137	0.77	0.77	0.93	0.00	0.73	−2.02	2.48
	SPARC	137	0.97	0.96	0.41	−0.06	0.28	−1.18	1.38
	ACD	137	0.95	0.95	0.44	−0.02	0.21	−2.57	2.56
primary amines	QC	106	0.73	0.53	0.99	−0.20	0.75	−2.92	1.64
	r-QC	106	0.74	0.74	0.74	0.00	0.54	−1.68	3.05
	SPARC	106	0.84	0.84	0.59	−0.06	0.41	−0.91	2.53
	ACD	106	0.85	0.84	0.57	0.11	0.36	−0.97	2.91
secondary amines	QC	43	0.78	0.42	1.12	−0.68	0.87	−2.40	0.98
	r-QC	43	0.77	0.77	0.70	0.00	0.58	−1.58	1.34
	SPARC	43	0.89	0.89	0.49	−0.03	0.39	−0.92	1.30
	ACD	42	0.90	0.90	0.47	0.00	0.33	−0.87	1.49
tertiary amines	QC	85	0.43	0.04	1.30	−0.16	0.93	−4.57	3.43
	r-QC	85	0.42	0.42	1.01	0.00	0.80	−3.24	2.60
	SPARC	85	0.85	0.79	0.74	−0.22	0.53	−1.96	2.06
	ACD	85	0.91	0.91	0.49	−0.04	0.34	−1.17	1.88
m/*p*-pyridines	QC	55	0.76	0.76	0.85	−0.07	0.60	−2.33	2.76
	r-QC	55	0.76	0.76	0.85	0.00	0.57	−2.04	3.17
	SPARC	55	0.91	0.89	0.57	−0.18	0.36	−3.01	0.88
	ACD	55	0.91	0.90	0.56	−0.17	0.31	−2.46	0.98
o-pyridines	QC	38	0.79	0.76	1.40	−0.07	0.96	−4.81	3.13
	r-QC	38	0.80	0.80	1.28	−0.06	1.01	−3.45	3.52
	SPARC	38	0.86	0.76	1.41	−0.31	0.53	−6.32	0.64
	ACD	38	0.90	0.82	1.21	−0.34	0.47	−5.23	1.21
pyrimidines	QC	16	0.80	0.77	1.19	−0.04	0.93	−2.39	2.22
	r-QC	16	0.78	0.78	1.15	−0.00	0.90	−2.23	1.94
	SPARC	16	0.91	0.83	1.00	−0.13	0.64	−3.31	0.72
	ACD	16	0.98	0.97	0.40	0.14	0.30	−0.74	0.89
imidazoles and benzimidazoles	QC	48	0.83	0.82	1.00	0.16	0.80	−2.13	2.67
	r-QC	48	0.82	0.82	1.00	0.00	0.81	−2.20	2.65
	SPARC	48	0.89	0.81	1.04	−0.40	0.73	−3.06	1.19
	ACD	48	0.80	0.80	1.10	−0.16	0.75	−2.63	2.88
quinolines	QC	35	0.58	0.56	0.91	−0.14	0.58	−3.29	2.22
	r-QC	35	0.58	0.58	0.89	−0.00	0.58	−3.18	2.47
	SPARC	35	0.86	0.85	0.53	−0.11	0.32	−1.22	1.64
	ACD	35	0.79	0.73	0.72	−0.21	0.45	−1.81	1.36

[a] The statistical parameters are: n = number of compounds; r^2 = squared correlation coefficient (eq 12); q^2 = predictive squared correlation coefficient (eq 13); rms = root-mean-square error; bias = systematic error, me = mean absolute error; mne = maximum negative error (largest underestimation); mpe = maximum positive error (largest overestimation). [b] The poor SPARC performance for *o*-substituted phenols with intramolecular H-bonding affects the overall SPARC statistics significantly. Without this subset of 29 compounds, the SPARC performance for the remaining 551 organic oxygen is: r^2 = 0.98, q^2 = 0.98, rms = 0.42, me = 0.06, mne = −1.22, and mpe = 2.42. [c] Calculated as average from subset-specific r^2 and q^2 and weighted according to subset size (see text).

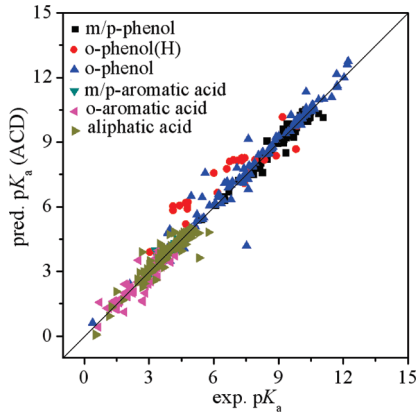


Figure 1. Predicted vs experimental pK_a for ACD applied to 580 organic oxygen acids, covering 6 subsets of phenols and carboxylic acids (*o*-phenol(H) represents *ortho*-substituted phenols with intramolecular H bonding).

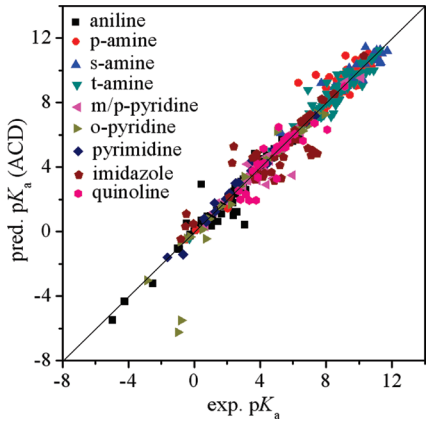


Figure 2. Predicted vs experimental pK_a for ACD applied to 562 organic nitrogen bases, covering 9 subsets of anilines, amines, and N-heterocycles.

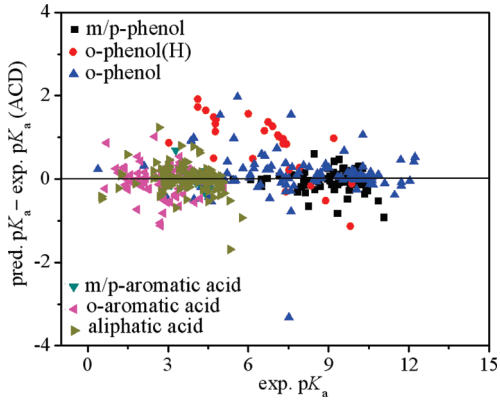


Figure 3. Prediction error vs experimental pK_a for ACD applied to 580 organic oxygen acids, covering 6 subsets of phenols and carboxylic acids (*o*-phenol(H) represents *ortho*-substituted phenols with intramolecular H bonding).

When comparing the acid and base statistics, the overall rms values indicate that all methods perform better for acids. This is particularly pronounced for QC (rms = 0.55 vs 1.07) and r-QC (0.55 vs 0.92) but applies also to SPARC (0.58 vs 0.71) and ACD (0.41 vs 0.65).

QC and r-QC Performances. Surprisingly, comparison of QC (original Tehan model suite)[16,17] and r-QC (recalibrated QC model suite) reveals that despite the substantially extended data set used for recalibration, r-QC yields an only marginal improvement over QC (Table 6). On the one hand,

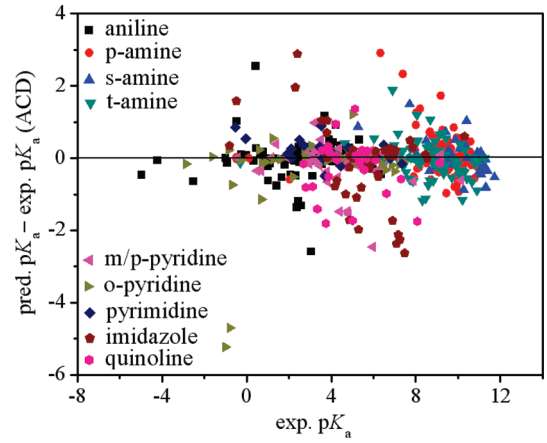


Figure 4. Prediction error vs experimental pK_a for ACD applied to 562 organic nitrogen bases, covering 9 subsets of anilines, amines and N-heterocycles.

this finding indicates a substantial robustness of the model suite developed by Tehan et al.,[16,17] keeping in mind that the currently used training set of 1143 compounds is larger by 444 compounds (163 acids and 281 bases) than the original training set of 699 compounds. On the other hand, the still substantially inferior performance as compared to SPARC and ACD (except for individual subsets) shows that the scope of this approach appears to be limited.

Note, however, that QC and r-QC employ the semiempirical AM1 scheme and thus may offer room for improvement through application of higher-level quantum chemical methods (then, however, requiring significantly more computation time). Another route for possible improvement is the exploration of other quantum chemical parameters, such as the local molecular parameters employed recently for modeling the H-bond donor and acceptor strengths of organic compounds.[36−38]

Coming back to the present level of QC and r-QC performance, the smallest rms values (0.21 and 0.20) are obtained for the subset of 70 *meta/para*-substituted aromatic carboxylic acids (where r-QC is competitive to SPARC), and rms values below 0.5 are observed for the 79 *meta/para*-substituted phenols and the 196 aliphatic carboxylic acids, the next best-performing subset being the 90 *ortho*-substituted aromatic carboxylic acids (rms 0.58 and 0.57). Among the oxygen acids, *ortho*-substituted phenols with intramolecular H-bonding yield the largest rms (0.90 and 0.87), but here QC and r-QC outperform both SPARC and ACD (see below).

With regard to the 563 nitrogen bases, the 55 *meta/para*-substituted pyridines yield the smallest rms for QC and r-QC (0.85), which however is still larger than for SPARC and ACD (0.57 and 0.56, respectively); note, however, the larger biases of the latter two methods for this subset (−0.18 and −0.17; see Table 6). The next best-performing subsets with QC and r-QC rms values around 0.9 are the 35 quinolines and the 106 primary amines (where rms for r-QC is much lower, 0.74), and for the following subsets, rms values around one pK_a unit are obtained: 137 anilines, 48 imidazoles and benzimidazoles, 43 secondary amines (where however r-QC is significantly superior to QC), and 16 pyrimidines.

Interestingly, imidazoles and benzimidazoles are the only base subset where QC and r-QC outperform (however only slightly) both SPARC and ACD, also with regard to biases

and maximum prediction errors. Moreover, r-QC outperforms SPARC for *ortho*-pyridines with regard to both rms (1.28 vs 1.40) and the bias (−0.06 vs −0.31; see Table 6), with a larger rms but smaller bias as compared to ACD (rms 1.21 and bias −0.34).

Subset-Specific Issues. SPARC and ACD yield the overall largest rms values with regard to acids (1.88 and 1.11; see Table 6) and also relatively large (and typically positive) pK_a prediction errors for *ortho*-substituted phenols with intramolecular H-bonding. Moreover, for this latter subset of 29 compounds, the by far largest positive biases (pK_a overestimations) are observed for both SPARC (1.66 pK_a units; see Table 6) and ACD (0.82). These findings suggest that both SPARC and ACD appear to overestimate the pK_a-raising (acidity-reducing) effect of intramolecular H bonding of the acidic H atom. Moreover, these prediction errors decrease with decreasing pK_a.

QC and r-QC do not show respective biases, indicating that the quantum chemical donor delocalizability D^E evaluated at the O atom carrying the acidic H (see eq 8 and Table 4) reflects the impact of intramolecular H bonding properly. While QC and r-QC are thus superior to SPARC and ACD for *ortho*-substituted phenols, their rms values (0.90 and 0.87) are nevertheless larger for these compounds than for all other subsets.

Interestingly, *ortho*-substituted phenols without intramolecular H bonds (116 compounds) show the second largest rms values for all four methods. This result appears to reflect the difficulty in modeling *ortho*-substituent effects, as is known since a long time for LFERs (that underlie both ACD and SPARC), but apply also to other structure-based calculation schemes. Indeed, *meta/para*-substituted phenols yield much better performances with rms values of 0.25 (ACD) to 0.42 (QC). Note further that with regard to aromatic carboxylic acids, the *ortho*-substituted subset of 90 compounds shows again lower prediction qualities (rms 0.35−0.57) than the *meta/para*-substituted subset of 70 compounds (rms 0.12−0.26).

The overall best-performing acid classes are the subsets of 196 aliphatic carboxylic acids (rms 0.26−0.43), the 79 *meta/para*-substituted phenols, and the just mentioned 70 *meta/para*-substituted aromatic carboxylic acids. For all acid subsets except *ortho*-substituted phenols with intramolecular H-bonding (see above), ACD yields the smallest rms.

Among the bases, the overall largest SPARC and ACD rms values (1.21 and 1.41) are obtained for *ortho*-substituted pyridines (n = 38), and in this case, QC and r-QC yield similar rms results (1.40 and 1.21). As with aromatic oxygen acids, the *meta/para*-substituted subset (n = 55) shows a much better performance for all four methods (rms 0.56−0.85), again reflecting the difficulty in addressing both steric and electronic effects of *ortho* substitution properly. The latter is also seen by the fact that both SPARC and ACD yield their by far largest pK_a underestimations for *ortho*-substituted pyridines (mne −6.32 and −5.23; see second to last column of Table 6), and here also show the largest negative bias (−0.34 and −0.31; see fourth to last column of Table 6). Systematic underestimation of the pK_a of (the conjugated acids of) bases means that their basicity is underestimated, suggesting an overestimation by SPARC and ACD of the degree of steric hindrance caused by *ortho* substitution.

In contrast to SPARC and ACD, QC and r-QC do not address any steric effect of *ortho* substitution but are confined to electronic effects without consideration of the spatial availability of the site of dissociation or protonation (see eqs 8 and 9 and Tables 4 and 5). From this viewpoint it is remarkable that they show only marginal biases (−0.07 and −0.06) for this subset of 38 *ortho*-substituted pyridines.

For the 16 pyrimidines, ACD performs significantly better (rms = 0.40) than do SPARC, r-QC, and QC (rms 1.00, 1.15, and 1.19), while all 4 methods provide similar rms values of about 1 pK_a unit for the 48 imidazoles and benzimidazoles. The overall best performances are observed for the subsets of 137 anilines (rms 0.44−1.03) and 43 secondary amines (rms 0.47−1.12), keeping in mind that ACD shows the smallest rms values for most base subsets, the only exceptions being anilines (where SPARC is slightly superior) and imidazoles and benzimidazoles (where QC and r-QC are in fact competitive to both SPARC and ACD as noted above).

Outliers. The maximum positive and negative errors of pK_a prediction (mpe and mne) are listed in the last two columns of Table 6. The largest acid outliers are observed for the subset of *ortho*-substituted phenols (without and with intramolecular H-bonding), with pK_a over- or underpredictions of 2−3 units for all 4 methods.

The outliers with the largest method-specific overpredictions of acid pK_a values are the following compounds: hexachlorophene (exp. pK_a = 4.85) for QC and r-QC; 3,5-dibromo-*N*-(4-chloro-2-nitrophenyl)-2-hydroxybenzamide (exp. pK_a = 4.11) for SPARC; and 2,2-methanediyl-bis(4,6-dichlorophenol)(exp. pK_a = 5.60) for ACD. Largest pK_a underestimations are obtained for the following compounds: *o*-hydroxybenzaldehyde (exp. pK_a = 8.34) for QC and r-QC; pentafluorobenzoic acid (exp. pK_a = 2.72) for SPARC; and 4-bromo-2,6-dichlorophenol (exp. pK_a = 6.21) for ACD.

For bases, the maximum prediction errors are much larger than for acids. Here, the largest QC and r-QC overestimations of pK_a are observed for tertiary amines and *ortho*-substituted pyridines (3.43 and 3.52 pK_a units, respectively), and the respective largest pK_a underestimations are obtained for tertiary amines and *ortho*-substituted pyridines (by up to 4.8 pK_a units). With SPARC, primary amines and *ortho*-substituted pyridines yield the largest over- and underestimations of pK_a (2.53 and −6.32), and the largest positive and negative ACD prediction errors are seen for primary amines (2.91) and *ortho*-substituted pyridines (−5.23).

The following bases yield the largest overestimations: *N*-[(8-*tert*-butyl-1,4-dioxaspiro[4.5]dec-2-yl)methyl]-*N*-ethylpropan-1-amine (exp. pK_a = 6.9) for QC; 2,6-dichloropyridine (exp. pK_a = −2.68) for r-QC; tryptophan (exp. pK_a = 7.38) for SPARC; and 4-[(4-methoxyphenyl)sulfonyl]furan-2-sulfonamide (exp. pK_a = 6.32) for ACD. Largest negative prediction errors of the pK_a (of the conjugated acids) of bases are obtained for 5-nitropyridine-2-amine (pK_a = 2.78) for QC and r-QC and pentachloropyridine (pK_a = −1.00) for SPARC and ACD.

Coming back to overall trends, the numbers of compounds with prediction errors exceeding 1 order of magnitude are as follows: QC: 39 acids (6.7%) and 146 bases (25.9%); r-QC: 38 acids (6.6%) and 131 bases (23.3%); SPARC: 39 acids (6.7%) and 43 bases (7.6%); and ACD: 25 acids (4.3%) and 48 bases (8.5%), respectively. Remarkably, more than half (23 and 14) of the acid outliers of SPARC and ACD

Table 7. Intercorrelation of pK_a Prediction Errors of the Two Quantum Chemical Models QC and r-QC and of SPARC and ACD in Terms of r^2 Values for the Test Set of 580 Oxygen Acids and 563 Nitrogen Bases[a]

	acids				bases			
method	QC	r-QC	SPARC	ACD	QC	r-QC	SPARC	ACD
QC	1.00				1.00			
r-QC	0.89	1.00			0.77	1.00		
SPARC	0.10	0.11	1.00		0.03	0.02	1.00	
ACD	0.15	0.15	0.46	1.00	0.06	0.06	0.31	1.00

[a] ACD intercorrelation statistics for organic bases refer to 562 compounds, because ACD could not handle methamphetamine.

are *ortho*-substituted phenols capable of forming internal H-bonds, and almost all of the respective pK_a values are overpredicted (see above).

When taking 2*rms as the threshold for the method-specific outliers, the following numbers of outliers are obtained: 30 acids (5.2%, rms = 0.55) and 32 bases (5.7%, rms = 1.07) for QC; 33 acids (5.7%, rms = 0.55) and 35 bases (6.2%, rms = 0.92) for r-QC; 31 acids (5.3%, rms = 0.58) and 21 bases (3.3%, rms = 0.71) for SPARC; and 39 acids (6.7%, rms = 0.41) and 32 bases (5.7%, rms = 0.65) for ACD.

Consensus Modeling Approach. Following the results of the current comparative analysis, the ACD methodology for predicting pK_a appears to be superior to SPARC as well as to the semiempirical quantum chemical approach (QC and r-QC) based mainly on local donor delocalizability. At the same time, there are also compounds with substantial ACD prediction errors up to 5 orders of magnitude. From this viewpoint, a consensus modeling approach looks attractive, where different methods are combined for predicting the property of interest (in our case: pK_a). If the different method predictions agree within an acceptable range for a given compound and if these methods differ (sufficiently) in their algorithms, then the overall confidence in the predicted value range increases. Conversely, if significantly different predictions are obtained, then this can be taken as an indication of problems that require further investigation.

Coming back to the four pK_a prediction schemes under current investigation, ACD, SPARC, and the quantum chemical approach, QC and r-QC, differ significantly with respect to their methodologies. As outlined above, ACD[3] is a model suite based on Hammett-type LFERs[1] with class-specific models of the general type shown in eq 11 (apparently without having published the model parameters). SPARC,[5] though also based on the LFER approach, has developed its own additive-constitutive scheme of quantifying resonance, electrostatic, steric, H-bonding, and aqueous solvation effects on pK_a for the relevant functional groups subject to dissociation or protonation, as formally summarized in eq 10. It is thus very likely that SPARC and ACD differ substantially in their model parameters and equations. A third method type is given by QC and r-QC that is based on parameters (and mainly on the donor delocalizability D^E, see Table 3) calculated at the semiempirical quantum chemical AM1 level from (optimized) 3D molecular geometries of the compounds in their electronic ground state.

Thus, the four methods can be grouped into three method types, providing opportunity for their application as parts of the consensus modeling approach. However, a straightforward means to check the actual degree of concordance (method prediction interdependence) is to evaluate the correlation in terms of r^2 values between the method-specific prediction errors for the current test set of 580 oxygen acids and 563 nitrogen bases. The respective intercorrelation results are summarized in Table 7.

As can be seen from the table, the QC and r-QC prediction errors have little to do with the ones of ACD and SPARC for both acids (r^2, 0.10−0.15) and bases (0.02−0.06) but show a significant intercorrelation with each other (acids: 0.89 and bases: 0.77). However, because of the overall significantly inferior performance of QC and r-QC, these methods would not be recommended as a general-purpose prediction tool. Instead, following Table 6, QC and r-QC are competitive for a few distinct compound classes where they partly outperform even ACD (see above).

The intercorrelation between ACD and SPARC is larger for acids (0.46) than for bases (0.31) but overall sufficiently low to consider the two method predictions as essentially independent from each other. The present findings thus suggest to use both ACD and SPARC, possibly augmented by r-QC in case of specific compound classes, as the general consensus strategy for predicting pK_a from molecular structures.

CONCLUSIONS

The present test set of 1143 organic compounds covering 580 oxygen acids and 563 nitrogen bases with an overall pK_a range from −5.00 to 12.23 provides a sound basis for evaluating the performance of pK_a prediction methods. Among the four methods under investigation, ACD (based on Hammett-type linear free-energy relationships) turned out to be significantly superior to SPARC, which in turn outperforms two different calibrations of a semiempirical quantum chemical AM1 approach based mainly on site-specific evaluations of the donor delocalizability. Moreover, the overall prediction performance is better for acids than for bases. However, the variation in root-mean-square errors (rms) across all six acid (phenols and aliphatic and aromatic carboxylic acids) and nine base (anilines, aliphatic amines, and aromatic N-heterocycles) subsets is substantial for all methods including ACD, and for the latter ranges from 0.12 to 1.11 pK_a units for acids and from 0.40 to 1.21 for bases. Accordingly, a consensus modeling approach is suggested through combined application of ACD and SPARC, thus increasing the level of confidence in the case of (sufficiently) similar predictions. Pitfalls for both ACD and SPARC concern *ortho* substitution of aromatics in general and intramolecular H-bonding accompanying *ortho* substitution in particular. With regard to the currently overall inferior quantum chemical approach (that is however competitive for a few subsets including *ortho*-substituted phenols), routes

for possible improvement include exploration of ab initio levels of calculation as well as local molecular parameters that have recently proven useful for modeling H-bond donor and acceptor strengths.[36−38]

ACKNOWLEDGMENT

Financial support was provided by the China Scholarship Council and the European Commission through the Integrated Project OSIRIS (contract no. 037017), which is gratefully acknowledged.

Supporting Information Available: A table listing all 580 acids and 563 bases with experimental and predicted values of pK_a with ACD, SPARC, QC, and r-QC methods is provided as additional information. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. *pK_a Prediction for Organic Acids and Bases*; Chapman & Hall: London, 1981.
(2) Lee, A. C.; Crippen, G. M. Predicting pK_a. *J. Chem. Inf. Model.* **2009**, *49*, 2013–2033.
(3) *ACD/Labs*, version 12.0; Advanced Chemistry Development Inc.: Toronto, Ontario, Canada, 2009.
(4) Karickhoff, S. W.; Mcdaniel, V. K.; Melton, C.; Vellino, A. N.; Nute, D. E.; Carreira, L. A. Predicting Chemical Reactivity by Computer. *Environ. Toxicol. Chem.* **1991**, *10*, 1405–1416.
(5) Hilal, S. H.; Karickhoff, S. W.; Carreira, L. A. A Rigorous Test for SPARC's Chemical Reactivity Models: Estimation of More Than 4300 Ionization pK_as. *Quant. Struct.-Act. Relat.* **1995**, *14*, 348–355.
(6) Hilal, S. H.; Karickhoff, S. W.; Carreira, L. A. Prediction of Chemical Reactivity Parameters and Physical Properties of Organic Compounds from Molecular Structure Using SPARC; EPA/600/R-03/030 March 2003; US Environmental Protection Agency, National Exposure Research Laboratory, Office of Research and Development: Research Triangle Park, NC, 2003.
(7) Meloun, M.; Bordovska, S. Benchmarking and Validating Algorithms That Estimate pK_a Values of Drugs Based on Their Molecular Structures. *Anal. Bioanal. Chem.* **2007**, *389*, 1267–1281.
(8) Liao, C. Z.; Nicklaus, M. C. Comparison of Nine Programs Predicting pK_a Values of Pharmaceutical Substances. *J. Chem. Inf. Model.* **2009**, *49*, 2801–2812.
(9) Schüürmann, G. Modelling pK_a of Carboxylic Acids and Chlorinated Phenols. *Quant. Struct.-Act. Relat.* **1996**, *15*, 121–132.
(10) Schüürmann, G.; Cossi, M.; Barone, V.; Tomasi, J. Prediction of the pK_a of Carboxylic Acids Using the Ab initio Continuum-Solvation Model PCM-UAHF. *J. Phys. Chem. A* **1998**, *102*, 6706–6712.
(11) 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.
(12) Takano, Y.; Houk, K. N. Benchmarking the conductor-like polarizable continuum model (CPCM) for aqueous solvation free energies of neutral and ionic organic molecules. *J. Chem. Theory Comput.* **2005**, *1*, 70–77.
(13) Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M. E. First Principles Calculations of Aqueous pK_a Values for Organic and Inorganic Acids Using COSMO-RS Reveal an Inconsistency in the Slope of the pK_a Scale. *J. Phys. Chem. A* **2003**, *107*, 9380–9386.
(14) Eckert, F.; Klamt, A. Accurate Prediction of Basicity in Aqueous Solution with COSMO-RS. *J. Comput. Chem.* **2006**, *27*, 11–19.
(15) Schüürmann, G. Quantum Chemical Analysis of the Energy of Proton Transfer from Phenol and Chlorophenols to H_2O in the Gas Phase and in Aqueous Solution. *J. Chem. Phys.* **1998**, *109*, 9523–9528.
(16) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Montana, J. G.; Manallack, D. T.; Gancia, E. Estimation of pK_a Using Semiempirical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids. *Quant. Struct.-Act. Relat.* **2002**, *21*, 457–472.
(17) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Gancia, E.; Manallack, D. T. Estimation of pK_a Using Semiempirical Molecular Orbital Methods. Part 2: Application to Amines, Anilines and Various Nitrogen Containing Heterocyclic Compounds. *Quant. Struct.-Act. Relat.* **2002**, *21*, 473–485.
(18) 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.
(19) 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.
(20) Xing, L.; Glen, R. C.; Clark, R. D. Predicting pK_a by Molecular Tree Structured Fingerprints and PLS. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 870–879.
(21) Lee, A. C.; Shedden, K.; Rosania, G. R.; Crippen, G. M. *J. Chem. Inf. Model.* **2008**, *48*, 1379–1388.
(22) Howard, P.; Meylan, W. *Physical/Chemical Property Database (PHYSPROP)*; Syracuse Research Corporation, Environmental Science Center: North Syracuse NY, 1999.
(23) Jover, J.; Bosque, R.; Sales, J. Neural Network Based QSPR Study for Predicting pK_a of Phenols in Different Solvents. *QSAR Comb. Sci.* **2007**, *26*, 385–397.
(24) Parthasarathi, R.; Padmanabhan, J.; Elango, M.; Chitra, K.; Subramanian, V.; Chattaraj, P. K. pK_a Prediction Using Group Philicity. *J. Phys. Chem. A* **2006**, *110*, 6540–6544.
(25) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and Original pK_a Prediction Method Using Grid Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, *47*, 2172–2181.
(26) Habibi-Yangjeh, A.; Danandeh-Jenagharad, M.; Nooshyar, M. Prediction Acidity Constant of Various Benzoic Acids and Phenols in Water Using Linear and Nonlinear QSPR Models. *Bull. Korean Chem. Soc.* **2005**, *26*, 2007–2016.
(27) Jover, J.; Bosque, R.; Sales, J. QSPR Prediction of pK_a for Benzoic Acids in Different Solvents. *QSAR Comb. Sci.* **2008**, *27*, 563–581.
(28) Tao, L.; Han, J.; Tao, F. M. Correlations and Predictions of Carboxylic Acid pK_a Values Using Intermolecular Structure and Properties of Hydrogen-Bonded Complexes. *J. Phys. Chem. A* **2008**, *112*, 775–782.
(29) Silverman, B. D.; Platt, D. E. Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superposition. *J. Med. Chem.* **1996**, *39*, 2129–2140.
(30) Brown, T. N.; Mora-Diez, N. Computational Determination of Aqueous pK_a Values of Protonated Benzimidazoles (part 1). *J. Phys. Chem. B* **2006**, *110*, 9270–9279.
(31) Halle, J. C.; Lelievre, J.; Terrier, F. Solvent effect on preferred protonation sites in nicotinate and isonicotinate anions. *Can. J. Chem.* **1996**, *74*, 613–620.
(32) Schüürmann, G. Quantum Chemical Descriptors in Structure-Activity Relationships - Calculation, Interpretation and Comparison of Methods. In *Predicting Chemical Toxicity and Fate.* Cronin, M. T. D., Livingstone, D. J., Eds.; CRC Press: Boca Raton, FL, 2004, pp 85−149.
(33) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. The Development and Use of Quantum-Mechanical Molecular-Models. 76. AM1 - A New General-Purpose Quantum-Mechanical Molecular-Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
(34) Karickhoff , S. W.; Carreira, L. A.; Hilal, S. H. *SPARC Performs Automated Reasoning in Chemistry*; University of Georgia: Athens, GA; http://ibmlc2.chem.uga.edu/sparc/. Accessed September 4, 2010.
(35) Schüürmann, G.; Ebert, R.-U.; Chen, J.; Wang, B.; Kühne, R. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient − Test Set Activity Mean vs Training Set Activity Mean. *J. Chem. Inf. Model.* **2008**, *48*, 2140–2145.
(36) Schwöbel, J.; Ebert, R.-U.; Kühne, R.; Schüürmann, G. Modeling the H bond Donor Strength of −OH, −NH, and −CH Sites by Local Molecular Parameters. *J. Comput. Chem.* **2009**, *30*, 1454–1464.
(37) Schwöbel, J.; Ebert, R.-U.; Kühne, R.; Schüürmann, G. Prediction of the Intrinsic Hydrogen Bond Acceptor Strength of Organic Compounds by Local Molecular Parameters. *J. Chem. Inf. Model.* **2009**, *49*, 956–962.
(38) Schwöbel, J.; Ebert, R.-U.; Kühne, R.; Schüürmann, G. Prediction of the Intrinsic Hydrogen Bond Acceptor Strength of Chemical Substances from Molecular Structure. *J. Phys. Chem. A* **2009**, *113*, 10104–10112.

CI100306K