Prediction of pK_a Values for Aliphatic Carboxylic Acids and Alcohols with Empirical Atomic Charge Descriptors†

Jinhua Zhang, Thomas Kleinöder, and Johann Gasteiger*

Computer-Chemie-Centrum and Institut für Organische Chemie, Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052, Erlangen, Germany

Received April 8, 2006

Two quantitative pK_a prediction models for aliphatic carboxylic acids and for alcohols were developed by multiple linear-regression (MLR) analysis with empirical atomic descriptors. The acid and alcohol molecules were described by a set of five and four atomic descriptors, respectively. For the pK_a model of 1122 aliphatic carboxylic acids, the squared correlation coefficient is 0.813 with a standard error of prediction of 0.423; for the pK_a model of 288 alcohols, the squared correlation coefficient is 0.817 with a standard error of prediction of 0.755, respectively. The good predictive abilities of the models obtained were indicated by both cross-validation and by external validation. An atomic descriptor was developed to model the inductive effect of the neighboring atoms for a central atom in a molecule. The ability of the descriptor to measure the inductive effect of substituent groups was demonstrated by a good correlation of this descriptor with Taft σ^* constants in aliphatic carboxylic acids. It provides a new approach to estimate Taft σ^* constants directly from molecular structures. An algorithm using Kohonen neural networks for splitting a data set into a training set and a test set is also presented.

INTRODUCTION

As one of the fundamental properties of an organic molecule the pK_a value determines the degree of dissociation in aqueous solution. The prediction of pK_a values for organic molecules is in great demand for rational drug design.[1] Some important properties of drugs, such as lipophilicity, solubility, and permeability, are all pK_a dependent. These properties have a profound influence on drug absorption and transport. pK_a is also important for understanding how a charged drug interacts with a receptor and in drug formulation for the choice of counterion and excipient. A method for the fast and accurate prediction of pK_a values from the chemical structure will greatly leverage the effort for screening large databases to find drug candidates.

Numerous pK_a prediction models have been developed with various approaches. The SPARC pK_a predictor applied a mechanistic perturbation method to estimate the pK_a through a number of models that accounted for electronic effects, solvation effects, hydrogen bonding effects, and the influence of temperature.[2] Xing and Glen used molecular tree structured fingerprints of key fragments and atom types in a hierarchical tree form to correlate pK_a values with the basic and acidic centers.[3,4] Dixon and Jurs built a QSPR pK_a model for oxy-acids with empirical atomic charges for protonated and deprotonated forms and several atom types to represent the charge stabilization abilities.[5] Schrödinger's Jaguar pK_a prediction module was built with an ab initio quantum chemical method and the self-consistent reaction field continuum treatment of solvation.[6] The model of ChemSilico's *CS*pK_a predictor was developed using artificial neutral networks and topological and E-state descriptors to encode molecules.[7] The ACD/pK_a module used fragment methods to build a large number of equations with experimental or calculated electronic constants to predict pK_a values.[8] Despite the numerous efforts that have been made, fast and accurate pK_a prediction models are still in demand for drug design in practice, due to the complexity of modeling acidity and the diversity of organic structures.

The principle of structure−property relationships is a basic concept in organic chemistry. A further important concept organic chemists use for interpreting reaction mechanisms is based on the distribution of partial charges in molecules. Organic chemists have a long history of estimating the relative acidity or reactivity of organic compounds by using the concept of partial charge. They can predict the relative acidity or reactivity by estimating the extent of delocalization of the charges based on molecular structure information. For a certain reaction, the structures of the substituents which can help in delocalizing the charges in the reaction intermediates or transition states will facilitate the reactions via that path. Historically, this approach is mostly used qualitatively to interpret the acidity or reactivities (e.g. why one mechanism is more favored than others). Quantitative methods developed with atomic partial charges have proved to be very successful in many quantitative structure−activity/property relationship (QSAR/QSPR) studies. Atomic charge descriptors that encode the electronic effects for substituents or the surrounding electronic properties of a reactive center will be very useful in modeling the acidity as well as the reactivity of organic compounds.

In this paper, we report a new quantitative structure−property relationship (QSPR) approach for the prediction of pK_a values of aliphatic carboxylic acids and of alcohols from their chemical structures with only empirical atomic charges

† Part of the Professor Johann Gasteiger special issue.

* Corresponding author phone: (49)-9131-85-26570; fax (49)-9131-85-26566; e-mail: Gasteiger@chemie.uni-erlangen.de.

10.1021/ci060129d CCC: $33.50
Published on Web 10/18/2006

and topological descriptors. The atomic charge descriptors used in these models reflect the intrinsic electronic properties of the acidic center and their substituents. The obtained models were validated with both cross-validation and external data set validation. A new algorithm which uses Kohonen neural networks for dividing a data set into a training set and a test set is also presented.

DATA SETS

A data set of 1123 aliphatic carboxylic acids was obtained from ChemSilico LLC. This data set is of a great structural diversity and includes 95 acids with pure alkyl (C, H only) substituents, 197 acids with halogen-containing substituents, 190 acids with unsaturated α-carbon atoms (sp or sp^2 hybridization), and 144 acids with an α-amino group. Formic acid is the only molecule in the data set that does not have an α-carbon, and it was excluded due to the descriptors we choose. The actual data set we used to build the model consisted thus of 1122 aliphatic carboxylic acids.

A data set of 288 alcohols was also obtained from ChemSilico LLC. Among them 26 alcohols are α-hydroxy-carboxylic acids and 86 alcohols are 2-amino alcohols. Enols are not included in this data set.

The data set of Taft σ^* constants of 130 aliphatic neutral substituents was taken from Table 9.1 in *Lange*'s *Handbook of Organic Chemistry*.[9]

METHODS

The pK_a value quantitatively indicates the acidity of a molecule referring to a specified acidic center in the molecule. Acidity is inherently determined by the electronic properties of the acidic center and the influences of substituents. To quantitatively model the pK_a values of molecules, the descriptors used must reflect the structural differences of molecules.

Structure Representation and Descriptor Selection. The fundamental hypothesis of QSPR modeling is that the properties of molecules are determined intrinsically by their structures. In a QSPR model, molecules are represented by structural descriptors, and the property to be modeled is a function of these descriptors. A descriptor represents either a physicochemical property or structural characteristics of a molecule. The task of modeling is to build a correlation between the property to be modeled and the structural descriptors. The correlations between the property and descriptors may have a linear or a nonlinear relationship.

A molecule must be transformed into a numerical number or a fixed-length vector of values before it can be used to conduct QSPR studies. Although molecular sizes may vary to a large extent, the vector representation must be in a fixed length for all molecules in a data set in order to apply a data analysis method. Various approaches have been developed to represent the structures of molecules for QSPR studies.[10]

To build a predictive QSPR model, one of the most important tasks is to select descriptors that are significantly related to the property of interest. There are two basic approaches for the selection of descriptors. The first one uses domain knowledge for calculating and selecting descriptors deemed important for the property of interest. The second approach is to apply mathematical methods for the selection step. Some practical approaches, such as stepwise inclusion, stepwise exclusion, and genetic algorithms have been applied for descriptor selections. In practice, these approaches will also be used in combination with our chemistry knowledge. In this study, we used the traditional stepwise inclusion method in combination with knowledge of the types of influences on acidity to select descriptors. Besides the consideration of those available molecular descriptors, it is always required to develop new molecular descriptors to meet the different requirements of various studies.

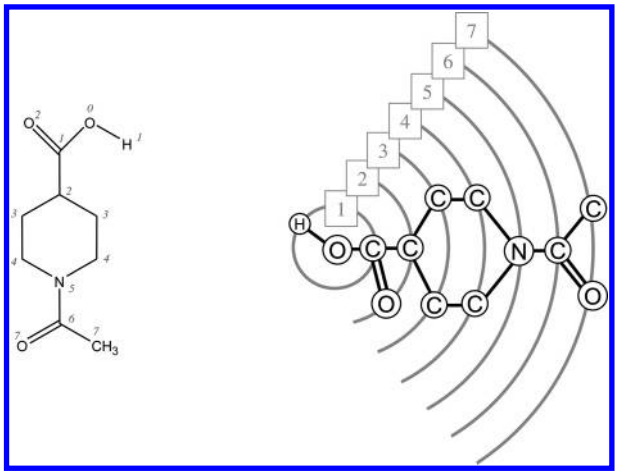


Figure 1. Graphical representation of the 7 topological spheres of *N*-acetylpiperidine-4-carboxylic acid centering at the hydroxyl oxygen atom (only the non-hydrogen atoms are shown).

a. Representation of Molecules with Topological Spheres. An organic molecule can be viewed as consisting of topological spheres with a specific atom as the center and all other atoms being at different levels of spheres of bond numbers around that central atom. According to this point of view, for example, *N*-acetylpiperidine-4-carboxylic acid can be represented as shown in Figure 1 with the hydroxyl oxygen atom as the center.

With the topological sphere representation, we can develop atomic descriptors that represent the molecules with a central atom being the reactive center. These descriptors are useful especially for modeling properties that are related to the reactivity of molecules with atoms being further away from the reactive center having less influence on reactivity.

b. PETRA Descriptors. The PETRA (Parameter Estimation for the Treatment of Reactivity Applications) program[11] collects a number of methods for the calculation of atomic, bond, and molecular descriptors such as charge distribution, polarizability, inductive and resonance effects, and other physicochemical properties. Many of these descriptors are calculated with the empirical partial equalization of orbital electronegativity (PEOE) method.[12] Many applications proved that the PETRA descriptors are very useful in modeling molecular activities and properties.

PETRA descriptors are calculated based on the neutral form of a molecule. Since we used some PETRA descriptors and molecular descriptors derived from PETRA descriptors to build our pK_a models, the structures of the acids of the model were in their neutral forms. For those acids where the acidic structures were not in the neutral forms, such as α-amino acids, we applied indicator descriptors to distinguish them from other neutral molecules.

c. Inductive Descriptor, Q_σ. To represent the impact of atoms from different topological spheres on the central atom,

we define an atomic charge descriptor, $Q_{\sigma,i}$, for a central atom i in a molecule, as the summation of the partial atomic sigma charges q_σ of all atoms from spheres with distance $d = 1..7$ with an attenuation factor of $1/d^2$

$$Q_{\sigma,i} = \sum_{d=1}^{7} \sum_{j \in \mathrm{TS}_d} \frac{q_{\sigma,j}}{d^2} \quad (1)$$

where d is the number of bonds (i.e. spheres) from an atom j (in a topological sphere TS_d) to the central atom i, and $q_{\sigma,j}$ is the atomic partial sigma charge on atom j calculated with the PEOE method.[12] By applying the factor $1/d^2$, $Q_{\sigma,i}$ should reflect the fact that an atom which is further away from the central atom will have less influence on the central reactive site.

Since the concept of atomic partial sigma charges is derived from electronegativity, it can be regarded as a quantitative measure of the sigma electron-withdrawing ability (i.e. inductive effect) of an atom. Similarly, the derived atomic descriptor $Q_{\sigma,i}$ can be used as a quantitative measure of the sigma electron-withdrawing or donating ability of neighboring atoms for a central atom. A larger positive $Q_{\sigma,i}$ value indicates that the neighboring atoms are more sigma electron-deficient and therefore having a higher potential to withdraw sigma electrons from the central atom while a larger negative value of $Q_{\sigma,i}$ means that the neighboring atoms have higher potential to donate sigma electrons to the central atom.

In aliphatic acid systems, the main electronic effect of a substituent is the inductive effect. This effect depends on the residual electronegativity of the substituents and decreases quickly as the distance from the substituents to the acidic center is increased. To investigate how the derived atomic descriptor $Q_{\sigma,i}$ can be used as a quantitative measure of a substituents' inductive effect, we built a data set consisting of 130 aliphatic carboxylic acids (R−COOH) for the substituents as their Taft σ^* constants were given in Table 9.1 of *Lange*'s *Handbook of Organic Chemistry*[9,13] and studied the correlation between $Q_{\sigma,i}$ for the acidic oxygen atoms in acids with Taft σ^* constants of substituents. The structures of 130 substituents, their Taft sigma constants, and the calculated $Q_{\sigma,o}$ for the acidic oxygen atoms in the corresponding acids are summarized in Table 1.

Linear regression analysis shows that the Taft σ^* constants and the inductive descriptor $Q_{\sigma,o}$ of the acidic oxygen atoms in 130 acids are quite well correlated. A simple regression equation was obtained:

$$\sigma^* = 24.05^* Q_{\sigma,o} - 11.04 \quad (2)$$

Calculated Taft σ^* constants obtained with eq 2 give a correlation with the experimental values with a squared correlation coefficient $R^2 = 0.848$ and a standard error of prediction $s = 0.436$. The *F-statistic* of the model is 385. In Figure 2 the atomic descriptor $Q_{\sigma,o}$ is plotted against experimental Taft σ^* constants.

For σ^* values higher than about 1.3 the correlation becomes less pronounced. Many of the substituents having σ^* values higher than 1.3 contain π-systems, in quite a few cases directly attached to the COOH group. It is believed that the $Q_{\sigma,o}$ values do not sufficiently well catch the inductive effects of these substituents which also exert an inductive effect through their π-charges. In separate studies we are investigating pK_a values of benzoic acids and phenols and are deriving models for their predictions.[14] Within these studies we are currently developing descriptors based on the π-charge distribution in order to tackle this deficiency of $Q_{\sigma,o}$ to capture inductive effects of π-systems.

Taft σ^* constants as well as Hammett σ constants proved to be very useful in characterizing electronic effects of substituents. However, experimentally determined σ^* values are only available for a small number of substituents, and their usage is limited when no experimental value is available.[15,16] Since $Q_{\sigma,o}$ is calculated directly from the chemical structures, eq 2 provides a general formula to estimate Taft σ^* constants for general neutral substituents. The descriptor $Q_{\sigma,o}$ can also be used to estimate the Hammett sigma constant σ_{meta} since σ_{meta} and σ^* are related by the expression $\sigma_{\mathrm{meta}} = 0.217\sigma^* - 0.106$.

d. Steric Descriptors, $A_{\mathrm{access},i}$(2D) and $A_{\mathrm{access},i}$ (3D). The steric effect is another factor that has influence on the pK_a values besides the electronic effects as steric effects will influence the ease of solvation of the acid and the anion. For this consideration, we propose two atomic accessibility descriptors to model the steric effect in organic molecules.

The accessibility to the central atom can be estimated from the degree of congestion of other atoms around the central atom in a molecule in 2D or in 3D space by eqs 3 and 4, respectively

$$A_{\mathrm{access},i}(2\mathrm{D}) = \sum_{n=1}^{5} \sum_{j \in \mathrm{TS}_n} \frac{V_{\mathrm{rel},j}}{N_n} \quad (3)$$

$$A_{\mathrm{access},i}(3\mathrm{D}) = \sum_{r_{ij}>0} \frac{V_{\mathrm{vdw},j}}{r_{ij}^2} \quad (4)$$

where $V_{\mathrm{rel},j}$ is the relative *van der Waals* volume for an atom j to a carbon atom; N_n is the number of carbon atoms at the topological sphere n (TS_n) in a diamond lattice with $N_n = \{4, 12, 32, 88, 240\}$; $V_{\mathrm{vdw},j}$ is the *van der Waals* volume of an atom calculated from the *van der Waals* radius; and r_{ij} is the distance between atoms i and j in a certain 3D conformation generated by the CORINA program.[17]

The difference between $A_{\mathrm{access},i}$ (2D) and $A_{\mathrm{access},i}$ (3D) is that $A_{\mathrm{access},i}$ (2D) is calculated from the topological graph according to the topological distance (number of bonds), while $A_{\mathrm{access},i}$ (3D) is calculated from a 3D structure.

e. E-State Index. For the purpose of comparison of the modeling capabilities of the inductive descriptor, Q_σ, introduced above we also calculated E-state indices for the acidic oxygen atoms. The E-state index, S_i, of an atom i is defined as

$$S_i = I_i + \sum_{j=1}^{N} \frac{I_i - I_j}{(d_{ij}+1)^2}, \quad \text{with } I = \frac{[(2/L)^2\delta^v + 1]}{\delta} \quad (5)$$

where L is the principle quantum number of a considered atom, δ^v is the number of valence electrons, δ is the number of sigma electrons, and d_{ij} is the topological distance between two atoms, i and j.[18] The E-state index has been used as descriptor for predicting a wide variety of properties such as aqueous solubility,[19,20] human oral absorption,[20] or Ames

Table 1. Taft σ^* Constants of 130 Substituents and the Calculated $Q_{\sigma,o}$ for the Acidic Oxygen Atoms in the Corresponding Carboxylic Acids

no.	substituent	Taft constant, σ^*	$Q_{\sigma,o}$ for **R**-COO*H	no.	substituent	Taft constant, σ^*	$Q_{\sigma,o}$ for **R**-COO*H
1	−Br	2.84	0.5468	66	−I	2.46	0.5156
2	$-CH_2Br$	1.00	0.4867	67	$-CH_2I$	0.85	0.4766
3	$-CH_3$	0.0	0.4573	68	$-N_3$(azide)	2.62	0.5833
4	$-CH_2CH_3$	−0.10	0.4577	69	$-NH_2$	0.62	0.4939
5	$-CH_2CH_2CH_3$	−0.12	0.4574	70	$-CH_2NH_2$	0.50	0.4686
6	$-CH(CH_3)_2$	−0.19	0.4580	71	$-N(CH_3)_2$	0.32	0.4896
7	$-CH_2CH_2CH_2CH_3$	−0.13	0.4570	72	$-NHCOCH_3$	1.40	0.5171
8	$-CH_2CH(CH_3)_2$	−0.13	0.4570	73	$-NHCOC_2H_5$	1.56	0.5168
9	$-CH(CH_3)CH_2CH_3$	−0.19	0.4577	74	$-NHCOC_6H_5$	1.68	0.5176
10	$-C(CH_3)_3$	−0.30	0.4582	75	−NHCHO	1.62	0.5181
11	$-CH_2CH_2CH(CH_3)_2$	−0.17	0.4566	76	$-NHCONH_2$	1.31	0.5206
12	$-CH_2CH_2CH_2CH_2CH_3$	−0.25	0.4567	77	$-NHCOOC_2H_5$	1.99	0.5260
13	$-CH_2C(CH_3)_3$	−0.12	0.4565	78	$-CH_2NHCOCH_3$	0.43	0.4777
14	$-CH_2(CH_2)_5CH_3$	−0.37	0.4558	79	$-NHSO_2C_6H_5$	1.99	0.5321
15	cyclohexyl-	−0.15	0.4577	80	−CN	3.30	0.5691
16	$-CH{=}CH_2$	0.56	0.4741	81	$-CH_2CN$	1.30	0.4935
17	$-CH{=}C(CH_3)_2$	0.19	0.4731	82	$-NO_2$	4.0	0.6245
18	$-CH{=}CHCH_3$, *trans*	0.36	0.4737	83	$-CH_2NO_2$	1.40	0.5128
19	$-CH_2CH{=}CH_2$	0.0	0.4621	84	$-CH_2CH_2NO_2$	0.50	0.4771
20	$-CH{=}CHC_6H_5$	0.41	0.4741	85	$-N(COCH_3)(COC_6H_5)$	1.37	0.5415
21	$-C{\equiv}CH$	2.18	0.5195	86	$-N(COCH_3)$(naphthyl)	1.65	0.5218
22	$-C{\equiv}CC_6H_5$	1.35	0.5196	87	−OH	1.34	0.5530
23	$-CH_2C{\equiv}CH$	0.81	0.4768	88	$-OCH_3$	1.81	0.5489
24	$-C_6H_5$	0.60	0.4823	89	$-OC_2H_5$	1.68	0.5484
25	naphthyl-(−1)	0.75	0.4839	90	$-OC_3H_7$	1.68	0.5480
26	naphthyl-(−2)	0.75	0.4819	91	$-OCH(CH_3)_2$	1.62	0.5479
27	$-CH_2C_6H_5$	0.22	0.4637	92	$-OC_4H_9$	1.68	0.5478
28	$-CH_2CH_2C_6H_5$	−0.06	0.4575	93	−O-cyclopentyl	1.62	0.5477
29	$-CH(CH_3)C_6H_5$	0.37	0.4639	94	−O-cyclohexyl	1.81	0.5473
30	$-CH(C_6H_5)_2$	0.41	0.4699	95	$-OCH_2$-cyclohexyl	1.31	0.5464
31	$-CH_2C_{10}H_7$	0.44	0.4630	96	$-OC_6H_5$	2.43	0.5583
32	3-indolyl-	−0.06	0.4901	97	$-ONO_2$	3.86	0.6113
33	2-thienyl-	1.31	0.4973	98	$-ON{=}C(CH_3)_2$	1.81	0.5343
34	2-thienylmethylene-	0.31	0.4684	99	$-CH_2OH$	0.31	0.4868
35	$-COCH_3$	1.65	0.5402	100	$-CH_2OCH_3$	0.52	0.4849
36	$-COCF_3$	3.7	0.5881	101	$-CH(OH)CH_3$	0.12	0.4868
37	$-COC_6H_5$	2.2	0.5476	102	$-CH(OH)C_6H_5$	0.50	0.4930
38	$-CONHC_6H_5$	1.56	0.5562	103	$-CH_2CH(OH)CH_3$	−0.06	0.4648
39	$-CONH_2$	1.68	0.5570	104	$-CH_2C(OH)(CH_3)_2$	−0.25	0.4642
40	$-CH_2COCH_3$	0.60	0.4829	105	−SH	1.68	0.5083
41	$-CH_2CONH_2$	0.31	0.4867	106	$-SCH_3$	1.56	0.5071
42	$-CH_2CH_2CONH_2$	0.19	0.4677	107	$-SCH_2CH_3$	1.56	0.5068
43	$-CH_2CH_2CH_2CONH_2$	0.12	0.4618	108	$-SCH_2CH_2CH_3$	1.49	0.5064
44	$-CH_2CONHC_6H_5$	0.0	0.4844	109	$-SCH_2CH_2CH_2CH_3$	1.44	0.5061
45	−COOH	2.08	0.5804	110	−S-cyclohexyl	1.93	0.5058
46	$-COOCH_3$	2.00	0.5780	111	$-SC_6H_5$	1.87	0.5154
47	$-COOCH_2CH_3$	2.12	0.5775	112	$-SC(C_6H_5)_3$	0.69	0.5066
48	$-CH_2COOCH_2CH_3$	0.82	0.4923	113	$-SCH_2C_6H_5$	1.56	0.5070
49	$-CH_2COOCH_3$	1.06	0.4927	114	$-SCH_2CH_2C_6H_5$	1.44	0.5048
50	−Cl	2.96	0.5636	115	$-CH_2SH$	0.62	0.4708
51	$-CCl_3$	2.65	0.5627	116	$-CH_2SCH_2C_6H_5$	0.37	0.4681
52	$-CHCl_2$	1.94	0.5273	117	−SCN	3.43	0.5478
53	$-CH_2Cl$	1.05	0.4923	118	$-SCONH_2$	2.07	0.5387
54	$-CH_2CH_2Cl$	0.38	0.4685	119	$-SOC_6H_5$	3.24	0.5537
55	$-CH_2CCl_3$	0.75	0.4902	120	$-CH_2SOCH_3$	1.33	0.4857
56	$-CH_2CH_2CCl_3$	0.25	0.4705	121	$-SO_2CH_3$	3.68	0.5849
57	$-CH{=}CCl_2$	1.00	0.4982	122	$-SO_2CH_2CH_3$	3.74	0.5845
58	$-CH_2CH{=}CCl_2$	0.19	0.4719	123	$-SO_2CH_2CH_2CH_3$	3.68	0.5841
59	−F	3.21	0.6030	124	$-SO_2C_6H_5$	3.55	0.5920
60	$-CF_3$	2.61	0.6093	125	$-CH_2SO_2CH_3$	1.38	0.4985
61	$-CHF_2$	2.05	0.5568	126	$-Si(CH_3)_3$	−0.81	0.4418
62	$-CH_2F$	1.10	0.5069	127	$-Si(CH_3)_2C_6H_5$	−0.87	0.4491
63	$-CH_2CF_3$	0.90	0.5043	128	$-Si(CH_3)_2OSi(CH_3)_3$	−0.81	0.4737
64	$-CH_2CF_2CF_2CF_3$	0.87	0.5089	129	$-CH_2Si(CH_3)_3$	−0.25	0.4559
65	−H	0.49	0.4629	130	$-CH_2CH_2Si(CH_3)_3$	−0.25	0.4567

genotoxicity.[20] The definition of the E-state index shares some characteristics with our inductive descriptor, Q_o, and was therefore considered in the present study.

Training and Test Set Selection by Kohonen Neural Networks. Kohonen neural networks[21] proved to be a successful method for splitting a data set into a training set and a test set.[22,23] After the training of a Kohonen neural network, the generated two-dimensional feature map will preserve the similarity in the data. Molecules which are represented by molecular descriptor vectors are mapped into

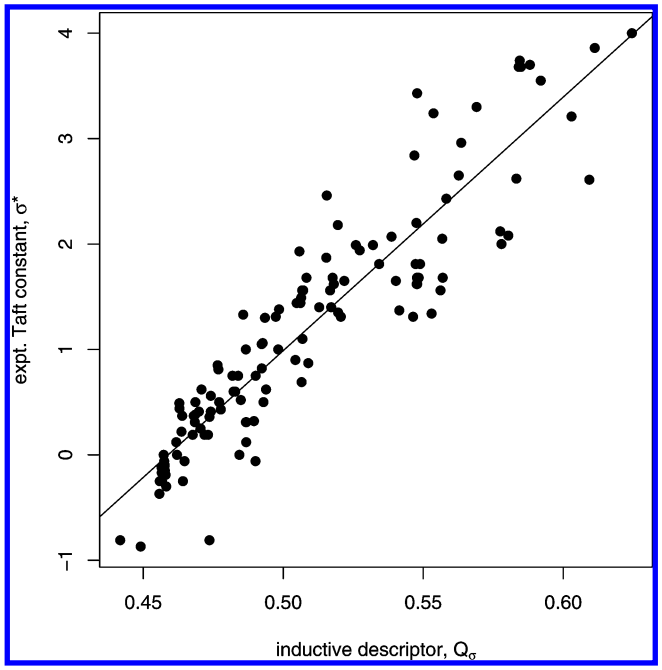


Figure 2. Inductive descriptor $Q_{\sigma,o}$ on the acidic oxygen atom of model compounds of type **R**COOH vs experimental Taft σ^* constants for 130 substituents, **R,** with $\sigma^* = 24.05 * Q_{\sigma,o} - 11.04$.

the various neurons of a Kohonen network with similar molecules being mapped into the same neuron or neurons close to each other in the two-dimensional map.

In the original algorithm using Kohonen networks to split a data set[22,23] (for simplicity we call it the original Kohonen network algorithm in this paper), for each occupied neuron, one molecule is assigned to the training set and the remaining molecule(s) from the same neuron is/are assigned to the test set. Since some neurons may be only occupied by one molecule and that will be assigned to the training set, the produced two subsets (the training set and the test set) may not cover the information space of the original data set as close as possible.

In this work we also developed a modified algorithm for using Kohonen networks to split a data set. In this algorithm, for each neuron that is occupied by multiple molecules, we randomly assigned half of those molecules into the training set and half of them into the test set. Those molecules in singly occupied neurons will be used for training a smaller sized Kohonen network. This procedure is repeated until less than 10 molecules are unassigned, and then we randomly assign half of the unassigned molecules into the training set and half of them into the test set. This modified Kohonen network algorithm may help us to obtain two subsets which can cover the information space of the original one more closely. For the two data sets, although we call one the training set and the other the test set, in practice we build a model with one set and test the built model with the other set and vice versa. If the two subsets cover a similar information space, then the equations and the statistics for the two models as well as the validation statistics should not be much in variance. The scheme of the modified algorithm using Kohonen networks to split a data set into a training set and a test set is illustrated in Figure 3 (the modified Kohonen network algorithm).

While applying Kohonen networks to split a data set, a selected 29 molecular descriptor vector was used to represent

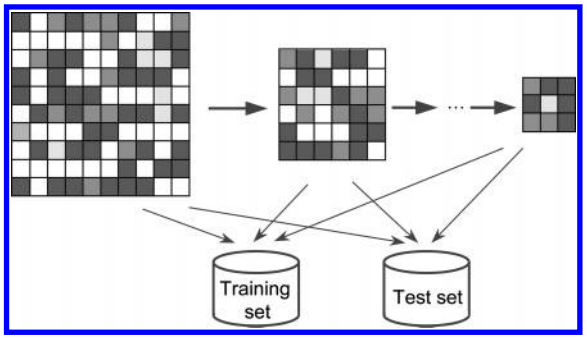


Figure 3. Scheme for splitting a data set into a training set and a test set with the modified algorithm using Kohonen neural networks.

a molecule. These molecular descriptors can represent molecules in various aspects, including physicochemical properties and electronic and topological characteristics. Table 2 gives detailed information on these descriptors.

Software Used. Several software packages have been used to conduct this study. Most of the work was carried out on our newly developed MOSES and VANESSA software packages. MOSES is a new programming framework written in C++ that can handle a variety of chemical structure formats and chemical reactions and which integrates software packages for property prediction such as CORINA[17] for 3D structure generation, PETRA[11] for physicochemical descriptor calculation, and methods for the calculation of other topological descriptors. VANESSA is an application software tool with a graphical user interface (GUI) built upon the MOSES framework. It can handle most formats of chemical structure files as inputs, calculate the selected molecular descriptors, perform the data analysis, build QSAR models using multiple linear-regression analysis (MLRA) and/or back-propagation (BPG) neural networks, and validate built models by cross-validation or external data as well as predict properties with built models. The CACTVS molecular browser and editor[30] were used to generate and handle chemical structures. The SONNIA software package[31] was used to build Kohonen neural networks for splitting a data set into a training set and a test set. E-state indices were calculated with a calculation module written in-house based on our C++ framework MOSES.

RESULTS AND DISCUSSION

Comparison of Q_σ and E-State Index for the Entire Data Set. In a preliminary study we investigated the modeling capabilities of the inductive descriptor, Q_σ, introduced in the present study and the E-state index that should also catch the inductive effect. A correlation analysis was performed between the experimental pK_a values of the entire data set of 1410 carboxylic acids and alcohols, respectively, and Q_σ values and E-state indices calculated for the acidic oxygen atoms. The results are shown in Table 3.

The correlation of the experimental pK_a values with the simple descriptor Q_σ is surprisingly high attesting to the great benefits of this descriptor. This correlation is significantly higher than the one obtained for the E-state index. Both descriptors are attempting to capture the electronic effects of the chemical environment on the acidic center. Although the E-state indices are calculated only from quantum numbers and numbers of valence electrons (see eq 5), it was argued that the intrinsic atom value, I, defined for representing the

Table 2. Molecular Descriptors Used for Feature Representation in the KNN for Splitting a Data Set into a Training Set and a Test Set

descriptor	description
M_AUTOCORR2D(0−5)q$_\sigma$	a molecular autocorrelation vector with the topological distance from 0 to 5 and the atomic property of q_σ (including hydrogen atoms)
M_ASA	molecular approximate surface area[24]
M_ALIPHATICINDICATOR	number of sp^3 carbons divided by the total number of carbon atoms in a molecule
M_AROMATICINDICATOR	number of aromatic atoms divided by the number of total atoms in a molecule (excluding hydrogen atoms)
M_HATOM_COUNT	total number of hydrogen atoms in a molecule
M_CHI0	molecular connectivity index chi0
M_CHI1	molecular connectivity index chi1
M_KAPPA1	molecular shape index Kappa1
M_COMPLEXITY	molecular complexity[25]
M_DIPOLE	molecular dipole moment based on partial total charges and 3D Cartesian coordinates
M_NPISYS	number of π-systems of the molecule (only π-systems spanning over more than one atom)
M_NSIGMASYS	number of σ-systems of the molecule
M_HDONOR_COUNT	number of hydrogen bond donors (OH and NH bonds)
M_HACCEPTOR_COUNT	number of hydrogen bond acceptors (O and N atoms)
M_HIGHEST_HBOND_ACC_POT	the highest hydrogen bond acceptor potential (the total number of lone-pair electrons on N, O, F atoms)
M_HIGHEST_HBOND_DON_POT	the highest hydrogen bond donor potential (the total H atoms in the groups −OH, −NH, −SH)
M_INTERNALHBOND	number of internal hydrogen bonds[26]
M_POLARIZABILITY	mean molecular polarizabilities based on atomic increments[27]
M_RANDICCINDEX	molecular Randic connectivity index
M_RING_COMPLEXITY	ring complexity of a molecule[28]
M_TPSA	topological polar surface area[29]
M_WEIGHT	molecular weight
M_WIENERINDEX	molecular Wiener index
M_XLOGP	molecular logP, based on atomic increments (XLogP 2.0)[26]

Table 3. Correlations[a] between the Experimental pK_a and the Inductive Descriptor Q_σ and E-State Indices, Respectively, Calculated for the Acidic Oxygen Atoms of 1410 Aliphatic Oxy-Acids

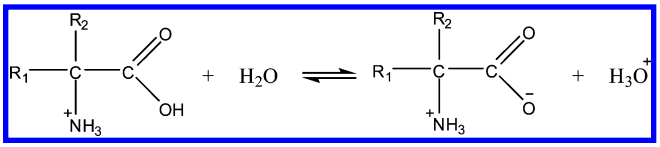
	Q_σ	E-state index
pK_a (exp)	−0.91	0.45
Q_σ		−0.45

[a] Given as Pearson correlation coefficients, r.

electronic nature of an atom in a molecule is correlated with valence-state electronegativity.[32] For defining the intrinsic electronic state in our approach we use the partial sigma charge obtained from partial equalization of orbital electronegativities (PEOE). The partial sigma charge is tightly related to the residual electronegativity of an atom representing the tendency of an atom to further attract or release electron density. The PEOE method is not only based on the topological information but has a sound physicochemical foundation in using valence-state orbital electronegativities obtained experimentally.[33−35] This higher level of physicochemical information is enclosed in our inductive descriptor, Q_σ, and explains its high correlation with the pK_a value. The opposite signs of the correlation coefficients are due to the fact that more electronegative atoms obtain a negative charge in the PEOE procedure while they are assigned a more positive intrinsic state value in calculating E-state indices.

The pK_a Model for Aliphatic Carboxylic Acids. We investigated the entire data set of carboxylic acids before building a model with a training data set and validating the model with a test set. To assist us in selecting descriptors that could mostly account for the acidity of the molecules and to build a predictive model, we performed an analysis of the chemical structures of acids in the data set.

Scheme 1. Dissociation of α-Amino Acids in Water



The inductive effect is the most important electronic effect which determines the pK_a value of an aliphatic carboxylic acid. Since the inductive descriptor $Q_{\sigma,o}$ can quantitatively measure the inductive effect of the aliphatic substituents, we started from this descriptor and applied a stepwise inclusion approach by including additional atomic descriptors to build a pK_a model for aliphatic carboxylic acids.

Except for formic acid, 1122 aliphatic carboxylic acids in the data set contain a α-carbon atom. The hybridization state of the α-carbon atom in an acid affects the delocalization of the negative charge on the carboxylate group (−COO$^-$) and therefore has influence on the pK_a value. The inclusion of π-electronegativity of the α-carbon atom, $\chi_{\pi,\alpha-C}$, can represent the different hybridization states (i.e. sp, sp^2, and sp^3) of the α-carbon atom in an acid. An amino acid contains both a basic −NH$_2$ group and an acidic −COOH group. During the dissociation of the carboxylic group of an amino acid in water, the molecule exists in an ionic form with a protonated amino group, as illustrated in Scheme 1. As an electron deficient group, a positively charged ammonium group is a stronger electron-withdrawing group than a neutral amino group. The values of the Taft σ^* constant and the Hammett σ_{meta} constant for an ammonium group are larger than those for an amino group (Table 4). Since the PEOE method we used to calculate atomic partial charges is based on the neutral molecular structures, there are deviations for using the calculated $Q_{\sigma,o}$ values for amino acids to represent the sigma charge environment for the oxygen atom in the

Table 4. Hammett Sigma Constants (σ_{meta}) and Taft σ^* Constants

substituent	Hammett constant, σ_{meta}	Taft constant, σ^*
$-NH_2$	−0.16	0.62
$-N^+H_3$	1.13	3.76
−COOH	0.36	2.08
$-COO^-$	−0.1	−1.06

Table 5. Selected 5 Descriptors and the Corresponding Regression Coefficients and *t-Statistic* in the Multiple Linear-Regression (MLR) Model for 1122 Aliphatic Carboxylic Acids

descriptor	coefficients	t-statistic
$Q_{\sigma,o}$	-37.54	62.1
$A_{access,o}$(2D)	12.27	25.1
$\chi_{\pi,\alpha c}$	0.11	17.6
α_o	−1.02	50.1
I_{amino}	−1.89	26.7

hydroxyl group in an acid. To take account of this situation in our MLR modeling, a classifier indicator, I_{amino}, was introduced to reflect the difference of $Q_{\sigma,o}$ values for α-amino acids and other acids. According to eq 1, the atomic contribution to the $Q_{\sigma,o}$ values decreases greatly with bond distance. An amino group attached to a α-carbon will have a large influence on the central oxygen atom, and this influence for an amino group attached to other carbons in the skeleton can be ignored. We assigned 1 to I_{amino} for α-amino acids and assigned 0 to I_{amino} for other amino acids as well as all other acids.

The best multiple linear-regression equation is obtained from five molecular descriptors: $Q_{\sigma,o}$, the inductive descriptor of the acidic oxygen atom in an acid; $A_{access,o}$(2D), the accessibility of the acidic oxygen atom in an acid; α_o, the polarizability of the acidic oxygen atom in an acid; $\chi_{\pi,\alpha c}$, the π-electronegativity for the α-carbon atom in an acid; and I_{amino}, for an α-amino acid the value is 1, otherwise it is 0.

A correlation analysis showed that no Pearson correlation coefficient (r) between any two descriptors in these five descriptors is larger than 0.6 for the entire set of 1122 aliphatic carboxylic acids.

The MLR model for aliphatic carboxylic acids is given by eq 6

$$pK_a = -37.54Q_{\sigma,o} + 12.27A_{access,o}(2D) + 0.11\chi_{\pi,\alpha c} - 1.02\alpha_o - 1.89I_{amino} + 19.10 \quad (6)$$

with a squared correlation coefficient R^2= 0.813 and a standard error of prediction s = 0.423. The *F-statistic* of the model is 809. The corresponding regression coefficients and *t-statistics* of five descriptors are summarized in Table 5.

Figure 4 shows a plot of predicted vs experimental pK_a values of the 1122 aliphatic carboxylic acids. The pK_a model for aliphatic carboxylic acids is mechanistically interpretable and intuitively consistent with the physicochemical meaning of the descriptors. As $Q_{\sigma,o}$ is a measure of the electronic inductive effect of substituents with the same physical meaning as the Taft σ^* constant, the negative sign of the coefficient of $Q_{\sigma,o}$ in eq 6 indicates that an aliphatic acid will have a small pK_a value if it has a large $Q_{\sigma,o}$ value; in other words a higher value of the inductive effect leads to higher acidity. The positive sign for $A_{access,o}$(2D) means that a large steric hindrance will increase the pK_a and therefore decrease the acidity of the acid as access of water

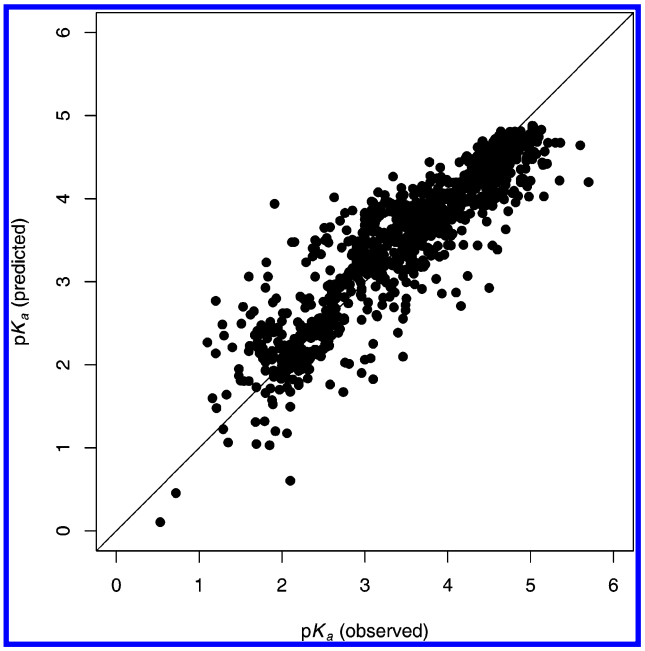


Figure 4. Predicted vs observed pK_a values of 1122 aliphatic carboxylic acids (R^2 = 0.813, s = 0.423, F = 809).

Table 6. Statistical Data for the pK_a Model of 1122 Aliphatic Carboxylic Acids

no. of molecules	no. of descriptors	model		5-fold cross-validation	
		R^2	s	R_{cv}^2	s
1122	5	0.813	0.423	0.810	0.426

to take up the proton is hindered. In our studies on building pK_a models for aliphatic carboxylic acids both descriptors $A_{access,o}$(2D) and $A_{access,o}$(3D) performed equally well.

The predictive power of the pK_a model is evaluated with both 5-fold cross-validation and external data set validation. During cross-validation, the total data set of 1122 molecules is divided into five groups; each of them is left out in turn to validate the model built with the other four groups. By repeating this procedure 100 times, we obtained an average cross-validated squared correlation coefficient R_{cv}^2 = 0.810 and an average standard error of prediction s = 0.426. Comparing these values to the squared correlation coefficient of the model R^2 = 0.813 and an s = 0.423 obtained from the entire data set, the small variance indicates that this is a well-behaved system and the pK_a model built should have a reasonable predictive power for the estimation of pK_a values of aliphatic carboxylic acids. Table 6 summarizes the statistical data for the pK_a model built with 1122 aliphatic carboxylic acids.

We then built a model with a training set and a test set. Both the original algorithm using Kohonen networks and its modification as described in the previous section were used to divide the 1122 data set into a training set and a test set. Each molecule was represented by a vector of 29 molecular descriptors, and this vector was used as input into the Kohonen network constructed with the SONNIA software. Details on the procedure of training Kohonen neural networks have been described in the literature.[22,23]

While applying the original algorithm using Kohonen networks, the 1122 aliphatic carboxylic acid data set was divided into two subdata sets of 612 and 510 molecules. With the modified algorithm, we obtained two data sets of 610

Table 7. Statistical Data for the pK_a Model of Aliphatic Carboxylic Acids Built with a Training Set and Validated with a Test Set[a]

	training set			test set		
algorithm	n	R^2	s	n	R^2	s
original KNN algorithm	612	0.801	0.435	510	0.822	0.415
	510	0.831	0.402	612	0.792	0.446
modified KNN algorithm	610	0.811	0.403	512	0.808	0.455
	512	0.820	0.437	610	0.795	0.421

[a] The training sets and test sets are created with the two algorithms using Kohonen neural networks (KNN). n: number of compounds, R^2: squared correlation coefficient, s: standard error of prediction.

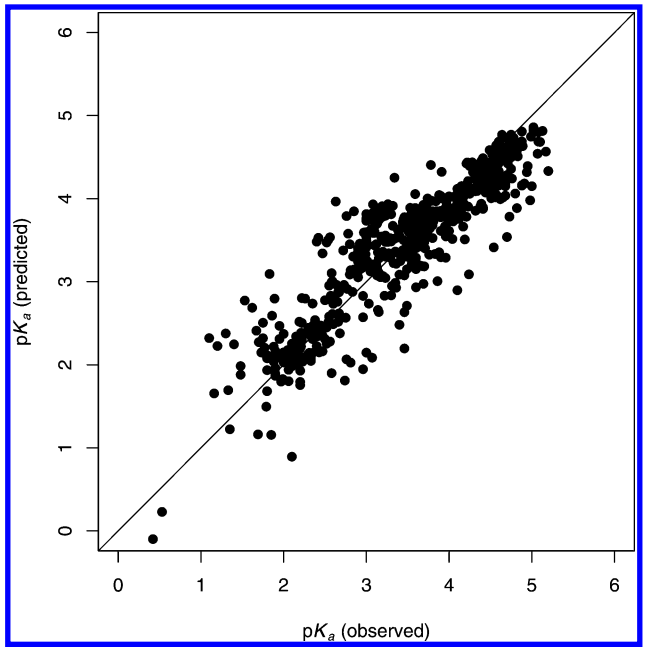


Figure 5. Predicted vs observed pK_a values of 610 aliphatic carboxylic acids in the training set ($n = 610$, $R^2 = 0.811$, $s = 0.403$, $F = 431$).

and 512 molecules. One data set was used to build a model, and the other one was used to test the built model. The statistical data of the models and the validations are summarized in Table 7.

Figures 5 and 6 show plots of predicted vs experimental pK_a values of the training set and the test set created by the modified algorithm. Figure 5 shows the plot of 610 aliphatic carboxylic acids as the training set to build a pK_a model with 5 descriptors. Figure 6 gives the plot of predicted vs experimental pK_a values of 512 aliphatic carboxylic acids as a test set for the model built with 610 training set. The distribution of the predicted errors of the 1122 aliphatic carboxylic acids is shown in Figure 7. For 80% of the total 1122 acids the prediction errors are within 0.5 units, and more than 95% errors are within 1.0 unit. Table 8 gives the structures of the carboxylic acids which have pK_a prediction errors larger than 1.5 units. In structures **1**−**3**, the carboxylic acid group is part of a conjugated π-system. We had already observed in the correlation of Q_σ with Taft σ^* constants that this descriptor apparently insufficiently expresses the influence of directly conjugated π-systems. For compound **4** we have doubts about the correctness of the experimental value. Taking the experimental pK_a values of CF_3CH_2COOH (3.1) and of lactic acid (3.9) and assuming roughly additivity of substituent effects we would expect an experimental pK_a value for compound **4** of about 0.9. From this point of view the pK_a value predicted with our model seems more consistent with other acids having a similar chemical structure. Also for compound **5** the experimental value seems to be doubtful. Carboxylic acids with large alkyl substituents have pK_a values in the range of 4.5−4.9. A carbonyl group in the γ-position as well as an aryl group in the δ-position should both enhance the acidity resulting in a lowering of the pK_a value rather than increasing it. A stabilization of the acid by hydrogen bonding seems improbable as this would ask for a seven-membered ring.

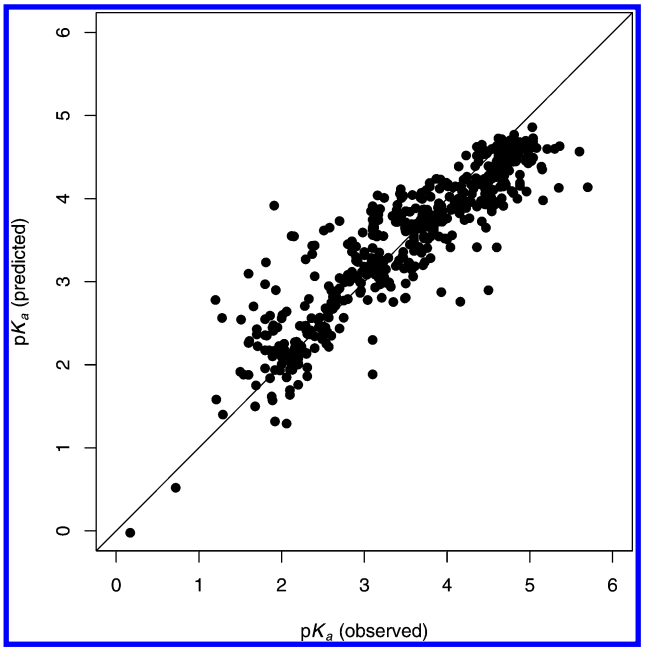


Figure 6. Predicted vs observed pK_a values of 512 aliphatic carboxylic acids in the test set for validating the model built with 610 aliphatic carboxylic acids (see Figure 5) ($n = 512$, $R^2 = 0.80$, $s = 0.455$).

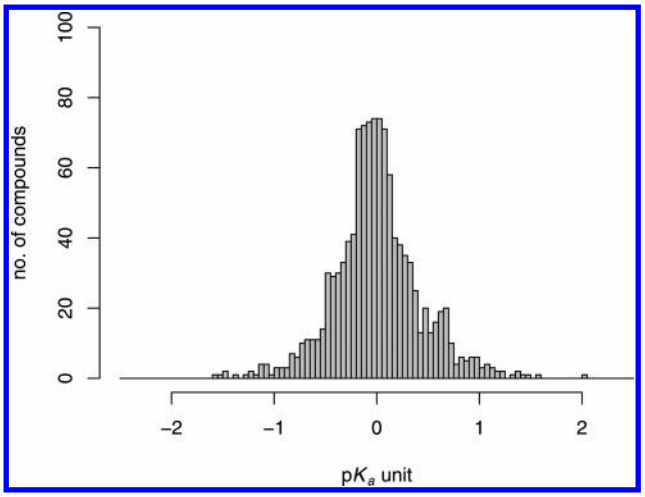
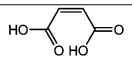
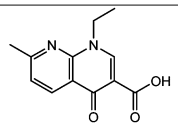
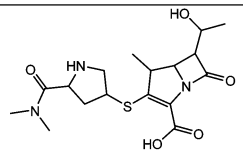
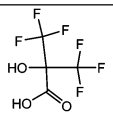
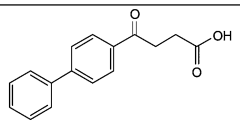


Figure 7. Distribution of pK_a prediction errors for 1122 aliphatic carboxylic acids.

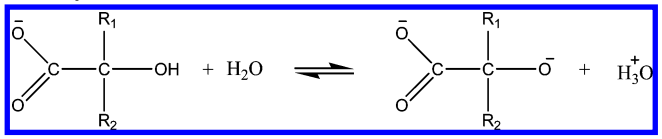
In summary, our model first allows the identification of physicochemical effects that have not yet been completely accounted for in our approach. Second, it also allows the identification of potential experimental errors.

The pK_a Model for Alcohols. While trying to build a MLR pK_a model with two atomic descriptors, $Q_{\sigma,o}$ and σ-electronegativity ($\chi_{\sigma,o}$), for 288 alcohols, we observed that

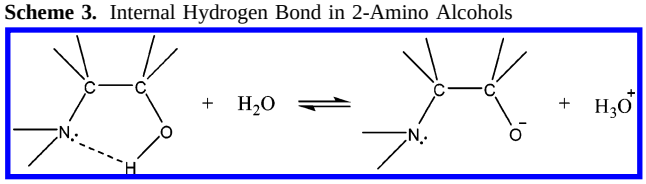
Table 8. Structures of 5 Aliphatic Carboxylic Acid Molecules with pK_a Prediction Errors Larger than 1.5 Units

no.	structure	pK_a (exp.)	pK_a (pred.)	ΔpK_a
1		1.91	3.95	+2.04
2		1.2	2.78	+1.58
3		4.5	2.94	-1.56
4		2.1	0.59	-1.51
5		5.7	4.20	-1.50

Scheme 2. Dissociation of the Alcoholic OH Group in α-Hydroxyl Carboxylic Acids in Water



Scheme 3. Internal Hydrogen Bond in 2-Amino Alcohols



two types of alcohols (α-hydroxy-carboxylic acids and 2-amino alcohols) display systematic large prediction errors. The predicted pK_a values for α-hydroxy-carboxylic acids are systematically lower than their experimental values, while the predicted pK_a values for 2-amino alcohols are systematically higher than their experimental values.

An α-hydroxy-carboxylic acid contains a more acidic −COOH group which will be easier to dissociate than the alcoholic −OH group. During the dissociation of the alcoholic −OH group in α-hydroxyl, the molecule is in a form of the negatively charged $-COO^-$, as illustrated in Scheme 2. As an electron abundant group, $-COO^-$ serves as an electron-donating group rather than the electron-withdrawing group as −COOH does. The Hammett σ_{meta} constant and the Taft σ^* constant for $-COO^-$ are all negative but are positive for −COOH in Table 4. A similar problem is encountered here as it was in the modeling of α-amino acids. This is due to the fact that $Q_{\sigma,o}$ is calculated from the neutral structural forms. Similar to the modeling of aliphatic carboxylic acids, we introduce a classifier indicator, $I_{carboxy}$, to take account of this situation.

In 2-amino alcohols, the internal hydrogen bonding effect may enhance the acidity of the −OH group, but the anion might even more be stabilized when a primary or secondary

Table 9. Statistical Data for the pK_a Model of Alcohol Compounds

no. of molecules	no. of descriptors	model		5-fold cross-validation	
		R^2	s	R_{cv}^2	s
288	4	0.817	0.755	0.805	0.780

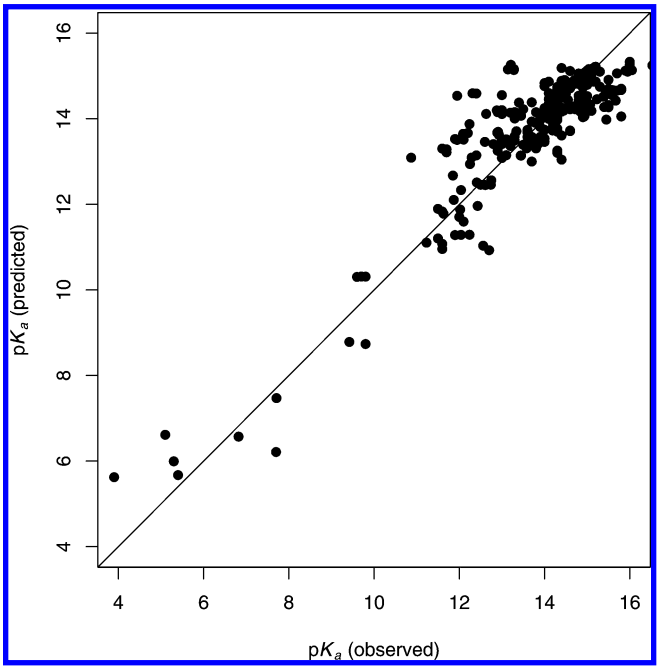


Figure 8. Predicted vs observed pK_a values of 288 alcohols (n = 288, R^2 = 0.817, s = 0.755, F = 252).

α-amino group is present (Scheme 3). Without representing this hydrogen bonding effect in the model, the predicted pK_a values for these types of molecules are systematically higher than their experimental values. A classifier descriptor, I_{amino}, is used to address this effect.

The best multiple linear-regression equation is obtained from four molecular descriptors: $Q_{\sigma,o}$, the inductive descriptor for the oxygen atom in the acidic hydroxyl group; $\chi_{\sigma,o}$, the σ-electronegativity for the oxygen atom in the acidic hydroxyl group; $I_{carboxy}$, for an α-hydroxy-carboxylic acid the value is 1, otherwise it is 0; and I_{amino}, for a 2-amino alcohols the value is 1, otherwise it is 0.

A correlation analysis showed that no Pearson correlation coefficient (r) between any two descriptors in these four descriptors is larger than 0.5 for the entire set of 288 alcohols.

The MLR model for alcohol is given by eq 7

$$pK_a = -19.07Q_{\sigma,o} - 4.68\chi_{\sigma,o} + 2.87I_{carboxy} - 0.66I_{amino} + 64.41 \quad (7)$$

with a squared correlation coefficient R^2 = 0.817, and a standard error of prediction s = 0.755. The *F-statistic* of the model is 252. The 5-fold cross-validation gave a cross-validated R_{cv}^2 = 0.805 and a standard prediction error s = 0.780 (Table 9). The corresponding regression coefficients and *t-statistic* of the four descriptors are summarized in Table 10. Figure 8 shows a plot of predicted vs experimental pK_a values of the 288 alcohols.

Similar to the pK_a equation for aliphatic carboxylic acids (eq 6), the negative coefficient sign for the atomic inductive descriptor $Q_{\sigma,o}$ in eq 7 is consistent with the physical meaning

Table 10. Selected 4 Descriptors and the Corresponding Regression Coefficients and *t-Statistic* in the Multilinear Regression (MLR) Model for Alcohols

descriptor	coefficients	t statistic
$Q_{\sigma,o}$	−19.07	31.2
$\chi_{\sigma,o}$	−4.68	9.3
I_{carbo}	2.87	18.5
I_{amino}	−0.66	6.5

Table 11. Statistical Data for the pK_a Model of Alcohols Built with a Training Set and Validated with a Test Set[a]

	training set			test set		
algorithm	*n*	R^2	*S*	*n*	R^2	*s*
original KNN algorithm	171	0.845	0.754	117	0.731	0.781
	117	0.744	0.745	171	0.836	0.787
modified KNN algorithm	143	0.818	0.794	145	0.836	0.769
	145	0.836	0.665	143	0.775	0.899

[a] The training sets and test sets are created with the two algorithms using Kohonen neural networks (KNN). *n*: number of compounds, R^2: squared correlation coefficient, *s*: standard error of prediction.

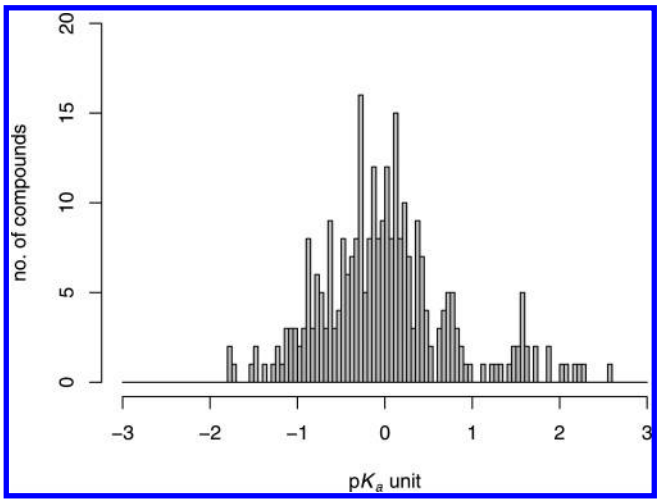


Figure 9. Distribution of pK_a prediction errors for 288 alcohols.

of this descriptor as a large $Q_{\sigma,o}$ value means a large inductive effect of the substituent.

The two Kohonen network algorithms were applied to split 288 alcohols into a training set and a test set. We built a model with one subset and validated the model with the other subset. The statistical results are summarized in Table 11.

The distribution of prediction errors is shown in Figure 9. For all 288 alcohol molecules, about 85% pK_a prediction errors are within 1.0 unit.

Comparison with Commercial Software. To compare our results for the calculation of pK_a values with other available software we have used the method provided by ChemSilico LLC on their Web site.[7] The pK_a values of 50 compounds can be calculated free of charge. We therefore randomly selected 50 structures from the entire data set of 1410 compounds such that the ratio of acids and alcohols (1122/288) was reflected by the test data set. Thus this test data set contained 37 acids and 13 alcohols. Our models were rebuilt excluding the selected compounds, and the pK_a values of the 50 test compounds were predicted with the newly built models. These values were compared with the results obtained with the ChemSilico Web service. The predicted pK_a values of our models show a standard error of prediction of $s = 0.48$, whereas the pK_a values obtained from ChemSilico have a slightly better standard error of $s = 0.41$. The slightly worse performance of our models can well be accepted if one takes into account that our models are based on only five and four descriptors for the acids and the alcohols, respectively, whereas the model of ChemSilico uses up to 70 descriptors.[36] This low number of descriptors in our approach underlines that we have explicitly modeled the physicochemical effects exerted onto acidity.

We presently extend our studies to handling acids that have charged sites (such as the amino acids or the amino alcohols) in order to remove the indicator variables and thus further reduce the number of descriptors.

CONCLUSIONS

In this study, two MLR pK_a prediction models for aliphatic acids and for alcohols were developed with five and four atomic descriptors, respectively. The obtained pK_a models are mechanistically interpretable and in consistence with intuitive chemistry knowledge. The predictive power was proved by both cross-validation and an external validation data set.

We demonstrated that the atomic inductive descriptor used for modeling pK_a has captured the substituents' inductive effect on the acidic center, and, therefore, it can also be used for QSAR/QSPR studies of reactivities in organic compounds. It provides a general approach to estimate Taft σ^* constants of substituents from molecular structures.

It was particularly gratifying that similar descriptor sets were able to model the pK_a values of both aliphatic carboxylic acids and aliphatic alcohols (eqs 6 and 7). The major differences lie in the use of different indicator variables required to account for the influence of charged groups and of intramolecular hydrogen bonding. We are presently extending the methods for charge calculation to also account for the influence of charged substituents thus hopefully obviating the use of indicator variables.

ACKNOWLEDGMENT

We thank Dr. Joe Votano of ChemSilico LLC for generously providing us with the data set for this study. We also acknowledge Dr. Stephan Sixt for some initial studies in this field.

REFERENCES AND NOTES

(1) *The Practice of Medicinal Chemistry*; Wermuth, C. G., Ed.; Academic Press: Amsterdam, 2003; pp 601−615.
(2) 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. http://ibmlc2.chem.uga.edu/sparc/ (accessed Aug 2006).
(3) Xing, L.; Glen, R. C. Novel Methods for the Predictions of log *P*, pK_a, and log *D*. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796−805.
(4) Xing, L.; Glen, R. C.; Clark, R. D. Predicting pK_a by Molecular Tree Structured Fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 870−879.
(5) Dixon, S. L.; Jurs, P. C. Estimation of pK_a for Organic Oxyacids Using Calculated Atomic Charges. *J. Comput. Chem.* **1993**, *14*, 1460−1467.
(6) http://www.schrodinger.com (accessed Aug 2006).
(7) http://chemsilico.com/CS_prpKa/PKAhome.html (accessed Aug 2006).
(8) http://www.acdlabs.com/products/phys_chem_lab/pka/batch.html (accessed Aug 2006).
(9) Dean, J. A. E. *Lange's Handbook of Chemistry*, 15th ed.; McGraw-Hill: 1998; pp 9.2−9.6.

(10) *Handbook of Chemoinformatics: From Data to Knowledge*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; pp 977−1078.
(11) PETRA program package. http://www2.chemie.uni-erlangen.de/software/petra (accessed Aug 2006). Available from Molecular Networks GmbH, Erlangen, Germany. http://www.mol-net.com (accessed Aug 2006).
(12) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − a Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(13) Two selenium containing substituents were not included in our data set due to the fact that the current PEOE implementation could not handle these compounds. The furoyl group was also not included in our data set because its Taft σ^* constant value given in this table was inconsistent with other literature.
(14) Pellegrini, E.; Kleinöder, T. Unpublished results.
(15) Hansch, C.; Leo, A.; Taft, R. W. A Survey of Hammett Substituent Constants and Resonance and Field Parameters. *Chem. Rev.* **1991**, *91*, 165−195.
(16) Ertl, P. Simple Quantum Chemical Parameters as an Alternative to the Hammett Sigma Constants in QSAR Studies. *Quant. Struct.-Act. Relat.* **1997**, *16*, 377−382.
(17) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(18) Kier, L. B.; Hall, L. H. An Electrotopological-state Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801−807.
(19) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.
(20) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B. New predictors for several ADME/Tox properties: Aqueous Solubility, Human Oral Absorption, and Ames Genotoxicity Using Topological Descriptors. *Mol. Diversity* **2004**, *8*, 379−391.
(21) (a) Kohonen, T. *Self-organization and Associative Memory*, 3rd ed.; Springer: Berlin, 1989; pp 105−176. (b) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, 1999; pp 81−100.
(22) Simon, V.; Gasteiger, J.; Zupan, J. A Combined Application of Two Different Neural Network Types for the Prediction of Chemical Reactivity. *J. Am. Chem. Soc.* **1993**, *115*, 9148−9159.
(23) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429−434.
(24) Labute, P. A Widely Applicable Set of Descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464−477.
(25) Hendrickson, J.; Huang, P.; Toczko, A. Molecular Complexity − A Simplified Formula Adapted to Individual Atoms. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 63−67.
(26) Wang, R.; Gao, Y.; Lai, L. Calculating Partition Coefficient by Atom-additivity Method. *Perspect. Drug Discovery Des.* **2000**, *19*, 47−66.
(27) Miller, K. J. Additivity Methods in Molecular Polarizability. *J. Am. Chem. Soc.* **1990**, *112*, 8533−8542.
(28) Gasteiger, J.; Jochum, C. An Algorithm for the Perception of Synthetically Important Rings. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 43−48.
(29) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(30) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Flexibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109−116.
(31) SONNIA is available from Molecular Networks GmbH, Erlangen, Germany. http://www.mol-net.com (accessed Aug 2006).
(32) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.
(33) Hinze, J.; Jaffé, H. H. Electronegativity. I. Orbital Electronegativity of Neutral Atoms. *J. Am. Chem. Soc.* **1962**, *84*, 540−546.
(34) Hinze, J.; Whitehead, M. A.; Jaffé, H. H. Electronegativity. II. Bond and Orbital Electronegativities. *J. Am. Chem. Soc.* **1963**, *85*, 148−154.
(35) Hinze, J.; Jaffé, H. H. Electronegativity. IV. Orbital Electronegativities of the Neutral Atoms of the Periods Three A and Four A and of Positive Ions of Period One and Two. *J. Phys. Chem.* **1963**, *41*, 1315−1328.
(36) http://www.chemsilico.com/CS_methods/methods.html (accessed Aug 2006).

CI060129D