JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Prediction of the Dissociation Constant pK_a of Organic Acids from Local Molecular Parameters of Their Electronic Ground State

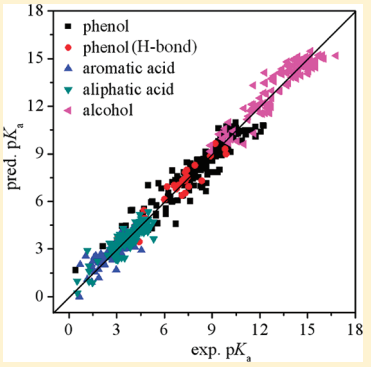
Haiying Yu,[†,‡] Ralph Kühne,[†] Ralf-Uwe Ebert,[†] and Gerrit Schüürmann*[,†,‡]

[†]UFZ Department of Ecological Chemistry, Helmholtz Centre for Environmental Research, Permoserstr. 15, D-04318 Leipzig, Germany

[‡]Institute for Organic Chemistry, Technical University Bergakademie Freiberg, Leipziger Str. 29, D-09596 Freiberg, Germany

Supporting Information

ABSTRACT: A quantum chemical method has been developed to estimate the dissociation constant pK_a of organic acids from their neutral molecular structures by employing electronic structure properties. The data set covers 219 phenols (including 29 phenols with intramolecular H-bonding), 150 aromatic carboxylic acids, 190 aliphatic carboxylic acids, and 138 alcohols, with pK_a varying by 16 units (0.38−16.80). Optimized ground-state geometries employing the semiempirical AM1 Hamiltonian have been used to quantify the site-specific molecular readiness to donate or accept electron charge in terms of both charge-associated energies and energy-associated charges, augmented by an ortho substitution indicator for aromatic compounds. The resultant regression models yield squared correlation coefficients (r^2) from 0.82 to 0.90 and root-mean-square errors (rms) from 0.39 to 0.70 pK_a units, corresponding to an overall (subset-weighted) r^2 of 0.86. Simulated external validation, leave-10%-out cross-validation and target value scrambling demonstrate the statistical robustness and prediction power of the derived model suite. The low intercorrelation with prediction errors from the commercial ACD package provides opportunity for a consensus model approach, offering a pragmatic way for further increasing the confidence in prediction significantly. Interestingly, inclusion of calculated free energies of aqueous solvation does not improve the prediction performance, probably because of the limited precision provided by available continuum-solvation models.



INTRODUCTION

The degree of dissociation of weak organic acids AH is an important determinant of their environmental fate and bioavailability.[1] This includes the sorption to organic matter,[2] affinity for biological membranes,[3] bioaccumulation,[4] and toxicity.[5−7] In dilute aqueous solution where solute activities can be approximated by their concentrations, evaluation of the proton transfer from AH to water, AH + $H_2O \rightleftarrows A^- + H_3O^+$, leads to the well-known definition of the acid dissociation constant $K_a = ([A^-]/[AH]) \bullet [H_3O^+]$ in terms of pK_a = −log K_a[8]

$$pK_a = pH + \log \frac{[AH]}{[A^-]} \quad (1)$$

While pK_a is uniquely defined for monofunctional acids, the presence of three or more acidic sites requires additional information or assumptions for determining the respective micro-pK_a constants.[9]

A recent review informs about methods for predicting pK_a from molecular structure for both small organic compounds and macromolecules.[10] The prominent LFER (linear free energy relationship) approach[8] has been computerized in the commercial ACD software (but without unraveling the individual model equations),[11] and SPARC[12] is based on a similar approach, with functional groups rather than compound classes serving as reference for quantifying fundamental pK_a values. Quantum chemical approaches include the use of continuum-solvation models[13−16] as well as a first-principle approach,[17] and there are also pK_a prediction methods employing chemoinformatics alone[18] or in combination with quantum chemistry.[19] In a recent comparative performance analysis using a test set of 1143 organic acids and bases, ACD was superior to SPARC and two quantum chemical schemes but also showed a substantial variation in prediction quality across compound classes.[20]

In the present study, a new quantum chemical approach for predicting pK_a is introduced. It employs the semiempirical AM1[21] Hamiltonian that enables fast calculations suitable for large chemical inventories. Local molecular parameters based on perturbational molecular orbital theory are used for quantifying site-specific reactivities, leading to simple regression equations for predicting pK_a from the three-dimensional molecular structure of the compounds (in their acid form). The parameters had originally been introduced by Klamt[22,23] and were already used successfully for predicting rate constants of the OH radical reaction with airborne compounds,[22−25] for

Received: May 25, 2011
Published: July 25, 2011

Table 1. Structural Characteristics and Subset-Specific pK_a Value Ranges of the Training Set of 697 Organic Oxygen Acids[a]

		pK_a		
atom composition and compound class	n	min	median	max
CHO(X)	471	0.51	4.62	16.80
CHO(X) + N	672	0.38	4.86	16.80
CHON(X) + PS	697	0.38	4.84	16.80
phenols w/o H-bond	190	0.38	8.88	12.23
phenols with H-bond	29	3.03	7.10	9.87
aromatic carboxylic acids	150	0.65	3.48	4.55
aliphatic carboxylic acids	190	0.51	3.98	5.32
alcohols	138	8.88	14.00	16.80
All				
phenols, aromatic + aliphatic carboxylic acids, alcohols	697	0.38	4.84	16.80

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

predicting the H-bond donor and acceptor strength,[26−29] and for deriving local electrophilicity parameters that in turn enable the prediction of reactive toxicity.[30,31]

The derived model suite is based on calculated site-specific reactivities for dissociation, outperforms previous AM1-based methods for predicting pK_a, and has been validated thoroughly to demonstrate its statistical robustness and prediction power.

MATERIALS AND METHODS

Data Set and Chemical Domain. For 697 organic oxygen acids carrying R−OH or R−COOH as the acidic group, experimental pK_a values were collected from the literature,[20] except for 138 aliphatic alcohols for which the pK_a data were provided by ChemSilico LLC. The measurement temperature ranged from 10 to 30 °C, and was in fact between 20 and 25 °C in the majority of the cases. Because pK_a is not sensitive to temperature, an explicit treatment of the temperature dependence was not undertaken.

Only compounds with one acidic group were included, except when two (or more) functional groups of the same type were present and could be allocated clearly to micro-pK_a values. The reason is that in case two (or more) different acidic groups are present, the resultant micro-pK_a values (the ones associated with each individual acidic group) that build the measurable macro-pK_a cannot be uniquely defined without knowing the relevant pK_b (basicity) values.[8] For compounds with two identical acidic groups, each of which have identical probabilities to become dissociated, however, the acidity is about twice as large as for the respective monofunctional compound, resulting in a pK_a decrease by log 2 (correspondingly, log n applies for a compound with n identical acidic groups in equivalent positions).

The data set compounds cover the atom types C, H, F, Cl, Br, I, N, O, and S, with dissociation being confined to both hydroxyl (−OH) and carboxyl (−COOH). The whole set of 697 compounds was divided into the following five subsets: 190 phenols (without internal hydrogen bonds), 29 ortho-substituted phenols forming intramolecular hydrogen bonds, 150 aromatic carboxylic acids, 190 aliphatic carboxylic acids, and 138 aliphatic alcohols according to dissociation function groups. In Table 1, the respective subsets are listed together with the associated pK_a value ranges.

All experimental pK_a values collected for the present study are listed in Table S1 of the Supporting Information.

Local Molecular Descriptors. Local molecular descriptors are site-specific parameters derived from quantum chemical calculations employing the LCAO-MO (linear combination of atomic orbitals−molecular orbital) approach. In contrast to conventional MO-based parameters such as the frontier orbital energies HOMO and LUMO (highest occupied and lowest unoccupied molecular orbital), local molecular descriptors are designed to extract from the delocalized MO wave functions and energies, energy and charge information that reflects the local characteristics of a given atomic site in the molecular environment.[22−31] Accordingly, such parameters may contain pertinent information about local bond strengths of H atoms attached to heteroatoms, which in turn depend on the ability of the bonding partners to donate or accept electron density and thus provide a direct link to the pK_a. Interestingly, this parameter type has already proven useful for predicting the strength of H bonding,[26−29] a process different from but related to the complete transfer of a proton from acid AH to water.

The energy-weighted donor energy $EE_{occ}(E_{ref}, r)$ describes the electron donor ability of a molecule at atomic site r, and is constructed through a sum of occupied MO energies E_i, weighted by exponential terms $w_i(E_{ref}, r)$ involving a reference energy E_{ref}:

$$EE_{occ}(E_{ref}, r) = \frac{\sum_{i=1}^{HOMO} E_i \times w_i(E_{ref}, r)}{\sum_{i=1}^{HOMO} w_i(E_{ref}, r)} \quad (2)$$

with

$$w_i(E_{ref}, r) = p_i(r)\exp\left(-\frac{E_i}{E_{ref}}\right) \qquad p_i(r) = 2\sum_{\mu(r)} c_{\mu i}^2$$

In eq 2, the LCAO-MO coefficient $c_{\mu i}$ quantifies the contribution of the μ-th AO at center r to the i-th MO. Unlike the HOMO energy, EE_{occ} also incorporates energetically lower MOs, weighting the E_i contribution according to the local electron density (p_i) and reference energy E_{ref} that in turn can be calibrated according to the property of interest. EE_{occ} ranges between the HOMO energy as delocalized limit (for $E_{ref} \rightarrow 0$) and the sum of the orbital energies weighted only by p_i (for $E_{ref} \rightarrow -\infty$), i.e., $(\Sigma E_i p_i)/(\Sigma p_i)$.

The energy-weighted acceptor energy $EE_{vac}(E_{ref}, r)$ is defined accordingly through unoccupied MOs. It characterizes the capability of the molecule to accept additional electron charge at atomic site r, and thus represents a localized generalization of the LUMO energy.

Another pair of local reactivity parameters are the charge-limited donor energy $EQ_{occ}(q, r)$ and the charge-limited acceptor energy $EQ_{vac}(q, r)$, the latter of which is defined as

$$EQ_{vac}(q, r) = \frac{1}{q}\sum_{k=LUMO}^{n} E_k w_k(q, r)$$

with

$$w_k(q, r) = \begin{cases} p_k & \text{if} \quad b_k + p_k \le q \\ q - b_k & \text{if} \quad b_k < q < b_k + p_k \\ 0 & \text{if} \quad q \le b_k \end{cases} \quad \text{and} \quad b_k(r) = \sum_{l=LUMO}^{k-1} p_l(r) \quad (3)$$

dx.doi.org/10.1021/ci200233s |J. Chem. Inf. Model. 2011, 51, 2336–2344

$EQ_{vac}(q, r)$ characterizes the energy gain upon accepting charge q (in units of electron charge) at atomic site r, and can be understood as a further local generalization of the LUMO energy. EQ_{vac} is calculated as weighted mean of the unoccupied molecular orbital energies. Starting with the LUMO, excess electron population p_k at center r is taken into account until the predefined charge limit q is achieved. For the sake of simplicity, the summed charge above the k-th MO at center r was abbreviated with $b_k(r)$ in eq 3. In case of $q \rightarrow 0$, eq 3 yields the LUMO energy as delocalized limit. As a general trend, EQ_{vac} becomes increasingly local with increasing amount of the charge penetration depth q. $EQ_{occ}(q, r)$ is defined analogously through occupied MOs and quantifies the energy associated with donating charge q from atomic site r.

EQ_{occ} and EQ_{vac} quantify the energy referring to donating or accepting a certain amount of charge q. A complementary approach is to evaluate, for a given energy loss or gain, the associated amount of charge released from or taken up at site r. Following this idea, an energy-limited donor charge $QE_{occ}(\varepsilon, r)$ can be defined as amount of charge being removed from center r when offering the energy ε

$$QE_{occ}(\varepsilon, r) = \sum_{i=1}^{HOMO} p_i(r) w_i(\varepsilon)$$

with

$$w_i(\varepsilon) = \begin{cases} 1 & \text{if} \quad \varepsilon \le E_i - 0.5 \\ E_i - \varepsilon + 0.5 & \text{if} \quad E_i - 0.5 < \varepsilon < E_i + 0.5 \\ 0 & \text{if} \quad E_i + 0.5 \le \varepsilon \end{cases} \tag{4}$$

As with E_{ref} (eq 2), the energy penetration depth ε can be calibrated for the target property of interest. As expected, atomic sites with high electron donor ability are characterized by large values for QE_{occ}. The correspondingly defined energy-limited acceptor charge QE_{vac} involves unoccupied MOs and quantifies the amount of accepted electron charge that is associated with a predefined energy gain ε.

Quantum Chemical Calculations. The geometries of all molecules were optimized in their neutral electronic ground state, employing the semiempirical AM1[21] Hamiltonian as implemented in MOPAC 2002.[32] All local molecular parameters described above were evaluated at the acidic sites of the compounds (see Results and Discussion) from their ground-state MO energies and LCAO-MO coefficients. At this stage, aqueous solvation was not addressed, thus focusing on intrinsic properties and in order to avoid confounding factors introduced through the further approximations associated with semiempirical solvation models. However, for evaluating the potential impact of including aqueous solvation on predicting pK_a, respective solvation free energies were calculated with the continuum-solvation model SM5.42[33] as implemented in AMSOL 7.1.[34]

Model Calibration. Multilinear regression (MLR) for calibration the pK_a prediction models was performed employing the software SPSS 17.0. The calibration performance was quantified in terms of the squared correlation coefficient, r^2, the root-mean-square error, rms, mean error (me), systematic error (bias), maximum positive error mpe (maximum overestimation of pK_a), and maximum negative error (maximum underestimation). In the context of model validation (see below), the prediction performance was quantified in terms of the predictive squared correlation coefficient, q^2.[35]

The calibration of the reference values of the six local molecular parameters (see above) was performed in a stepwise manner. To this end, every parameter was calculated for a series of test reference values followed by a multilinear regression calibration of the linear model parameters, and the reference value or values (in case of more than one local molecular parameter) yielding the best overall statistics were selected as final reference value(s).

Model Validation. Three validation studies were undertaken to evaluate the statistical robustness and prediction power of the derived regression models. First, the data set was split temporarily into a training set and a prediction set, covering 70% and 30% of the compounds while preserving the overall compound class composition. To this end, the compounds were grouped according to the five compound classes mentioned above (see also Table 1), and then from each group 30% were randomly selected for the temporary prediction set, leaving the remaining 70% as training set compounds (stratified random selection). Subsequently, the derived pK_a prediction models were recalibrated for the 70% subset and applied to the 30% subset that had been temporarily left out for calibration. The resultant q^2 of the latter application—evaluated as for an external data set[35]—informs about the prediction capability, and comparison of r^2 and rms for the 70% and 30% subsets provides information about the statistical robustness.

Second, cross-validation was performed in a leave-10%-out manner, again preserving the overall compound class composition through stratified random selection of the 10 subsets of which each covered 10% of the compounds. In this case, the pK_a values of a given 10% subset are predicted from a model calibrated on the other nine subsets covering 90% of the compounds. The resultant cross-validated q^2 (calculated as described earlier)[35] provides further information about the prediction capability and statistical robustness.

Third, a permutation test—also called target value scrambling test—was performed to explore the potential presence of overfitting and chance correlation. Permutation means that for a given pair of two compounds, their pK_a values are temporarily exchanged. Respective permutations were undertaken to generate wrong compound-pK_a allocations for different fractions of the data set up to 100% reallocation. The degree of scrambling is quantified through evaluating the q^2 between the original (unscrambled) and permuted pK_a values, keeping in mind that a 20% permutation does not necessarily result in 20% numerically wrong pK_a values because not all compounds have necessarily different pK_a values (and some may have very similar pK_a values). For regression models based on mechanistically sound relationships, increasing the degree of such target value scrambling should decrease the calibration r^2, with 100% permutation yielding r^2 values usually around or below 0.1 (because usually the target values could only be properly predicted from the descriptor values belonging to the correctly assigned compounds).

Influential Compounds and Outliers. In the descriptor space, a high-leverage compound is one with extreme (large or small) descriptor values. Such compounds have a significantly larger influence on the calibration of the regression coefficients than compounds with typical descriptor values. The leverage h_i of a given compound i is calculated as respective diagonal element of the so-called Hat matrix (that relates the target values to their prediction counterparts),[36] and the threshold or warning leverage h^* for identifying a statistically high leverage is usually taken

Table 2. Regression Results for Predicting pK_a of Five Subsets, Employing AM1-Based Local Molecular Parameters and An Indicator Variable *I* for Aromatic Compounds[a]

data sets	linear regression equations: pK_a =	*n*
phenols w/o H-bond	$3.02 \times EQ_{vac}(0.28e, H) - 4.48 \times QE_{vac}(6.0\ eV, H) - 0.46 \times I + 5.37$	190
phenols with H-bond	$2.55 \times EE_{vac}(1.0\ eV, H) + 3.24 \times EE_{occ}(-2.2\ eV, O) + 39.01$	29
aromatic acids	$1.69 \times EE_{vac}(1.3\ eV, H) - 0.73 \times I - 1.03$	150
aliphatic acids	$1.75 \times EE_{vac}(1.4\ eV, H) + 2.77 \times QE_{occ}(-14.8\ eV, =O) - 8.92$	190
alcohols	$1.25 \times EQ_{vac}(0.45e, A) - 1.81 \times QE_{vac}(2.8\ eV, H) + 10.74$	138

[a] EQ_{vac} = charge-limited acceptor energy (eq 3); QE_{occ} = energy-limited donor charge (eq 4); QE_{vac} = energy limited acceptor charge (see text); EE_{occ} = energy-weighted donor energy (eq 2); EE_{vac} = energy-weighted acceptor energy (see text), all specified further through reference values (see text) and evaluated at the one of the following atoms: acidic H (H), oxygen bonded to acidic H (O), carbonyl oxygen of the acidic carboxyl group (=O), and atom bonded to the acidic OH (A). Indicator variable $I = 1$ for ortho substitution at the aromatic ring carrying the acidic site (−OH or −COOH), $I = 0$ otherwise (meta or para or no substitution).

Table 3. Statistical Performance of the Model Suite for Predicting pK_a Employing AM1-Based Local Molecular Parameters[a]

data set	*n*	r^2	q^2	rms	bias	me	mne	mpe
all	697	0.98	0.98	0.54	−0.01	0.39	−2.09	2.08
all (subset-weighted performance)[b]	697	0.86	0.86	0.52	−0.01	0.39	−1.74	1.75
phenols w/o intramolecular H-bonding	190	0.90	0.90	0.70	−0.02	0.53	−2.09	2.08
phenols with intramolecular H-bonding	29	0.92	0.92	0.52	−0.00	0.41	−1.04	0.75
aromatic carboxylic acids	150	0.82	0.82	0.39	−0.01	0.26	−1.62	1.50
aliphatic carboxylic acids	190	0.82	0.82	0.39	−0.02	0.28	−1.68	1.75
aliphatic alcohols	138	0.88	0.88	0.61	0.00	0.49	−1.63	1.80

[a] The statistical parameters are *n* = number of compounds; r^2 = squared correlation coefficient; q^2 = predictive squared correlation coefficient;[35] rms = root-mean-square error; bias = systematic error, me = mean absolute error; mne = maximum negative error (largest underestimation); and mpe = maximum positive error (largest overestimation). [b] Calculated as average from subset-specific r^2 and q^2 and weighted according to subset size as introduced earlier.[20]

as $3p/n$ where p = number of regression model parameters, and n = number of compounds (observations).

A high-leverage compound may but need not be an outlier of the model. Outliers can be characterized statistically by a standardized prediction residual larger than 3, and thus inform about limitations of the application domain of the regression model. Thus, the Williams plot of standardized prediction residuals versus leverages was used to identify both high-leverage compounds and outliers.

RESULTS AND DISCUSSION

Data Set Characteristics. Table 1 shows the decomposition of the 697 organic acids into atom types and major compound classes. A total of 471 compounds (68%) are built from C, H, O, and possibly halogen (X), with pK_a ranging from 0.51 to 16.80 and a median of 4.62. The N atom as further atomic constituent applies for 201 compounds (29%), and there are 25 acids (4%) containing P or S (or both). All acids are oxygen acids involving dissociation from −OH attached to aliphatic or aromatic carbon or as part of a carboxylic group (−COOH) of respective aliphatic or aromatic acids.

Among the 697 organic acids, there are 219 phenols, including 29 with the opportunity for intramolecular H bonding, 150 aromatic and 190 aliphatic carboxylic acids, and 138 aliphatic alcohols. Overall, pK_a varies from 0.38 (2,4,6-trinitrophenol) to 16.8 (2-butanol), and has a median of 4.84 (2-methyl-propanoic acid, pentanoic acid, 4-methyl-pentanoic acid).

Model Calibration. Table 2 lists the regression equations for the five compound classes under analysis, and Table 3 provides the respective performance statistics, with r^2 ranging from 0.82 to 0.92, and rms ranging from 0.39 (aliphatic and aromatic carboxylic acids) to 0.70 (phenols without intramolecular H bonding). At the same time, the class-specific models are both simple (up to two local reactivity parameters and up to one indicator variable) and mechanistically sound as discussed below.

For assessing the overall prediction performance, both the standard r^2 and the subset-weighted r^2 (averaging the subset-specific r^2 values weighted according to subset size)[20] are listed in Table 3. As pointed out earlier,[20] we consider the latter approach more suitable for characterizing the statistical performance of composite regression models consisting of subset-specific regression equations. For the present set of 697 organics, the subset-weighted r^2 indicates that 86% of the pK_a variance can be explained through class-specific regressions on local reactivity parameters derived from the electronic ground state of the molecules.

Interestingly, the prediction precision for alcohols (rms 0.61 aliphatic, 0.52 and 0.70 aromatic) is significantly inferior to the one for carboxylic acids (rms 0.39 both aliphatic and aromatic). Moreover, the subset of 190 phenols without intramolecular H bonding contains the largest outliers regarding both pK_a underestimation (−2.09: 3, 5-dinitrophenol) and overestimation (2.08: hexachlorophene). Nevertheless, the aliphatic and aromatic alcohols yield larger r^2 values than the respective carboxylic acids (0.88−0.92 vs 0.82), which is driven by their larger range of pK_a values included (16 units vs 5 units).[35]

Note further the smaller maximum pK_a of phenols with intramolecular H bonding as compared to phenols without intramolecular H bonding (9.9 vs 12.2), illustrating the decreased acidity of H-bonded H atoms.

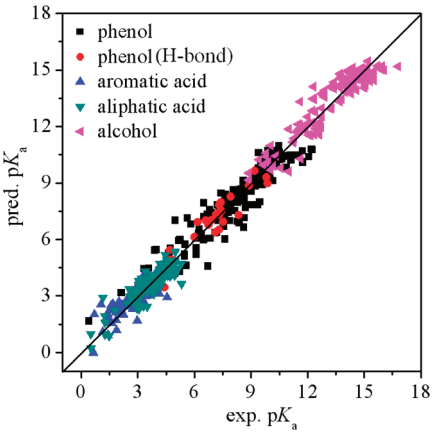


Figure 1. Predicted vs experimental pK_a for 697 organic oxygen acids with acidic sites −OH and −COOH, respectively, employing the model suite of Table 2 with AM1-based local molecular parameters.

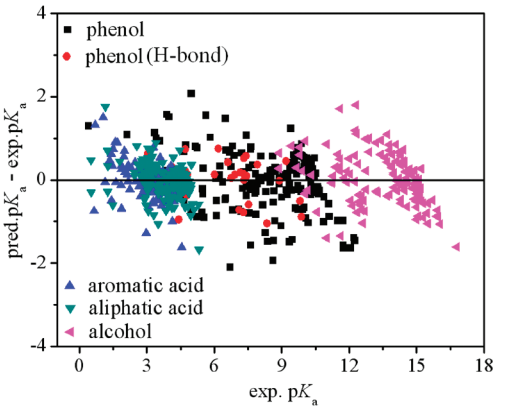


Figure 2. Prediction error vs experimental pK_a for 697 organic oxygen acids with acidic sites −OH and −COOH, respectively, employing the model suite of Table 2 with AM1-based local molecular parameters.

The calculated versus experimental pK_a data distribution for all 697 organic acids is shown in Figure 1, and the distribution of the pK_a prediction error versus experimental pK_a is shown in Figure 2. The latter reveals that with the presently introduced regression models based on local molecular reactivity parameters, there is no global dependence of the pK_a prediction error on pK_a.

Model Interpretation. The pK_a prediction model for phenols without intramolecular H bonding contains the local parameters EQ_{vac} and QE_{vac} evaluated at the acidic H atom, and an indicator variable I indicating ortho substitution (I = 1; 0 otherwise; see Table 2). The charge-limited acceptor energy EQ_{vac}(0.28e, H) quantifies the energy gain through accepting 0.28 electron charge at the H atom, and increases with an increasing corresponding energy gain. Thus, H atoms with a large EQ_{vac} prefer more strongly to retain electron charge and accordingly provide more resistance in donating charge to their bonding partner in order to become dissociated. It follows that increasing EQ_{vac} correlates with decreasing acidity and thus increasing pK_a, which is reflected by the positive sign of the regression coefficient (+3.02; see Table 2).

The second reactivity parameter of the regression model for phenols without intramolecular H bonds is the energy-limited acceptor charge QE_{vac}(6.0 eV, H), quantifying the amount of electron charge associated an energy gain of 6.0 eV at the acidic H atom. Larger QE_{vac} values reflect larger amounts of electron charge per unit energy transferred to H and thus a larger polarizability of H in its bonding situation. Because increasing polarizability indicates a decreasing resistance to changing local electron density, QE_{vac} increases with increasing readiness of the acidic H atom to become ionized upon dissociation. Accordingly, QE_{vac} increases with increasing acidity and thus with decreasing pK_a as indicated by the negative sign of its regression coefficient (−4.48; Table 2).

The pK_a prediction model derived for the 29 phenols with intramolecular H bonding employs EE_{vac} evaluated at the acidic H atom and EE_{occ} evaluated at the oxygen atom bonded to this H atom. The energy-weighted acceptor energy EE_{vac}(1.0 eV, H) increases with increasing electron acceptor strength of H, indicating a larger energy demand for ionizing H to become H^+. Thus, EE_{vac} increases with decreasing acidity and accordingly increasing pK_a, which is reflected by its positive regression coefficient (+2.55; Table 2). The energy-weighted donor energy EE_{occ}(−2.2 eV, O) evaluated at the oxygen bonded to H increases with increasing oxygen donor strength, which in turn reflects an increasing O−H bond strength and thus a lower tendency for bond fission. Accordingly, EE_{occ} correlates with decreasing acidity and thus with increasing pK_a, and has a positive regression coefficient (+3.24; Table 2).

EE_{vac} evaluated at H is also used as local reactivity parameter of the pK_a prediction models calibrated for the 150 aromatic and 190 aliphatic carboxylic acids, supported by a corresponding mechanistic reasoning. For the latter compound class, the energy-limited donor charge evaluated at the carbonyl oxygen of the carboxylic group, QE_{occ}(−14.8 eV, =O), is used as second molecular parameter. QE_{occ} increases with increasing amount of loose electron charge ready for donation, which in turn stabilizes the carboxylic O−H bond. It follows that QE_{occ} increases with decreasing acidity (decreasing O−H bond fission tendency) and thus with increasing pK_a and accordingly has a positive regression coefficient (+2.77; Table 2).

The indicator variable I (1 for ortho substitution, 0 otherwise) is used in the pK_a predictions models for phenols (without intramolecular H bonding) and aromatic carboxylic acids. In both cases, the regression coefficient is negative, indicating an increased acidity for ortho-substituted compounds. A possible explanation is that ortho substitution destabilizes the molecular ground state as compared to meta and para substitution, and that the associated steric repulsion decreases upon dissociation, thus supporting the cleavage of H^+.

Furthermore, the t-test indicates that every descriptor selected in Table 2 is statistically significant at the α = 0.001 level (p < 0.001) for the respective regression models except for QE_{vac}(2.8 eV, H) in the alcohol model, which is significant at the α = 0.1 level (p = 0.092).

Validation. First and as outlined above, the total compound set (n = 697) was split in a stratified random manner into a temporary training set covering 70% of the compounds (n = 489) and a temporary prediction set built from 30% of the compounds (n = 208), with both subsets reflecting the overall composition into the five compound classes in terms of their relative proportions. Subsequently, the models were recalibrated using the 70% subset, and their recalibrated versions were used for predicting the pK_a of the 30% subset compounds (that for this exercise were kept external from model calibration, hence simulated external validation).

Table 4. Statistical Validation Results for the Model Suite Employing AM1-Based Local Molecular Parameters, Using Simulated External Validation and Leave-10%-out Cross-Validation[a]

	training set			test set				cross-validation	
data set	n	r^2	rms	n	r^2	q^2	rms	q^2	rms
all (subset-weighted performance)	489	0.86	0.53	208	0.86	0.85	0.54	0.85	0.54
phenols w/o intramolecular H-bonding	133	0.90	0.69	57	0.88	0.88	0.78	0.88	0.73
phenols with intramolecular H-bonding	20	0.93	0.52	9	0.97	0.83	0.65	0.88	0.60
aromatic carboxylic acids	106	0.81	0.41	44	0.84	0.84	0.34	0.81	0.40
aliphatic carboxylic acids	134	0.82	0.40	56	0.83	0.82	0.37	0.82	0.39
aliphatic alcohols	96	0.88	0.61	42	0.87	0.86	0.65	0.87	0.62

[a] The temporary training and test sets used for the simulated external validation were generated from the total data set as subsets containing 70% and 30% of the compounds, respectively. The respective subsets were generated through stratified random selection, preserving the relative composition with respect to all five compound classes (see also text). The 10 subsets used for the leave-10%-out cross-validation were also generated in a correspondingly stratified manner.

As shown in Table 4, the training set (70% subset) r^2 values are similar to both r^2 and q^2 of the prediction set (30% subset). In particular, the subset-weighted r^2 is the same for both subsets (0.86) as well as for the full data set, with the rms being almost the same for all three sets (0.53−0.54). Concerning the individual compound classes, both r^2 and rms of the 70% subset come close to the full data set calibration statistics (Table 3), which holds correspondingly for the individual regression equations (details not shown). It follows that the presently derived model suite can be considered to be statistically robust and predictive.

Second, a leave-10%-out cross-validation was performed. For each of the corresponding 10 runs, the 10% left out from calibration but used for prediction were selected randomly under the constraint of compound class stratification (implying that for each run, 10% of each compound class was left out). The resultant cross-validated q^2 and rms values are again similar to the total set calibration statistics, confirming the statistical robustness and prediction power.

Third, a permutation test (target value scrambling; see Material and Methods) was performed to evaluate the statistical relevance of properly representing the compounds by their true descriptor values. For regression models based on a mechanistically sound relationship, increasing the degree of scrambling should decrease the calibration quality, and the completely scrambled data set (where each compound is associated with the pK_a of another compound) should yield r^2 values below 0.1, provided the pK_a values differ sufficiently between the different compounds. By contrast, if fully scrambled data sets yield r^2 values substantially larger than 0, this indicates that even noise could predict the target property to the r^2 level obtained, and thus suggests the presence of overfitting or chance correlation or both.

We generated 10 different degrees of scrambling, ranging from about 10% reallocated pK_a values to 100% permutation in steps of about 10%. To this end, the degree of scrambling actually achieved with random permutations was characterized through evaluating the q^2 between the original and scrambled pK_a data. In Figure 3, the respective dependence of r^2 on the degree of scrambling is shown for the 190 phenols without intramolecular H bonding; respective plots for the other compound classes are given in Figure S1 of the Supporting Information. At the left-hand side of the x axis, $q^2 = -1$ corresponds to 100% scrambling, while $q^2 = 1$ at the right-hand side corresponds to using the original data set (0% scrambling). The observed trend demonstrates that for the phenol model, r^2 decreases systematically with

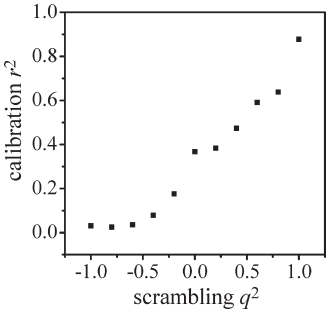


Figure 3. Permutation test results with the subset of 190 phenols without intramolecular H-bonding, employing the model with AM1-based local molecular parameters (Table 2). The calibration performance r^2 for predicting pK_a is plotted against the degree of target value scrambling, the latter of which is quantified as q^2 between the scrambled and original (unscrambled) pK_a values of the data set (see text).

increasing degree of scrambling (when going from right to left in Figure 3) down to values below 0.05, confirming the absence of overfitting and chance correlation. Similar results are obtained for the other compound classes, confirming that the presently introduced model suite has a sound mechanistic basis.

Overall, the three validation studies undertaken suggest a global prediction capability in terms of rms of around 0.4 for carboxylic acids, 0.6 for aliphatic alcohols, and 0.8 for phenols.

Influential Compounds and Outliers. Application of the Kolmogorov−Smirnov test confirmed that the prediction residuals of all five subsets (phenols without and with intramolecular H bonding, aromatic and aliphatic carboxylic acids, and aliphatic alcohols) follow normal distributions at the 95% confidence interval. Accordingly, the leverage approach and regression residuals can be employed to identify influential compounds and outliers.

In Figure 4, a respective Williams Plot is shown for the subset of 190 aliphatic carboxylic acids (corresponding plots for the other four subsets are shown in Figure S2 of the Supporting Information). As shown in the plot, there is one entry above three standardized residual units (top left in Figure 4) and one below (bottom left), representing the two outlying compounds trinexapac and trichloroacrylic acid (see also Scheme 1). Moreover, three entries at the right side of the plot have leverage values larger than the warning leverage

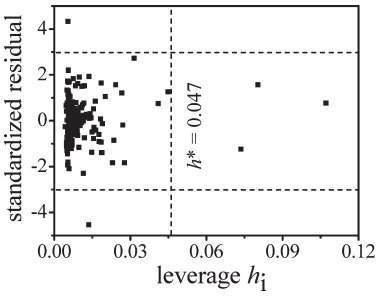
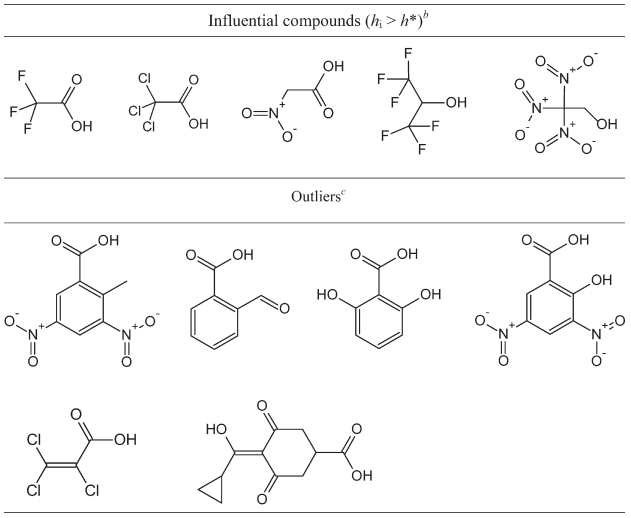


Figure 4. Williams plot for the subset of 190 aliphatic acids (see text), employing the model with AM1-based local molecular parameters (Table 2). The two data points exceeding 3 standardized residuals (y axis) represent outliers, while the three high-leverage data points at the right-hand side (that exceed the warning leverage h^*) represent compounds with higher influence on the regression model coefficients. The respective molecular structures are shown in Scheme 1.

Scheme 1. Influential Compounds and Outliers of the Model Suite Predicting pK_a from AM1-Based Local Molecular Parameters[a]



[a] Influential compounds are identified through their leverage h_i being larger than the warning leverage h^* (see text and Figure 4). [b] Compounds are from left to right: trifluoroacetic acid, tricloroacetic acid, nitroacetic acid, 1,1,1,3,3,3-hexafluoro-propan-2-ol, and 2,2,2-trinitroethanol, respectively. [c] Compounds are from top left to bottom right: 2-methyl-3,5-dinitro-benzoic acid, 2-formyl-benzoic acid, 2,6-dihydroxy-benzoic acid, 2-hydroxy-3,5-dinitro-benzoic acid, trichloroacrylic acid, and trinexapac, respectively.

$h^* = 3p/n$ (p = # model parameters, n = # subset compounds) and represent the compounds trifluoroacetic acid, trichloroacetic acid, and nitroacetic acid.

In Scheme 1, all five high-leverage compounds and all six outliers are shown with their molecular structures. Interestingly, the former group is characterized by the presence of strong electron-withdrawing groups ($-CF_3$, $-CCl_3$, $-NO_2$) close to the acidic site, making extreme values of the respective local molecular descriptors at the acidic site reasonable. At the same time, these compounds are no statistical outliers, indicating that the derived regression models are indeed applicable to such descriptor value combinations.

The lower part of Scheme 1 shows the six statistical outliers as identified through correspondingly large residuals. These include four aromatic acids with both acceptor ($-CHO$, $-NO_2$) and π-donor ($-OH$) substituents, and two aliphatic carboxylic acids, of which one is α,β-unsaturated with inductive electron-withdrawing substituents ($-Cl$), and the other one carries an additional OH group that is acidified through resonance interaction with an unsaturated moiety. At present, we have no mechanistic explanation for their outlying behavior but take these as out-of-domain reference compounds to guide a proper use of the derived model suite. Here, a possible procedure is to use the ACF (atom-centered fragment) approach[37] for identifying potential outliers based on their ACF similarity to these six compounds.

Solvation Energy Contribution to pK_a. Coming back to the dissociation process in terms of the proton transfer from acid AH to water, $AH + H_2O \rightleftarrows A^- + H_3O^+$, the respective reaction free energy ΔG – that is proportional to pK_a – can be written as

$$\Delta G = \Delta G_f(A^-) + \Delta G_f(H_3O^+) - [\Delta G_f(AH) + \Delta G_f(H_2O)] \quad (5)$$

where ΔG_f denotes the free energy of formation of the respective molecular species, and AH and A^- the neutral (protonated) and dissociated form of the acid under consideration. Each of the four terms on the right-hand side of eq 5 can be decomposed into a gas-phase part and a contribution from aqueous solvation, such as $\Delta G_f(AH) = \Delta G_{f,gas}(AH) + \Delta G_s(AH)$ for the acid AH. Approximating the gas-phase free energies of formation by AM1 heat of formations, $\Delta G_{f,gas} \approx \Delta H_{f,gas} \equiv \Delta H_f$ (AM1 refers to the gas phase by definition), then leads to

$$\Delta G \approx \Delta\Delta H_f + \Delta\Delta G_s \quad (6)$$

with

$$\Delta\Delta H_f = \Delta H_f(A^-) + \Delta H_f(H_3O^+) - [\Delta H_f(AH) + \Delta H_f(H_2O)] \quad (7)$$

$$\Delta\Delta G_s = \Delta G_s(A^-) + \Delta G_s(H_3O^+) - [\Delta G_s(AH) + \Delta G_s(H_2O)] \quad (8)$$

Equation 8 represents the solvation contribution $\Delta\Delta G_s$ to pK_a, resulting from the free energies of solvation, ΔG_s, of the relevant molecular species. Equations 6–8 actually offer a direct quantum chemical approach for predicting pK_a, with the prediction quality depending heavily on the computational level available for quantifying $\Delta\Delta H_f$ and $\Delta\Delta G_s$.

Employing AM1[21] for quantifying the gas-phase proton transfer energy ($\Delta\Delta H_f$, eq 7) and the continuum-solvation model SM5.42[33] for quantifying the solvation contribution ($\Delta\Delta G_s$, eq 8), however, yields relatively poor prediction statistics as summarized in Table 5.

With solution-phase ΔG approximated through eq 6, the overall subset-weighted r^2 is 0.71, which is significantly inferior to the one of our currently introduced model suite (r^2 = 0.86, Table 3) and even inferior to taking the gas-phase reaction energy alone ($\Delta\Delta H_f$ according to eq 7, r^2 = 0.74, Table 5). Moreover, offering an additional degree of freedom for calibration through using $\Delta\Delta H_f$ and $\Delta\Delta G_s$ as linear combination, yields only a marginal improvement (r^2 = 0.75) over taking $\Delta\Delta H_f$ alone.

These findings suggest that the semiempirical continuum-solvation approach employed is not sufficiently precise to quantify the solvation contribution to pK_a. Accordingly, attempts

Table 5. Performance Statistics of Predicting pK_a through Calculated Gas-Phase and Solution-Phase Proton Transfer Energies and Associated Solvation Free Energies Employing the Semiempirical Quantum Chemical AM1 and SM5.42 Schemes, Respectively[a]

		r^2 between pK_a and			
data set	n	$\Delta\Delta H_f$	$\Delta\Delta G_s$	ΔG	linear combination of $\Delta\Delta H_f$ and $\Delta\Delta G_s$
all (subset-weighted performance)	697	0.74	0.45	0.71	0.75
phenols w/o intramolecular H-bonding	190	0.86	0.58	0.81	0.87
phenols with intramolecular H-bonding	29	0.79	0.62	0.80	0.80
aromatic acids	150	0.55	0.30	0.62	0.55
aliphatic acids	190	0.73	0.34	0.68	0.75
alcohols	138	0.78	0.55	0.70	0.78

[a] $\Delta\Delta H_f$ = gas-phase proton transfer energy from acid AH to water calculated through AM1 heats of formation (eq 7); $\Delta\Delta G_s$ = solvation contribution to pK_a calculated from SM5.42 free energies of solvation of the relevant molecular species (eq 8); ΔG = solution-phase proton transfer energy employing AM1 and SM5.42 (eq 6); and linear combination of $\Delta\Delta H_f$ and $\Delta\Delta G_s$ = multilinear regression on the individual gas-phase and solution-phase contributions to pK_a.

Table 6. Statistical Performance of Quantum Chemical and LFER-Based Models for Predicting pK_a of Organic Oxygen Acids[a]

method	n	r^2	q^2	rms	bias	me	mne	mpe
new model	697	0.86	0.86	0.52	−0.01	0.39	−1.74	1.75
QC	559[b]	0.79	0.78	0.58	−0.02	0.42	−2.45	1.62
r-QC	559[b]	0.82	0.82	0.53	0.00	0.39	−1.70	2.02
SPARC	697	0.83	0.57	0.82	0.19	0.59	−1.68	3.38
ACD	697	0.91	0.90	0.44	0.06	0.25	−2.44	1.83

[a] New model = currently introduced model (see Table 2); QC and r-QC = AM1-based models employing donor delocalizability as major descriptor in original and recalibrated version;[20] r^2 and q^2 calculated as averages from respective subset-specific values, weighted according to subset size.[20] [b] QC and r-QC are not available for alcohols, thus reducing the number of compounds by 138 compounds.

to further improve the prediction quality of our presently introduced model suite (Tables 2 and 3) through additional inclusion of $\Delta\Delta G_s$ failed (details not shown). Comparison of Table 5 with Table 3 demonstrates further the scope of local molecular parameters to extract, in a mechanistically sound manner, site-specific molecular reactivity that drives the dissociation of acids, resulting in overall statistics significantly superior to building on calculated global molecular energies.

Consensus Model Strategy. In Table 6, the prediction performance of our presently introduced model is compared with the ones of existing AM1-based models[20] as well as of SPARC[12] and ACD[11] for the present set of 697 organic acids. The new model outperforms all other methods except ACD (r^2 0.86 vs 0.91, rms 0.52 vs 0.44) that employs a large (and in fact publically unknown) number of locally parametrized LFER equations. At the same time, the ACD prediction quality varies also significantly across different compound classes as already noted above,[20] and the overall intercorrelation between the prediction errors of ACD and of our present model is only $r^2 < 0.14$.

These findings suggest using both the present model and ACD in a consensus model approach. In case both methods—that differ methodologically and with regard to their prediction errors—provide similar pK_a estimates, the respective confidence in prediction is significantly increased. Correspondingly, substantial differences in pK_a prediction would call for a more detailed investigation. A respective consensus model approach thus provides additional information about assessing the prediction quality for individual compounds, the latter of which may—naturally—differ substantially from the one expected from global regression statistics.

CONCLUSIONS

The developed pK_a prediction model suite based on quantum chemical local reactivity parameters outperforms existing AM1-based approaches. For 697 oxygen acids covering phenols, aliphatic and aromatic carboxylic acids, and aliphatic alcohols, the subset-weighted overall squared correlation coefficient (r^2) is 0.86, and the root-mean-square error (rms) is 0.54 pK_a units. Comprehensive validation through simulated external prediction, cross-validation, and target value scrambling demonstrate the statistical robustness and prediction power, suggesting rms values in the prediction mode of around 0.4 for carboxylic acids, 0.6 for aliphatic alcohols, and 0.8 for phenols belonging to the chemical domain of the present data set. The difference in methodology and concerning prediction errors to the ACD model suite suggests further to use both methods in combination following a consensus model approach, thus increasing the confidence in prediction significantly for those compounds where both methods yield similar pK_a estimates. Comparison with solution-phase dissociation energy calculations employing AM1 and SM5.42 demonstrate the superiority of using site-specific reactivity parameters when deriving pK_a prediction models based on multilinear regression. Potential room for improvement is provided by quantum chemical ab initio calculations, keeping in mind their substantially larger computation time. A further line of development concerns the calibration of local reactivity parameters for the basicity of organic bases, thus allowing one to predict also multifunctional acids and bases that are subject to various macro- and micro-pK_a values.

ASSOCIATED CONTENT

Supporting Information. A table listing all 697 organic oxygen acids with experimental and predicted pK_a values (Table S1),

and two figures showing further permutation test results (Figure S1) and Williams plots (Figure S2). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author
*E-mail: gerrit.schuurmann@ufz.de.

ACKNOWLEDGMENT

We thank Dr. Joe Votano of ChemSilico LLC for generously providing pK_a values of alcohols for this study. This research was financially supported by the China Scholarship Council (CSC) and by the EU integrated project OSIRIS (Contact 037017), which is gratefully acknowledged.

REFERENCES

(1) Franco, A.; Trapp, S. A multimedia activity model for ionizable compounds: Validation study with 2,4-dichlorophenoxyacetic acid, aniline, and trimethoprim. *Environ. Toxicol. Chem.* **2010**, *29*, 789–799.

(2) Franco, A; Trapp, S. Estimation of the soil−water partition coefficient normalized to organic carbon for ionizable organic chemicals. *Environ. Toxicol. Chem.* **2008**, *27*, 1995–2004.

(3) Escher, B. I.; Schwarzenbach, R. P. Evaluation of liposome−water partitioning of organic acids and bases. 1. Development of a sorption model. *Environ. Sci. Technol.* **2000**, *34*, 3954–3961.

(4) Fu, W.; Franco, A.; Trapp, S. Methods for estimating the bioconcentration factor of ionizable organic chemicals. *Environ. Toxicol. Chem.* **2009**, *28*, 1372–1379.

(5) Divkovic, M.; Pease, C. K.; Gerberick, G. F.; Basketter, D. A. Hapten-protein binding: From theory to practical application in the in vitro prediction of skin sensitization. *Contact Dermatitis* **2005**, *53*, 189–200.

(6) Schüürmann, G.; Somashekar, R. K.; Kristen, U. Structure−activity relationships for chloro- and nitrophenol toxicity in the pollen tube growth test. *Environ. Toxicol. Chem.* **1996**, *15*, 1702–1708.

(7) Schüürmann, G.; Aptula, A. O.; Kühne, R.; Ebert, R.-U. Stepwise discrimination between four modes of toxic action of phenols in the *Tetrahymena pyriformis* assay. *Chem. Res. Toxicol.* **2003**, *16*, 974–987.

(8) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. *pK_a Prediction for Organic Acids and Bases*; Chapman and Hall: London, 1981.

(9) Ullmann, G. M. Relations between protonation constants and titration curves in polyprotic acids: A critical view. *J. Phys. Chem. B* **2003**, *107*, 1263–1271.

(10) Lee, A. C.; Crippen, G. M. Predicting pK_a. *J. Chem. Inf. Model.* **2009**, *49*, 2013–2033.

(11) *ACD/Labs*, version 12.0; Advanced Chemistry Development, Inc.: Toronto, Ontario, Canada.

(12) 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.

(13) 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.

(14) 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.

(15) 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.

(16) 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.

(17) 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.

(18) 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.

(19) 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.

(20) Yu, H.; Kühne, R.; Ebert, R.-U.; Schüürmann, G. Comparative analysis of QSAR models for predicting pK_a of organic oxygen acids and nitrogen bases from molecular structure. *J. Chem. Inf. Model.* **2010**, *50*, 1949–1960.

(21) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(22) Klamt, A. Estimation of gas-phase hydroxyl radical rate constants of organic compounds from molecular orbital calculations. *Chemosphere* **1993**, *26*, 1273–1289.

(23) Klamt, A. Estimation of gas-phase hydroxyl radical rate constants of oxygenated compounds based on molecular orbital calculations. *Chemosphere* **1996**, *32*, 717–726.

(24) Böhnhardt, A.; Kühne, R.; Ebert, R.-U.; Schüürmann, G. Indirect photolysis of organic compounds: prediction of OH reaction rate constants through molecular orbital calculations. *J. Phys. Chem. A* **2008**, *112*, 11391–11399.

(25) Böhnhardt, A.; Kühne, R.; Ebert, R.-U.; Schüürmann, G. Predicting rate constants of OH radical reactions with organic substances: Advances for oxygenated organics through a molecular orbital HF/6-31G** approach. *Theor. Chem. Acc.* **2010**, *127*, 355–367.

(26) 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.

(27) 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.

(28) 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.

(29) Schwöbel, J.; Ebert, R.-U.; Kühne, R.; Schüürmann, G. Prediction models for the Abraham hydrogen bond donor strength: Comparison of semi-empirical, ab initio and DFT methods. *J. Phys. Org. Chem.* **2011**, DOI: 10.1002/poc.1834, in press.

(30) Wondrousch, D.; Böhme, A.; Thaens, D.; Ost, N.; Schüürmann, G. Local electrophilicity predicts the toxicity-relevant reactivity of michael acceptors. *J. Phys. Chem. Lett.* **2010**, *1*, 1605–1610.

(31) Schwöbel, J.; Wondrousch, D.; Koleva, Y.-K.; Madden, J.-C.; Cronin, M.; Schüürmann, G. Prediction of Michael-type acceptor reactivity toward glutathione. *Chem. Res. Toxicol.* **2010**, *23*, 1576–1585.

(32) Stewart, J. J. P. *MOPAC 2002*; Fujitsu Limited: Tokyo, 1999.

(33) Giesen, D. J.; Gu, M. Z.; Cramer, C. J.; Truhlar, D. G. A Universal Organic Solvation Model. *J. Org. Chem.* **1996**, *61*, 8720–8721.

(34) Hawkins, G. D.; Giesen, D. J.; Lynch, G. C.; Chambers, C. C.; Rossi, I.; Storer, J. W.; Li, J. B.; Zhu, T. H.; Thompson, J. D.; Winget, P.; Lynch, B. J.; Rinaldi, D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *AMSOL*, version 7.1, MN, 2004.

(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) Chatterjee, S.; Hadi, A.-S. Influential observations, high leverage points, and outliers in linear regression. *Stat. Sci.* **1986**, *1*, 379–416.

(37) Kühne, R.; Ebert, R.-U.; Schüürmann, G. Chemical domain of QSAR models from atom-centered fragments. *J. Chem. Inf. Model.* **2009**, *49*, 2660–2669.