New and Original pK_a Prediction Method Using Grid Molecular Interaction Fields

Francesca Milletti,[†] Loriano Storchi,[‡] Gianluca Sforna,[‡] and Gabriele Cruciani*[,†]

Laboratory for Chemometrics and Cheminformatics, Department of Chemistry, Università degli Studi di Perugia, via Elce di Sotto 10, 06123 Perugia, Italy, and Molecular Discovery Limited, 215 Marsh Road, Pinner, Middlesex, London HA5 5NE, United Kingdom

Received January 18, 2007

One of the most important physicochemical properties of a molecule is pK_a. It is known that two parameters imperative in ADME profiling, solubility, and lipophilicity are governed by pK_a, and receptor binding can be influenced by pK_a. Because most drugs are ionized in physiological conditions, pK_a is particularly relevant to medicinal chemistry. Despite the numerous advances in high-throughput measurements, in silico determination is still the fastest and cheapest way of obtaining pK_a. This paper presents a new original computational method for pK_a prediction of organic compounds. Descriptors were generated using the program GRID, and these descriptors are based on molecular interaction fields precomputed on a set of molecular fragments. The new method was developed, trained, and cross-validated by using a large and diverse data set of 24 617 pK_a values. This paper presents the results for a class of 421 acidic nitrogen compounds (RMSE = 0.41, $r^2 = 0.97$, $q^2 = 0.87$) and for a class of 947 six-membered N-heterocyclic bases (RMSE = 0.60, $r^2 = 0.93$, $q^2 = 0.85$). For external validation 28 novel compounds were selected that covered nine different ionizable groups, and 39 pK_a values could be experimentally determined by spectral gradient analysis (SGA). Comparison of experimental pK_a with calculated pK_a demonstrated that the predictive ability of the method is good (external set, $r^2 = 0.85$, RMSE = 0.90).

INTRODUCTION

The acid dissociation constant (pK_a) measures the strength of an acid and is of general interest in both chemistry and biology. It has been suggested that 95% of all drugs are ionizable.[1] The degree of ionization controls lipophilicity and solubility, two properties widely used in pharmaceutical research to predict absorption and distribution of a compound. Therefore, the pK_a is one of the most important physicochemical properties of a molecule.

The ionization state is a key parameter in ADME profiling, but the effect of a charge on the biological behavior of a molecule can also elicit other types of effects. Ionizable groups affect the ability of a molecule to interact with a target. When the latter is a metabolic enzyme, pK_a can be important in determining the rate and the site of metabolism.[2] In drug formulation the ionization constant is important in choosing the correct excipient and counterion. In addition, pK_a is often a relevant descriptor in QSAR models.[3] All these factors explain why there is a growing interest in the development of better pK_a prediction methods.

Numerous authors have addressed the importance of computational tools in pK_a calculation so that the risk of late-stage attrition in drug discovery is reduced.[4,5] There have been several attempts at predicting pK_a. Theoretical studies based on ab initio could provide a general and accurate way of calculating pK_a by accounting for the solute−solvent interaction, but the latter increases the complexity of the problem considerably. Therefore, because of their great computational cost, ab initio,[6−8] density functional theory, and semiempirical[9,10] methods have not proved to be feasible for virtual screening applications.

Comparative molecular field analysis (CoMFA) has been used to model pK_a values for small sets of structures drawn from specific chemical series,[11−13] but this application is conformation and alignment dependent. In the past methods based on linear free energy relationships have been thoroughly investigated using the Hammett-Taft equations,[14] an approach applied extensively in commercial programs.[15]

In the last 10 years new QSPR pK_a prediction methods have been investigated, and the descriptors used include atomic charges,[16,17] topological[18,19] and E-state descriptors,[20] chemical reactivity models[21] (SPARC), atom types,[22,23] and group philicity.[24] A common framework in the most recent methods for pK_a prediction is the use of molecular-tree structured fingerprints to describe the neighborhood of an ionizable center. This approach was introduced by Xing who used Sybyl atom types as descriptors and has received much attention. Furthermore, numerous authors[25−27] have taken the investigation of the use of topological distances to model pK_a further by using not only atom types but also descriptors based on atomic charges.

The authors present a new QSPR pK_a prediction method that describes the molecular structure around an ionizable center using topological distances. In other words, atomic descriptors are used to map the chemical space of a molecule at different topological distances from the ionizable center. Although the pK_a is predicted by the 2D structure, the descriptors are derived from 3D molecular interaction fields (MIFs) generated by the program GRID[28] on a very large set of fragments. Many applications have demonstrated that

* Corresponding author phone: +390755855629; fax: +3907545646; e-mail: gabri@chemiome.chm.unipg.it.
[†] Università degli Studi di Perugia.
[‡] Molecular Discovery Limited.

10.1021/ci700018y CCC: $37.00
Published on Web 10/02/2007

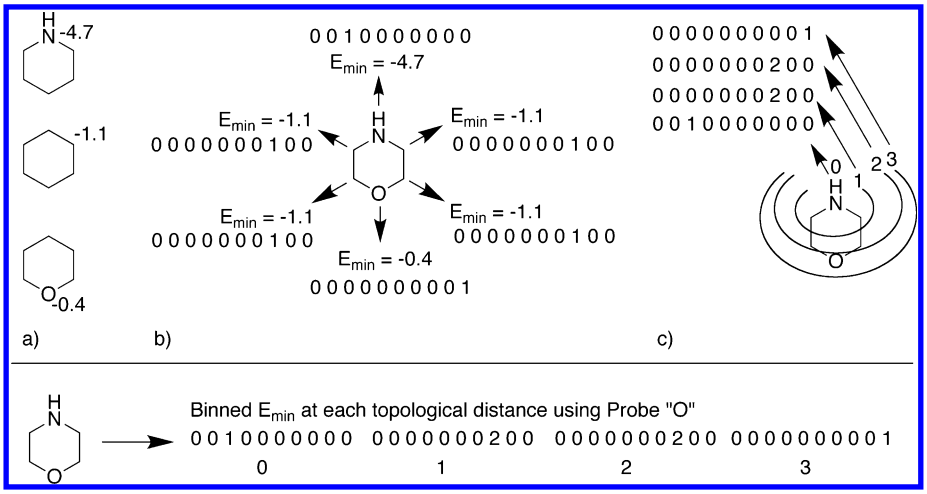


Figure 1. Example of the workflow to generate the descriptors: (a) fragments and energy minima (kcal/mol) calculated for the reference atoms using probe "O"; (b) binned energies used to describe each atom of morpholine; and (c) sum of the bins for the atoms at the same level.

GRID MIFs can be useful in drug discovery and ADME optimization.[29] The GRID force field has been applied first to predicting molecular properties that influence the passive transport across different membranes (Volsurf[30]), second to predicting the site of metabolism (MetaSite[31]), and last to generating fingerprints for ligands and proteins (FLAP[32]).

The data used as the training set include a very large collection of mono- and multiprotic molecules (18 882) for an overall total of 24 617 pK_a values. Twenty-eight external compounds covering nine different ionizable groups (phenols, anilines, amides, pyridines, sulfonamides, imides, carboxylic acids, amines, and oximes) were used to validate the method.

The procedure to calculate pK_a is neither CPU time-intensive nor memory demanding and is feasible for virtual screening applications. Moreover, it may also be relatively easy to extend it in the future to larger molecules of biological interest.

METHODS

The acid dissociation constant measures the acidity of a specific site in a molecule. In general, the development of a pK_a prediction model based on a QSPR approach can be summarized as follows: (1) accurately collecting a large amount of diverse experimental data; (2) identifying descriptors that explain how a change in the molecular structure affects pK_a; and (3) finding a relationship between the experimental pK_a and the selected descriptors using appropriate statistical models.

Several approaches to pK_a prediction are successful when developed over small series, but they may fail when applied to large and chemically different data sets. To overcome this limitation the method presented in this paper was applied over a large and very diverse data set of ionizable sites. Figure 1 summarizes the procedure, which is composed of the following steps: (1) building a large database of fragments and computing energy minima (MIFs) for each fragment by using the GRID force field [Figure 1a shows a small sample of the fragments used.]; (2) describing every atom of the molecule to be analyzed by using precomputed MIFs from the database of fragments [Figure 1b illustrates this step for morpholine and shows that MIFs were binned.]; (3) describing the molecular structure around each ionizable site by using binned MIFs summed at each topological distance (Figure 1c); and (4) building statistical tables containing molecular descriptors and experimental pK_a values for different classes of ionizable sites.

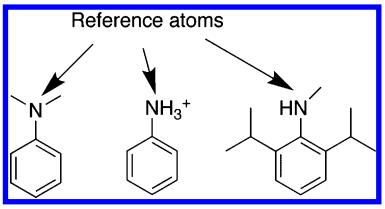


Figure 2. Three of the fragments used. The nitrogens are the "reference atoms". The carbons are the "surroundings".

It is important to stress that ionization constants are not global properties because they depend on the molecular structure around the site that undergoes ionization. Therefore, the descriptors selected accounted for the environment around the ionizable site. Molecular structures were described using topological distances, that is, the properties of atoms positioned at increasing distances (in terms of bond numbers) from the site of ionization were described, and GRID MIFs were used to obtain these properties.

MIFs require the 3D structure of a molecule, and from the perspective of an application in virtual screening this can be a disadvantage because generating the 3D structure and computing the MIFs can be a relatively time-consuming task. This limitation was solved using MIFs precomputed on molecular fragments, so pK_a was calculated by only using a 2D connectivity matrix. However, this approach implies that for each atom of the molecule the MIFs precomputed from appropriate fragments need to be retrieved.

Fragments Database. We selected a set of 466 small and semirigid molecular fragments and then generated and minimized a single conformer of each using internal software. The smallest fragment used was composed of two atoms, while the largest was composed of 30 atoms (excluding hydrogens). We identified in each fragment a center of interest that will be called the "reference atom", composed of either one or two/three atoms, and the chemical surroundings, composed of atoms in proximity to the center of interest (Figure 2).

The reference atom could be an ionizable group in a neutral, positively, or negatively charged state or a nonion-

izable group. Therefore, we selected ionizable and nonionizable fragments. Fragments containing a nonionizable reference atom were necessary to describe the effects of nonionizable substituents, such as an alkoxy or a nitro group. Each ionizable fragment had its charged counterpart.

Charged Fragments. Fragments containing charged groups (such as a protonated pyridine or a carboxylate) were necessary to treat multiprotic systems because the molecule had to be represented in the ionization state at pH = pK_a of the ionizable group of interest. It was important to identify the correct ionization state of the molecule because the neutral form and the charged form of the same group exert a different effect on the center of reaction. This is exemplified by the values of Hammett σ_m for NH_2 (−0.16[33]), NH_3^+ (0.86[33]), COOH (0.37[33]), and COO^- (−0.10[33]). Therefore, to predict the pK_a of the amino group of glycine (pK_{a1} = 2.35, pK_{a2} = 9.78), the carboxyl was set in the negatively charged form because when the amino group titrates, the carboxyl is fully deprotonated.

Albeit most multiprotic compounds have one predominant ionization route, such as glycine (cation ⇆ zwitterion ⇆ anion), some molecules show multiple ionization patterns. For example, for m-aminobenzoic acid, $[^+H_3NC_6H_4COO^-]/[H_2NC_6H_4COOH] = 1.5$[34] in aqueous solution at 25°. The loss of a proton to give $H_2NC_6H_4COO^-$ can be attributed to two distinct species, one bearing the NH_3^+ group, the other the COOH group, so that two ionization routes occur in solution (cation ⇆ zwitterion ⇆ anion and cation ⇆ uncharged ⇆ anion). To handle such a complex system by using the method presented in this paper, pK_a values obtained by pH-metric titration (macroconstants) are inadequate because they do not distinguish between single species; therefore, we used microconstants,[35] which refer to the ionization of the single species. However, because of the limited availability of literature data for microconstants, for most compounds only the main ionization route was considered along with the experimental pK_a (macroconstant).

Recognition of Fragments. To recognize the fragments contained in a molecule among those selected, a computational procedure was developed using the computer programming language C. The algorithm that recognizes fragments was not based on an exact substructural search. In fact, the aim was to attribute each atom of a molecule to the fragment whose reference atom (1) was composed of the same chemical element as the query atom and (2) had a microenvironment most similar to that of the chemical surroundings of the query atom.

If the MIFs calculated for a fragment produced the same binning of an existing fragment, there was no need to use the new fragment. This concept is exemplified by the fragments "piperidine" and "morpholine", for which very similar MIFs were calculated that fell into the same bins. When the molecule of interest contained a morpholine, as shown in Figure 1, descriptors calculated for piperidine were used to describe the nitrogen atom. Therefore, the fragment "morpholine" was not necessary, and we did not include it in the database of known fragments. This procedure was useful to reduce the number of fragments that, otherwise, would theoretically have been infinite.

Computation of MIFs. The GRID force field was used to generate descriptors for each fragment (Figure 1a). The GRID program calculates interaction energies between a target molecule and a "probe". MIFs are obtained for each point distributed throughout and around the target molecule (the fragment, in this method) by summing the following different types of contributes: electrostatic effects, hydrogen bonds, and van der Waals interactions.

Table 1. Ten Probes Used for the Computation of the MIFs

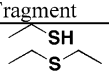
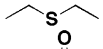
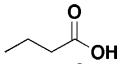
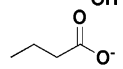
probes	charge	description
OH2	neutral	water
N3+	+	cationic sp^3 nitrogen bonded to three hydrogens
N1:	neutral	sp^3 nitrogen with one lone pair bonded to one hydrogen
N1=	+	cationic sp^2 N with one hydrogen, one single, and one double bond
N:	neutral	sp^3 nitrogen with one lone pair bonded to three non-hydrogen atoms
N-:	-	anionic sp^2 nitrogen with one lone pair, one non-hydrogen single, one double bond
O	neutral	sp^2 carbonyl oxygen bonded to one atom and with two lone pairs
O::	-	explicit resonating sp^2 oxygen with two lone pairs (carboxylate)
CL-	-	chloride anion
LI+	+	lithium cation

The MIFs were used to describe the properties of the "reference atom" and not the properties of the entire fragment. For each fragment, GRID first flagged all points whose computed energy had mainly to be attributed to the reference atom; second, from the energies computed in these points, the one with the minimum value was selected and used as descriptor.

Since pK_a depends on the environment around the site of ionization, a QSPR method needs to characterize such an environment. GRID energy minima are useful to describe the properties of an atom in different chemical surroundings according to the probe used. To take full advantage of GRID MIFs, energy minima were binned by adopting an equal-frequency scheme. The discretization method used generates intervals containing an equal number of observations; therefore, the size of each of the 25 intervals was defined so that approximately the same number of precalculated interaction energies would fall into each interval. By converting continuous descriptors into discrete values, it was possible to achieve superior predictive ability, because different properties of the atoms were categorized according to the values of the energy minima.

MIFs: Probes Used. Ten out of the 77 probes available in the GRID force field were used (Table 1). To account for different types of interactions while keeping the computational time to a minimum, four probes in neutral state, three positively charged, and three negatively charged were selected. Therefore, the resulting MIFs were able to describe different chemical properties of the atoms of a molecule. Values of the energy minima described the strength of the interaction. Usually, very weak energies (approximately between 0.0 and −2.0 kcal/mol) are related to van der Waals forces only. The strongest interaction obtained corresponded to the fragment "methylsulfonyl(phenyl)amide" (reference atom: "N", deprotonated form, probe: "LI+", $E_{min} = -8.9$ kcal/mol). If repulsive interactions were calculated between a fragment and a probe, the corresponding energy was set to zero. Therefore, energy minima ranged from zero to −8.9 kcal/mol.

Table 2. Energy Minima (kcal/mol) Calculated Using Probes "LI+" and "O" [a]

Fragment	LI+	O
1 	-0.4	-1.9
2 	-0.5	-1.1
	-4.1	-3.9
3		
	-5.2	-1.2
4		

[a] The "reference atoms" are indicated in bold (S and O). Fragments with similar chemical properties have similar interaction energies.

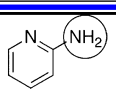
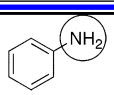
The Information in the MIFs. The MIFs were used to produce ten interaction energies for each atom of the examined molecule (except molecular hydrogens). Depending on the probe used, these energies could quantitatively measure the various effects of the substituents. The "LI+" probe only accounts for electrostatic interactions, described by a Coulombic potential, so it is able to express the distribution of charges in each atom of the molecule. Table 2 shows that probe "LI+", only able to account for electrostatic effects, interacts poorly with the sulfur atom of a thioether ($E_{min} = -0.5$ kcal/mol) and of a thiol ($E_{min} = -0.4$ kcal/mol). The interaction between "LI+" and a carboxyl is much stronger (−4.1 kcal/mol), and it is even stronger for a carboxylate (−5.2 kcal/mol). Because it is a hydrogen bond acceptor, probe "O" is useful in taking into account the effects of hydrogen bonding. It is worth noting that the interaction energy with aliphatic thioethers ($E_{min} = -1.1$ kcal/mol) is approximately the same as the interaction energy with a carboxylate ($E_{min} = -1.2$ kcal/mol) because none of the two fragments is a hydrogen bond donor. MIFs also reflect the fact that compared to carboxyls (E_{min} ("O") = −3.9 kcal/mol), thiols are very weak hydrogen bond donors (E_{min} ("O") = −1.9 kcal/mol).

Atoms having similar chemical surroundings also had similar MIFs. As Table 3 shows, when the environment of a reference atom is only slightly changed (aniline and o-aminopyridine), the interaction energies show a small deviation (less than 0.3 kcal/mol), and the same information (binned MIFs) is produced. Therefore, energies computed for fragment "aniline" could also be used to describe the anilinic nitrogen of o-aminopyridine, and this explains why a finite number of fragments was used. In summary, if the reference atom of an unknown fragment had very similar chemical properties to the reference atom of a known fragment, the MIFs produced were very similar and a replacement was possible.

A further demonstration of the information obtained by using MIFs is given in Figure 3, where energy minima calculated between probe "CL-" and nine fragments correlate with Hammett σ_m constants of the substituents (R) used as reference atoms in these fragments. The correlation found demonstrated that atomic descriptors derived from MIFs could be used as a quantitative measure of the electronic effect of the substituents. However, this type of information alone was insufficient to explain the pK_a for all ionizable centers, and so another nine probes had to be used to fully characterize a molecule.

The Descriptors. QSPR prediction methods rely on the representation of a molecule by means of descriptors that contain information about physicochemical properties or structural characteristics. The binned MIFs were used to describe the physicochemical properties of a molecule at each topological distance from the site of ionization. What is more, they implicitly reflect structural characteristics. The effects of substituents on pK_a are additive, and they depend on the distance from the site of ionization; therefore, the atomic descriptors corresponding to the same topological level could be summed together. Figure 1c shows that the binned MIFs calculated for atoms having the same number of bond numbers from the site of ionization were summed at each topological level. A numerical vector described each level, characterizing the strength of the interactions found by each probe according to the position and the numerical value of the bins. Since 10 probes were used, the overall description

Table 3. Interaction Energies Reported (kcal/mol) Were Computed Only in Points of the Grid Where the Major Contributor Was the Reference Atom (in the Circle)[a]

PROBES			
OH2	-5.3	-5.1	-5.7
N3+	-1.8	-1.7	-5.0
N1:	-6.0	-5.7	-6.6
N1=	-1.7	-1.7	-5.4
N:	-6.3	-6.0	-1.1
N-:	-4.7	-4.6	-1.5
O	-5.4	-5.3	-0.9
O::	-5.9	-5.6	-1.9
CL-	-2.5	-2.4	-2.4
LI+	-0.4	-0.5	-0.9

[a] Therefore, o-aminopyridine and aniline had similar E_{min} ($\Delta E < 0.3$ kcal/mol) because the anilino group was the reference atom for both fragments.

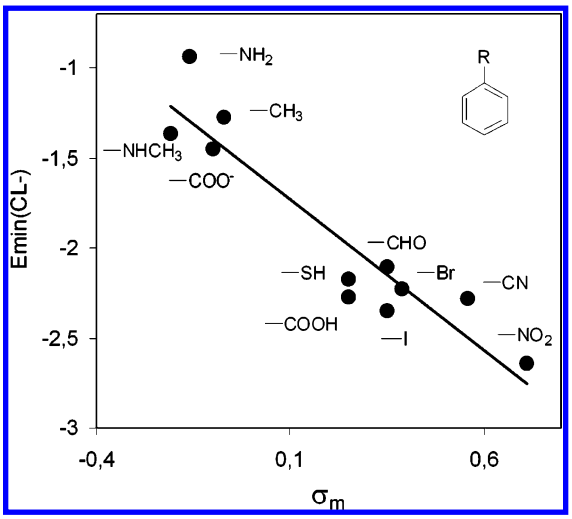


Figure 3. Correlation between Hammett σ_m[33] and E_{min} (kcal/mol, probe "CL-") calculated for the corresponding substituents.

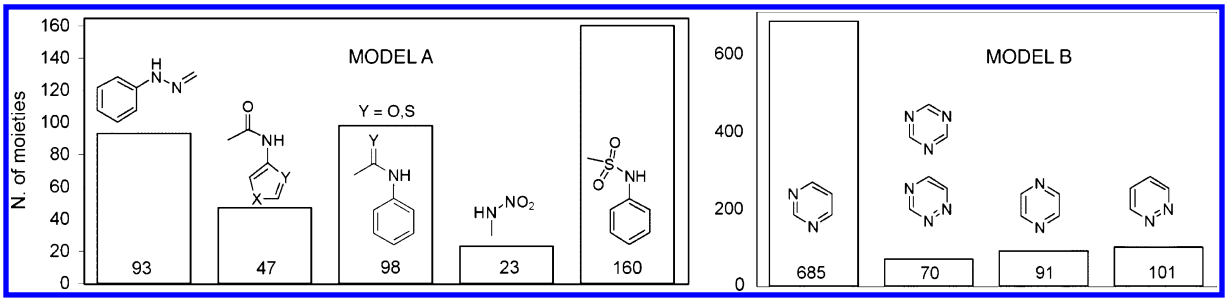


Figure 4. Number of moieties included in the two pK_a prediction models presented. On the left: model A with the 421 acids; benzenic rings can be replaced with heteroaromatic systems. The five-membered ring represents a generic five-membered heteroaromatic system. On the right: model B with the 947 bases.

of an ionizable site was composed of 10 of these vectors. A level was also defined for the starting ionizable site, because most of the pK_a prediction models included different types of ionizable sites, each having different flags (position of the bins) expressed by the MIFs. If at the topological distance n there were m atoms with a different binned MIF, the sum of active (=nonzero) descriptors within the block n was 10 $\times$ m (because 10 probes were used).

The maximum number of topological distances was related to the type of ionizable site, because the transmission of electronic effects is different for aromatic and aliphatic systems. Thus, the number of descriptors ranges from a minimum of 1750 = 25 (binned energy of a probe) $\times$ 10 (number of probes) $\times$ 7 (maximum distance from the ionizable site) up to a maximum of 3250 = 25 (binned energy of a probe) $\times$ 10 (number of probes) $\times$ 13 (maximum distance from the ionizable site).

The descriptors used did not encode information about the entire molecular structure, but they only refer to atoms in proximity to the site of ionization. In addition, cis−trans and R/S relationships and the conformation of the molecule of interest were disregarded because the descriptors were calculated by the connectivity matrix only.

RESULTS AND DISCUSSION

Classes of Ionizable Sites. The ionizable sites were classified into 33 pK_a prediction models. For example, separate classes for phenols and benzoic acids were defined, and all five-membered heterocycles were included in the same class. Because a statistical model requires the same number of descriptors for each object, only ionizable sites with the same number of descriptors could form part of the same model. Some classes included only one of the 166 ionizable sites; other classes had more than one type of starting ionizable site. Hypothetically, it would have been possible to build either a unique global model that included all the different ionizable sites or 166 different models for the 166 different ionizable sites. The selection of a smaller number of classes has the advantage of building more flexible and robust models. Nevertheless, if the interaction effects of different ionizable sites with the rest of the molecule were not of the same type, the resulting model showed a decrease in terms of r^2 and q^2. A larger number of classes was also used, but because fewer objects were included in each model, this resulted in poorer predictive abilities in cross-validation. The selection of 33 pK_a prediction models was based on the best compromise between robustness of the models and accuracy of the predictions.

Data Set. Some of the pK_a data were collected from the literature (including pK_a compilations[36,37]), some were obtained in the laboratories of the authors, and some were measured in external laboratories. Most pK_as were measured in water at 25 °C, but some were determined in a range of temperatures between 15 and 30 °C. This does not represent a source of noise as changing the temperature up to 10 degrees only produces very small shifts[14] (as low as 0.1 pK_a units) in the ionization constant. The aqueous pK_a values for a limited number of compounds were obtained by extrapolation from measurements obtained in cosolvent.

The results for two of the 33 pK_a prediction models are presented. The first model (A) is composed of a set of 421 pK_as of compounds bearing an acidic nitrogen; the second model (B) includes 947 pK_as and is composed of six-membered basic heteroaromatic systems having two or three nitrogen atoms in the ring. Pyridines were treated in a different model. The statistical distribution of the moieties used to build the two models is presented in Figure 4.

Training and Internal Validation. The GOLPE[38] program was run in order to build the 33 pK_a prediction models using Partial Least-Square (PLS[39]). The matrix of descriptors was autoscaled, and variables having zero variance were automatically removed. Consequently, 732 active variables were used for model "A", and 1123 active variables were used for model "B". The predictive power of the pK_a models was evaluated using both cross-validation and external data set validation. The cross-validated q^2 was obtained by randomly dividing the data set into five groups, with the model then predicting each group by deriving the prediction from compounds in the other four groups and averaging across all 100 trials. The initial partitioning was repeated 20 times. A critical part of a PLS model is the optimal number of components (NC). When NC increases, RMSE always decreases. Nevertheless, models with too many components might be overfitted, resulting in poorer predictive ability in cross-validation. The optimal number of components was selected by choosing the model with the highest q^2. Predicted versus experimental pK_a values for the two models presented are plotted in Figure 5. For model A (acidic nitrogen compounds):

$$R^2 = 0.97,\ \mathrm{RMSE} = 0.41,\ \mathrm{NC} = 8,$$
$$Q^2 = 0.87,\ \mathrm{RMSE}_{\mathrm{PRED}} = 0.91,\ N = 421$$

And for model B (six-membered heteroaromatics):

$$R^2 = 0.93,\ \mathrm{RMSE} = 0.60,\ \mathrm{NC} = 10,$$
$$Q^2 = 0.85,\ \mathrm{RMSE}_{\mathrm{PRED}} = 0.86,\ N = 947$$

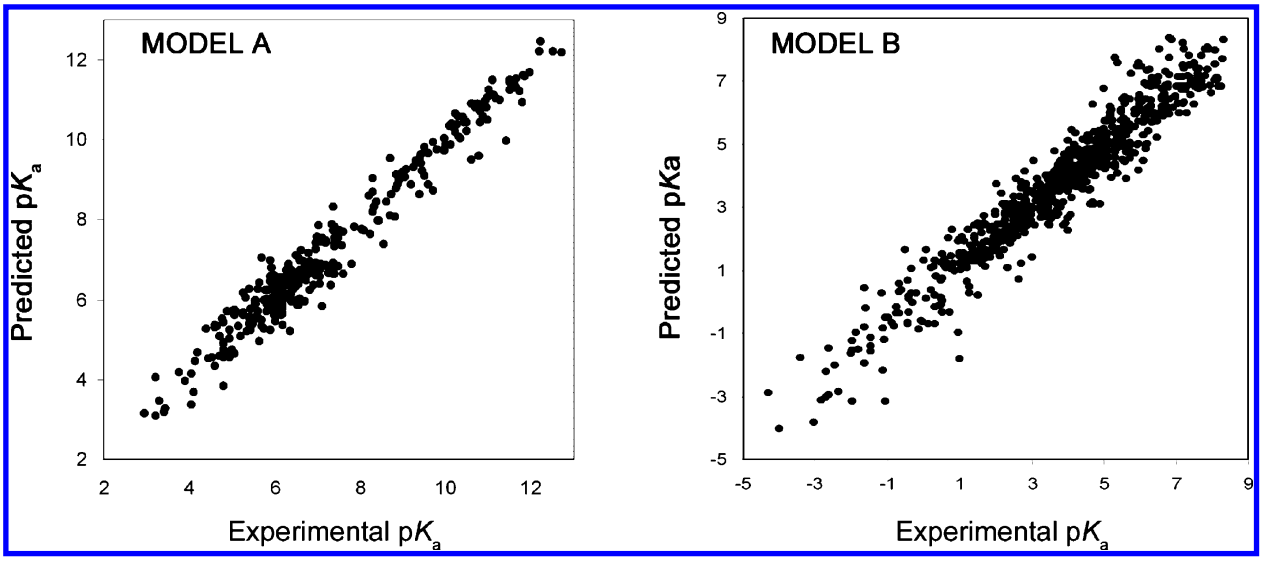


Figure 5. Predicted vs experimental pK_as for model A (421 acidic nitrogen compounds) and for model B (947 six-membered heterocyclics) used in the training set.

For both models, the best results from cross-validation were obtained by using a large number of components, and the absence of overfitting was double-checked by scrambling of the pK_a values over the matrix of descriptors. For the model with the acids, the first component yielded $r^2 = 0.1$ and $q^2 = -0.02$, whereas for the model with the bases $r^2 = 0.01$ and $q^2 = -0.06$. The highest error was found in model B, being for pyrimidine-2,5-diamine (exp pK_a: 1.0,[40] pred pK_a: −1.81), a multiprotic compound with two pK_as; the first protonation of the heterocyclic nitrogen was predicted with sufficient accuracy (experimental pK_a: 4.0,[40] predicted pK_a: 4.13). It is worth noting that the experimental pK_a of 1.0 in the original paper was reported as being approximate.[40] Most experimental techniques are not able to determine extreme pK_a (below 2 and above 12), and the pK_a values outside this range are not usually very accurate. Therefore, models cannot be expected to display the same level of accuracy throughout the whole range of pK_a, and the lack of experimental accuracy for extreme pK_a is reflected in the pK_a calculated at extreme values.

External Validation of Method. To test the validity of the new method, 28 novel and chemically diverse compounds bearing moieties of pharmaceutical interest (Figure 6) were purchased from SPECS[41] screening collection, and their pK_a values were determined using Spectral Gradient Analysis[42] (SGA) from Sirius Analytical Instruments Ltd. (U.K.).

SGA is a high-throughput method used for accurate pK_a determination, widely applied in pharmaceutical research. The largest experimental error is estimated to be around 0.3 pK_a units, but most pK_a values can be determined with an accuracy of 0.1 pK_a units. Because SGA is a UV based method, it is only able to detect a pK_a if a group shows a change of absorbance between the neutral and the charged species; therefore, a chromophore moiety in proximity to the ionizable site is essential. All the compounds selected included at least one aromatic ring; therefore, almost all pK_a in the range between pH 2 and 12 could be determined. Only the pK_a for the amino group of compound 5 and the pK_a for the sulfonamide of compound 15 could not be detected because of this limitation.

Twelve out of the 28 compounds were diprotic, and it was possible to measure an overall total of 39 pK_as. Compound 18 might exist in multiple tautomeric forms (keto−enol), but the species shown is estimated to be by far the most stable because the ortho ester stabilizes the hydroxy form via the intramolecular hydrogen bond.[43] The set of 28 compounds included a very diverse collection of chemical groups, and 14 different pK_a prediction models were used to predict their ionization constants. It is remarkable that every compound of the test set bears at least one moiety that is not represented in the training set at the same topological distance from the site of ionization. In fact, external compounds were selected to test the robustness of the method in the presence of new original functionalities. The good agreement found in the comparison with experimental pK_as indicates the new method is valid and performs well ($N = 39$, $r^2 = 0.85$, RMSE = 0.90).

Comparison with ACD/pKa. The pK_a prediction method presented was trained over a very large data set. The advances in high-throughput pK_a measurements made up to the time of writing provide larger and more consistent data. SGA Profiler can run approximately 300 assays per day. Multiplexed capillary electrophoresis is another advantageous high-throughput technique for measuring pK_a.[44] Nevertheless, experiments are much more expensive than in silico predictions, making computational tools to predict pK_a very desirable especially when the compound is either chemically unstable or toxic, its solubility is too low, or it has extreme pK_a values.

The data set used for external validation included novel compounds whose pK_as were not available from the literature. Most compounds contained moieties that might be found in potential drugs, and it was therefore important to predict their pK_a accurately so that the method presented here can be used for drug discovery applications. In drug research it is very important to have robust predictors for novel libraries, so a benchmarking data set only containing new original structures was selected. Table 4 presents the results obtained by ACD/pKa[45] ($r^2 = 0.76$; RMSE = 1.36) and the method presented in this paper ($r^2 = 0.85$; RMSE = 0.90).

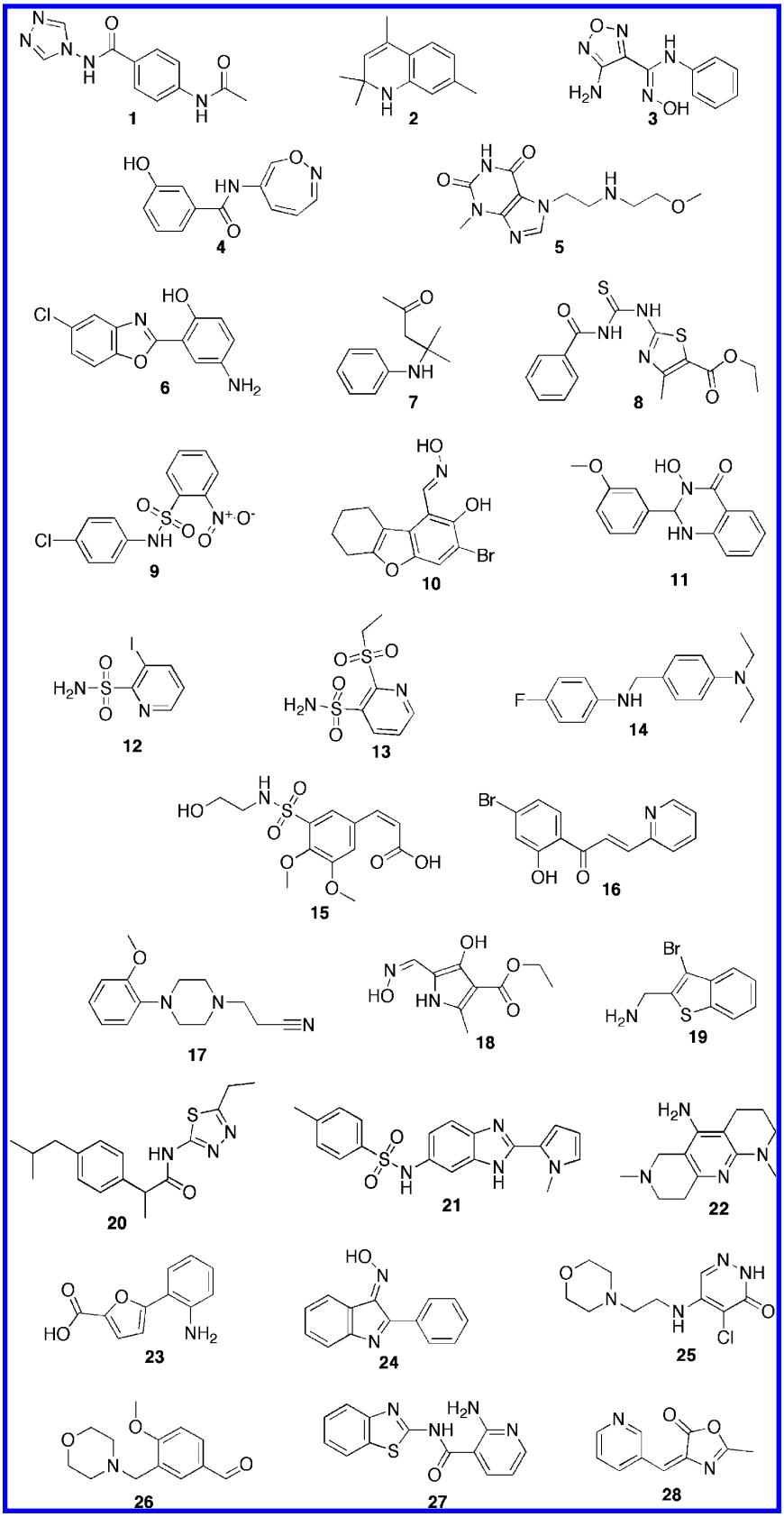


Figure 6. Compounds used as external test set.

It was found that the pK_a values in the benchmarking set used in this paper were particularly difficult and challenging to predict because of the novelty of the compounds used, and because very little knowledge is available in the literature about the pK_a of some of the structural domains explored. The distribution of the residuals plotted in Figure 7 highlights the largest deviations found, which can be explained by the fact that the pK_a of molecules with the same substructure is not available in the literature. Therefore, the lack of information about the effect of unexplored domains is still an

Table 4. Experimental and Predicted pK_a Values for Compounds in Figure 6[a]

compd	CAS	exp pK_a	pred pK_a	ACD/ pKa[46]	ionizable site	acid (A), base (B)
1	333431-72-4	6.29	6.60	10.75	amide (triazole)	A
2	1810-62-4	4.14	5.07	4.96	aniline	B
3	154557-01-4	10.10	9.70	10.15	oxime	A
4	663197-21-5	8.81	8.92	9.23	phenol	A
		11.20	11.04	12.77	amide	A
5	354795-15-6	9.62	8.10	9.85	imide	A
6	293737-93-6	4.64	5.4	4.01	aniline	B
		9.81	9.62	8.17	phenol	A
7	88187-84-2	4.62	5.95	6.15	aniline	B
8	83584-24-1	7.80	6.91	9.04	imide	A
9	63228-67-1	6.55	7.63	7.19	sulfonamide	A
10	342391-55-3	8.33	8.13	6.66	phenol	A
		11.46	10.44	11.11	oxime	A
11	438235-59-7	8.07	9.39	8.67	hydroxyl	A
12	94527-47-6	8.96	8.62	8.82	sulfonamide	A
13	56825-37-7	8.51	9.54	8.66	sulfonamide	A
14	723753-72-8	4.66	3.30	3.37	aniline (II)	B
		6.51	6.88	6.29	aniline (III)	B
15	326882-27-3	3.77	4.14	3.71	carboxylic acid	A
16	5250-15-7	3.58	3.45	2.64	pyridine	B
		8.44	7.02	6.55	phenol	A
17	21103-25-3	5.50	5.58	6.55	aniline	B
18	172753-18-3	7.85	7.56	9.93	hydroxyl	A
		9.74	10.55	11.82	oxime	A
19	337469-92-8	7.94	6.87	8.95	amine	B
20	666817-82-9	8.94	7.32	8.86	amide	A
21	700849-00-9	4.67	5.18	5.63	imidazole	B
		8.75	9.12	8.46	sulfonamide	A
22	297757-67-6	6.54	6.62		amine (III)	B
		9.51	11.1	8.35	pyridine	B
23	65172-75-0	2.44	3.97	2.84	carboxylic acid	A
		3.93	2.98	3.79	aniline	B
24	4676-99-7	7.69	8.58	10.03	oxime	A
25	691867-24-0	6.29	6.54	7.16	amine	B
		10.50	10.09		amide	A
26	128501-81-5	6.77	5.54	6.72	amine	B
27	723743-86-0	5.00	4.23	4.54	pyridine	B
		7.87	9.04	10.39	amide	A
28	76315-20-3	4.66	3.81	3.13	pyridine	B

[a] Data obtained by ACD/pKa Software (v. 8.03) are also reported.

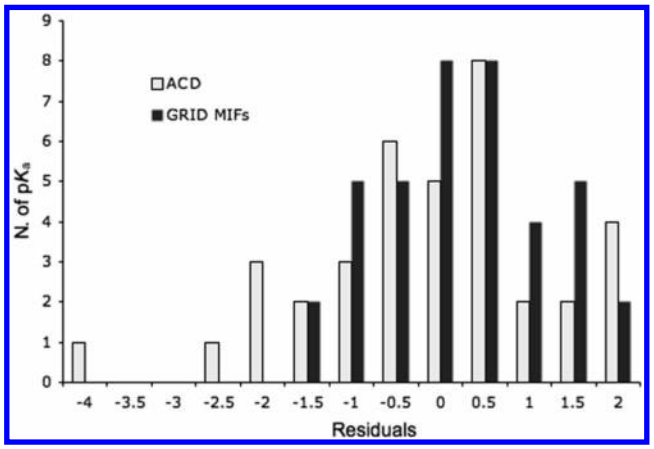


Figure 7. Distribution of residuals for the compounds used as external set.

important matter in pK_a prediction and at the time of writing the authors are working to obtain better results for novel structures.

CONCLUDING REMARKS

This paper introduces a new approach to predicting pK_a of single or multiple ionizable organic compounds. The idea underlying this method was to describe atoms around the site of ionization with GRID precalculating energy minima from a large database of fragments. The energies were binned so that the description of the molecular structure was improved. This approach is suitable for computing pK_a quickly because it only needs to take the connectivity matrix of a molecule into account. Thirty-three pK_a prediction models were developed, each one composed of different types of ionizable sites. Accurate pK_a predictions were obtained, and we presented the results for two of the 33 models, one composed of a set of acids (acidic nitrogen compounds, RMSE = 0.41), and the other being a set of bases (six-membered heterocyclics, RMSE = 0.60). The overall training set is composed of a large and diverse collection of 24 617 pK_a values. To demonstrate the suitability of the method, 28 novel compounds bearing moieties of pharmaceutical interest were selected, and their pK_a were experimentally measured using Sirius SGA. The comparison of experimental data with predicted data confirmed the validity of the method, yielding an external correlation coefficient of 0.85 and an RMSE of 0.90.

Experimental. As indicated by the supplier, compound purity was over 95%. Sirius SGA (spectral gradient analysis) was used to measure the pK_a. Using a flowing pH gradient eliminates the need for pH measurement, which is the rate-limiting step in conventional titration. During the gradient

the pH varies linearly and can be predicted from the time elapsed since the start of gradient generation by calibration using a set of standards. A UV spectrophotometer enables pK_a values to be determined from changes in absorption as a function of pH. At the start of the experimental run, five blank titrations were carried out in which water was injected directly into the gradient. The gradient was then calibrated by injecting solutions of standards with known pK_a values and strong UV absorbance. Ten standards were used in this study, five acids (benzoic acid, phenol, 4-nitrophenol, KHP, 4-chlorophenol) and five bases (4-chloroaniline, 2-methoxybenzylamine, 2-dimethylaminopyridine, 8-methlyquinoline, and 1-phenylpiperazine). The concentration of sample in the gradient was 3.7×10^{-5} M, and the gradient contained 0.4%/vol DMSO and 10%/vol methanol. The compounds were titrated at 25 °C from acidic to basic buffer and from basic to acidic buffer. Ionic strength was adjusted at 0.15 M using a mixture of buffer systems. SGARefine v. 3.1.0.3 software (Sirius) was used to calculate the results.

We gratefully acknowledge Pierre Alain Carrupt (Université de Genève), Manfred Kansy, and Bjoern Wagner (Roche, Basel) for providing experimental pK_a of compounds used for training and external validation.

Supporting Information Available: Matrix of descriptors and experimental pK_a values for models A and B. This material is available free of charge from the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Wells, J. I. *Pharmaceutical Preformulation*; Ellis Horwood Ltd.: London, 1998; p 25.

(2) Upthagrove, A. L.; Nelson, W. L. Importance of Amine pK_a and Distribution Coefficient in the Metabolism of Fluorinated Propranolol Derivatives. Preparation, Identification of Metabolite Regioisomers, and Metabolism by CYP2D6. *Drug Metab. Dispos*. **2001**, *29*, 1377−1388.

(3) Wan, H.; Ulander, J. High-throughput pK_a screening and prediction amenable for ADME profiling. *Expert Opin. Drug. Metab. Toxicol.* **2006**, *2*, 139−155.

(4) van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.

(5) Livingstone, D. J. Theoretical Property Predictions. *Curr. Top. Med. Chem*. **2003**, *3*, 1171−1192.

(6) Silva, C. O.; Silva, E. C.; Nascimento, M. A. C. Ab Initio Calculations of Absolute pK_a Values in Aqueous Solution I. Carboxylic Acids. *J. Phys. Chem. A* **1999**, *103*, 11194−11199.

(7) Silva, E. F.; Svendsen, H. F. Prediction of the pK_a Values of Amines Using ab Initio Methods and Free-Energy Perturbations. *Ind. Eng. Chem. Res*. **2003**, *42*, 4414−4421.

(8) Schuurmann, 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.

(9) Otha, K. Prediction of pK_a Values of Alkylphosphonic Acids. *Bull. Chem. Soc. Jpn*. **1992**, *65*, 2543−2545.

(10) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W .R.; Gancia, E.; Manallack D. T. Estimation of pK_a Using Semiempirical Molecular Orbital Methods. Part 2: Application to Amines, Anilines and Various Nitrogen Containing Heterocyclic Compounds. *Quant. Struct.-Act. Relat.* **2002**, *21*, 473−485.

(11) Kim, K. H.; Martin, Y. C. Substituent Effects from 3D Structures Using Comparative Molecular Field Analysis. 1. Electronic Effects of Substituted Benzoic Acids. *J. Org. Chem*. **1991**, *56*, 2723−2729.

(12) Kim, K. H.; Martin, Y. C. Direct Prediction of Dissociation Constants (pK_a's) of Clonidine-like Imidazolines, 2-Substituted Imidazoles, and 1-Methyl-2-Substituted Imidazoles from 3D Structures Using a Comparative Molecular Field Analysis (CoMFA) Approach. *J. Med. Chem*. **1991**, *34*, 2056−2060.

(13) Gargallo, R.; Sotriffer, C. A.; Liedl, K. R.; Rode, B. M. Application of Multivariate Data Analysis Methods to Comparative Molecular Field Analysis (CoMFA) data: Proton Affinities and pK_a Prediction for Nucleic Acids Components. *J. Comput.-Aided Mol. Des*. **1999**, *13*, 611−623.

(14) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. *pK$_a$ Prediction for Organic Acids and Bases*; Chapman and Hall: New York, 1981.

(15) http://www.acdlabs.com/products/phys_chem_lab/pka/ (accessed December 2006).

(16) Dixon, S. L.; Jurs, P. C. Estimation of pK_a for organic oxyacids using calculated atomic charges. *J. Comput. Chem*. **1993**, *14*, 1460−1467.

(17) Clark, F. H.; Cahoon, N. M. Ionization constants by curve fitting: Determination of partition and distribution coefficients of acids and bases and their ions. *J. Pharm. Sci*. **1987**, *76*, 611−620.

(18) Hall, L. H.; Kier, L. B. Molecular Connectivity Chi Indices for Database Analysis and Structure-Property Modeling. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Reading, U.K., 1999.

(19) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(20) Hall, L. H.; Kier, L. B. *Molecular Structure Description: The Electrotopological State*; Academic Press: New York, 1999.

(21) Hilal, S. H.; Karickhoffs, 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.

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

(23) 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.

(24) Parthasarathi, R.; Padmanabhan, J.; Elango, M.; Chitra, K.; Subramanian, V.; Chattaraj, P. K. pK_a Prediction Using Group Philicity. *J. Phys. Chem. A* **2006**, *110*, 6540−6544.

(25) Zhang, J.; Kleinöder, T.; Gasteiger, J. Prediction of pK_a Values for Aliphatic Carboxylic Acids and Alcohols with Empirical Atomic Charge Descriptors. *J. Chem. Inf. Model*. **2006**, *46*, 2256−2266.

(26) 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.

(27) Kogej, T.; Muresan, S. Database Mining for pK_a Prediction. *Curr. Drug Discovery Tech.* **2005**, *4*, 221−229.

(28) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(29) *Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction*, 1st ed.; Cruciani, G., Ed.; Wiley-VCH: Weinheim, Germany, 2005.

(30) Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular Fields in Quantitative Structure-Permeation Relationships: the VolSurf Approach. *J. Mol. Struct.: THEOCHEM* **2000**, *503*, 17−30.

(31) Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: Understanding Metabolism in Human Cytochromes from the Perspective of the Chemist. *J. Med. Chem*. **2005**, *48*, 6970−6979.

(32) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model*. **2007**, *47*, 279−294.

(33) Hansch, C.; Leo, A.; Hoekman, D. Hammett sigmas. In *Exploring QSAR*; Americal Chemical Society: Washington, DC, 1995.

(34) Christensen, J. J.; Wrathall, D. P.; Izatt, R. M.; Tolman, D. O. Thermodynamics of proton dissociation in dilute aqueous solution. IX. pK, ΔH°, and ΔS° values for proton ionization from o-, m-, and p-aminobenzoic acids and their methyl esters at 25°. *J. Phys. Chem*. **1967**, *71*, 3001−3006.

(35) Avdeef, A. *Absorption and Drug Development*; Wiley-Interscience: Hoboken, NJ, U.S.A., 2003; pp 33−35.

(36) *Dissociation Constants of Organic Bases in Aqueous Solutions*; Perrin, D. D., Ed.; International Union of Pure and Applied Chemistry, Page Bros. Ltd.: Norwich, 1965.

(37) *Dissociation Constants of Organic Acids in Aqueous Solutions*; Kortum, G., Vogel, W., Andrussow, K. D., Eds.; International Union of Pure and Applied Chemistry, Butterworths: London, 1961.

(38) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9−20.

(39) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J. The collinearity problem in linear regression-the partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Statist. Comput*. **1984**, *5*, 735−743.

(40) Brown, D. J.; Harper, J. S. The Dimroth rearrangement. Part I. Some alkylated 2-iminopyrimidines. *J. Chem. Soc*. **1963**, 1276−1284.

(41) http://www.specs.net (accessed Oct 2006).

(42) Box, K.; Bevan, C.; Comer, J.; Hill, A.; Allen, R.; Reynolds, D. High-Throughput Measurement of pK_a Values in a Mixed-Buffer Linear pH Gradient System. *Anal. Chem.* **2003**, *75*, 883−892.
(43) Friedrichsen, W.; Traulsen, T.; Elguero, J; Katritzky, A. R. Tautomerism of Heterocycles: Five-Membered Rings with One Heteroatom. In *Advances in Heterocyclic Chemistry*; 2000; Vol. 76, p 120.
(44) Zhou, C.; Jin, Y; Kenseth, J. R.; Stella, M; Wehmeyer, K. R.; Heineman, W. R. Rapid pK_a estimation using vacuum-assisted multiplexed capillary electrophoresis (VAMCE) with ultraviolet detection. *J. Pharm. Sci.* **2005**, *94*, 576−89.
(45) *ACDpKa, v. 8.03*; Advanced Chemistry Development: Toronto, Canada, 2007.
(46) The predicted values of pK_a were obtained using the ACD/I-Lab Web service (ACD/pKa 8.03).

CI700018Y