Estimation of pK_a for Druglike Compounds Using Semiempirical and Information-Based Descriptors

Stephen Jelfs,* Peter Ertl, and Paul Selzer

Novartis Institutes for BioMedical Research, Basel, Switzerland

Received July 10, 2006

A pragmatic approach has been developed for the estimation of aqueous ionization constants (pK_a) for druglike compounds. The method involves an algorithm that assigns ionization constants in a stepwise manner to the acidic and basic groups present in a compound. Predictions are made for each ionizable group using models derived from semiempirical quantum chemical properties and information-based descriptors. Semiempirical properties include the partial charge and electrophilic superdelocalizabilty of the atom(s) undergoing protonation or deprotonation. Importantly, the latter property has been extended to allow predictions to be made for multiprotic compounds, overcoming limitations of a previous approach described by Tehan et al. The information-based descriptions include molecular-tree structured fingerprints, based on the methodology outlined by Xing et al., with the addition of 2D substructure flags indicating the presence of other important structural features. These two classes of descriptor were found to complement one another particularly well, resulting in predictive models for a range of functional groups (including alcohols, amidines, amines, anilines, carboxylic acids, guanidines, imidazoles, imines, phenols, pyridines, and pyrimidines). A combined RMSE of 0.48 and 0.81 was obtained for the training set and an external test set compounds, respectively. The predictive models were based on compounds selected from the commercially available BioLoom database. The resultant speed and accuracy of the approach has also enabled the development of Web application on the Novartis intranet for pK_a prediction.

1. INTRODUCTION

Since the majority of known drugs are ionizable at physiological pH-levels[1−6] (ca. 1−8), a knowledge of the ionization constants of compounds is particularly important in the drug discovery process. These constants can have a profound effect on the physicochemical properties of a compound and are therefore essential for the optimization of *absorption, distribution, metabolism, and excretion* (ADME) characteristics. Notably, compounds in their un-ionized form are less soluble but can more easily penetrate lipophilic barriers encountered on the way to a biological target. Knowledge of the ionization state of a compound is also required for determining the correct binding-site interactions that occur and the development of reliable *structure−activity* relationships (SAR). Furthermore, ionization constants allow the enumeration of likely chemical species (i.e., that are present at ca. pH 7) prior to protein−ligand docking studies.

Ionization constants, typically represented as pK_a, provide an insight into the degree of dissociation of hydrogen ions from a compound at a given pH. These acidity constants can be derived for both the acidic groups (HA) and the conjugate-acid of basic groups (BH^+) in a compound

$$pK_a = pH + \log \frac{[HA]}{[A^-]} \quad pK_a = pH + \log \frac{[BH^+]}{[B]}$$

where low values indicate the presence of strongly acidic or weakly basic groups and high values indicate the presence of weakly acidic or strongly basic groups. pK_a equals the pH at which a drug is 50% ionized and 50% un-ionized. Experimental methods for deriving these constants involve exposing a compound to an environment of changing pH and monitoring changes that occur to a property dependent on the ionization state of the compound. Modern automated methods based on UV absorption are now available for the *high-throughput* (HT) measurement of pK_a.[7] The development of HT-pK_a methods is particularly important for drug discovery, where there is a need to profile a large number of compounds using only a small amount of sample. Despite the reported accuracy of these techniques, however, ionization constants are often missed for groups not in close proximity to a UV-chromophore. Measurements can also be made using alternative, traditional titration experiments with glass pH electrodes. However, the physical size and sensitivity of this apparatus, and the time required for the electrode to stabilize, often limits their wider application in drug discovery.

In silico methods for pK_a estimation are desirable, aiding the design of experiments and providing predictions for missing ionization constants and 'virtual' compounds that have not yet been synthesized. Unfortunately, pK_a values

* Corresponding author phone: +41 61 6961748; fax: +41 61 6967416; e-mail: stephen.jelfs@novartis.com.

10.1021/ci600285n CCC: $37.00
Published on Web 01/13/2007

remain one of the most challenging physicochemical properties to predict, with tautomerism, charge transfer in conjugated systems, and multiple ionization centers having a complex affect on the ionization of a particular group. Nonetheless, reasonable predictions can be made using methods such as *linear free energy relationships* (LFER) based on Hammett and Taft equations, with one of the most popular commercial tools, ACD/pK_a,[8] using this approach. However, LFER are typically derived from congeneric series of simple organic molecules, and their applicability to druglike compounds is often limited. Ab initio simulations have also been used to predict pK_a,[9] but the computational complexity of these approaches often limits their applicability to relatively small compounds. Furthermore, the representation of the solvent is particularly problematic. Properties derived using other quantum chemical and semiempirical methods have also shown good correlation to pK_a,[1,10] with the latter allowing predictions to be made in a reasonable amount of time. However, for many of the published methods, results were limited to a series of monoprotic structures, and, indeed, the relationships were often found to be unsuitable for charged compounds, i.e. when a more basic or acidic group is present in addition to the ionizable group of interest. Molecular tree structured fingerprints,[3] similar to *hierarchically ordered spherical description of environment* (HOSE) codes,[11] have successfully been used to predict ionization constants for more diverse sets of compounds. These information-based descriptors are particularly good at exploiting large amounts of experimental data to model complex affects on pK_a. However, complex definitions of these descriptors were required to model some conjugated systems, for which the electronic (mesomeric and inductive) effects of substituents are dependent on their point of attachment. This was illustrated by improvements observed when separate molecular tree descriptors were derived for *ortho*-, *meta*-, and *para*-substituents present in anilines, for example.[4]

The approach presented in this paper was developed for the estimation of pK_a of druglike compounds. Importantly, unlike many of the published methods, the focus was to develop a method that could be applied to multiprotic compounds. An algorithm has therefore been developed that applies multiple predictive models in a manner which reproduces the correct ionization order of different groups within such compounds. The predictive models used by the algorithm could essentially be derived using any appropriate set of descriptors. In this study, two classes were selected from the literature. Semiempirical properties including the partial charge and electrophilic superdelocalizability of atoms were used to model electronic effects. Unlike the original publication,[1] however, these properties have been extended to provide models that are applicable to charged compounds. Information-based descriptors including molecular tree structured fingerprints[3] and various 2D substructure flags to indicate the presence of other important structural features were also used. These latter descriptors are particularly important if the vast amounts of data provided by modern HT-pK_a methods are to be fully exploited in the future. The descriptors used in this study were found to complement one another particularly well and, combined with the prediction algorithm, were capable of reproducing the correct ionization order of groups present in the compounds studied. The resultant speed and accuracy of the approach has also led to the development of Web application for the prediction of pK_a, an important addition to the existing suite of cheminformatics tools available on the Novartis intranet.[12]

2. METHODOLOGY

2.1. Software. The CACTVS toolkit[13] was used throughout this project for chemical data manipulation and SMILES[14] parsing. 3D coordinates were generated for structures using the CORINA automated structure generation program.[15] The semiempirical properties were calculated using a modified version of Mopac 6.01[16] (available from Peter Bladon, Interchem Chemical Services, Glasgow).

2.2. Compound Data Sets. The predictive models were trained using compounds selected from the commercially available BioLoom database.[5] BioLoom contains physicochemical data and structures, including over 10K compounds with associated pK_a values, extracted from the literature. The data used in this study included values determined in aqueous solution at temperatures of around 20 °C and excluded salts, solvent mixtures, and approximate values.

Predictive models were developed for a range of ionizable groups including alcohols, amidines, amines, anilines, carboxylic acids, guanidines, imines, phenols, pyridines, and pyrimidines. The current data sets contained compounds with either one or a combination of two or more of these groups. For compounds having multiple ionizable groups, the pK_a values had to be assigned to the correct groups and any stronger acidic or basic groups identified. This procedure was aided by simple models derived using only the semiempirical properties. These preliminary models allowed incorrectly assigned values to be identified and corrected without the risk of overfitting to previously incorrect assignments. Compounds containing multiple ionizable groups of the same functional group type were often excluded unless these groups were identical. This avoided the incorrect assignment of values to the former groups, while, in the latter case, the assignment of the values to the identical groups was arbitrary. However, a statistical factor was subtracted from these observations prior to training. For example, if two identical acid groups are present, then these both have an equal chance of losing a proton, increasing their effective acidity, and therefore decreasing the observed pK_a by log2. The loss of a proton from the second group, in contrast, results in two groups that have an equal chance of being reprotonated, therefore increasing the observed pK_a by log2. For structures that could form multiple tautomers, the tautomer forms that produced the most predictive models were used in each case.

The critical distillation of the experimental data led to the development of a so-called 'star-list' of pK_a values. Each value was associated with the SMILES string of the parent structure in its neutral state, the ionizable atom of interest (undergoing protonation or deprotonation), any atoms acting as stronger bases (protonated prior to the ionization of the atom of interest) or acids (deprotoned prior to the ionization of the atom of interest), and any applicable statistical factors affecting the observed pK_a. The influence of functional groups with similar micro-pK_a constants that may interfere to produce different than expected macro-pK_a (observed) constants[9] was ignored in this current study. This should have little effect on the predictive models since compounds

containing such groups, i.e., with very similar pK_a values that were difficult to assign, were generally excluded.

Finally, the data sets were each divided into a training and external test set of experimental data. A simple *k*-means clustering was performed based on the Euclidean distance between the molecular tree descriptors derived for each ionizable group. The number of clusters produced was set to the number of test set compounds required, with the compounds closest to the cluster centroid being selected in each instance. The final test set contained approximately 20% of the observations for each of the ionizable groups studied.

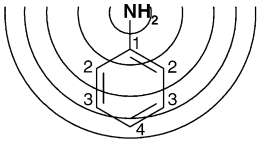
2.3. Semiempirical Properties. The semiempirical properties were generated for each compound in multiple ionization states using AM1.[17] Based on previously published results,[1] both the partial charge and *electrophilic superdelocalizability* (SE) of the ionizable atoms of interest were used in this study. SE is based on frontier electron theory[18] and derived from the eigenvectors $c_{\alpha j}$ and eigenvalues λ_j of the atomic orbitals α and molecular orbitals j of a structure

$$\mathrm{SE}(p) = 2 \sum_{j=1,m} \sum_{\alpha=1,q} (c_{\alpha j}{}^2/\lambda_j)$$

where the sum is calculated over the atomic orbitals for a given atom p and the occupied molecular orbitals. This property is often correlated with the pK_a of monoprotic compounds,[1,2] but, unfortunately, these relationships are not applicable to multiprotic compounds with an additional charged group. To overcome this limitation, each SE value was replaced by a series of values derived from the structure in various ionization states. This series starts with the value derived from the neutral structure followed by relative values derived from various ionized states, where different numbers of the stronger acidic and basic groups are ionized in each case. This allowed the effect of the various ionized groups on the ionization of the group of interest to be modeled effectively. The relative values in the series being the change in SE affected when a given number of either acidic or basic groups were ionized, i.e., starting with one acid or base up to a given maximum of acids or bases, but not combinations of these. Whenever there was a choice of groups to ionize, the groups that produced the largest difference with respect to the neutral structures were selected. Using this series of relative values resulted in far superior models compared to using the absolute values of SE for the groups of interest in the correctly charged species. In contrast, only the partial charge derived for the structure in its correct ionization state was used in the predictive models.

The time required to derive the series of SE properties increases with the number of ionizable groups being compared. It was therefore beneficial to limit these to the most likely groups to have a strong affect, i.e., to groups in close proximity or part of the same conjugated system as the group of interest. This was achieved by ranking the groups in descending order of the minimum number of isolating carbons (along all paths) followed by the minimum path length between these groups and the group of interest. The more time-consuming computation of the affects on SE was then, optionally, limited to only the highest ranking groups.

The performance was also greatly increased by only using the 3D coordinates generated by CORINA, i.e., without any further geometry optimization. Preliminary studies showed



Distance, *d*:	0	1	2	3	4
Atom-type, *t*:	N.pl	C.ar	C.ar	C.ar	C.ar
Frequency (*t*,*d*):	1	1	2	2	1

Figure 1. Molecular tree structured fingerprint generated for an aniline group with one planar nitrogen (N.pl) and six aromatic carbon (C.ar) atoms.

that performing a geometry optimization using AM1 in Mopac only produced a modest improvement in the models, with an increase in RMSE of around 0.01, while the computational time required increased 10-fold. In comparison, using a much faster molecular mechanical minimization resulted in no improvement to the models. A full conformational analysis was not, of course, performed since the increased computational time would have been unfeasible.

2.4. Information-Based Descriptors. The information-based descriptors included molecular-tree structured fingerprints and a small selection of additional 2D substructure flags. The molecular-trees encode the frequency of occurrence of different atom-types at distances moving away from the ionizable atom of interest. In this study, Sybyl atom-types[19] and through-bond distances were used to generate the descriptors. The Sybyl atom-types encode numerous atomic properties that affect the ionization constant of a neighboring atom, including element type, hybridization, and formal charge. Figure 1 shows the atom-type frequencies used in the molecular-tree generated for aniline which consists of aromatic carbon (C.ar) and planar nitrogen (N.pl) atom-types, for example.

Molecular-trees spanning five bonds away from the central atom were found to be adequate for modeling pK_a. These descriptors were generated using both the neutral and also the appropriately ionized structure for each observation. 2D substructural flags were also used to indicate the presence or absence of other important structural features and provided useful corrections for otherwise common outliers, including ionizable groups involved in ring closure or the formation of internal hydrogen bonds. The substructure flags used in each of the models are discussed further in the results section.

2.5. Model Generation. The predictive models were derived using *partial least-squares* (PLS) and validated using 7-fold cross-validation. All of the descriptors were standardized (autoscaled) prior to model generation. For comparison, models were generated using both the semiempirical properties and information-based descriptors independently and in combination. Consecutive latent variables were only added to the PLS models if the resultant cross-validated correlation coefficient increased by greater than 5%.

2.6. Prediction Algorithm. A series of substructural patterns was used to assign each predictive model to the correct ionizable groups within a structure. If reliable predictions are to be made, however, then other groups in the structure that are either more acidic or basic must also be identified and protonated accordingly. A simple algorithm was therefore developed to assign the pK_a values in a stepwise manner to all of the ionizable groups present, i.e.,

starting with the most basic (least acidic) group and ending with the most acidic (least basic) group. At the beginning of each step, any remaining acids were considered to be more acidic (and deprotoned), while previously assigned bases were considered more basic (and protoned).

The determination of the next 'most basic' group at each step was initially done by simply predicting the pK_a for each of the remaining groups. However, this did not always reproduce the correct ionization order for the training compounds, particularly when the pK_a values of the ionizable groups present were similar. A more robust method was therefore developed which involved a further comparison of the top two ranking (most basic) groups. For each of these groups, a further prediction was made but assuming that the other top ranking group, in each case, was more basic (less acidic), i.e. in contrast to the initial prediction. Essentially, this should provide at least one prediction for each group for which the correct ionization state of the compound was used. The highest average pK_a was then used to select the final most basic group, with the pK_a from the initial prediction being assigned to this group as a result.

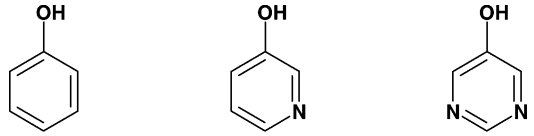
As described, for multiprotic compounds calculations were performed for each structure in multiple ionization states. The performance of the algorithm can therefore be improved if the number of states being compared at each step is reduced. This was achieved using a combination of methods. First, initial predictions were made for the neutral structure, and any groups that appeared to be particularly weak were simply ignored in later comparisons. Other heuristics could also be used to reduce the number of comparisons at each step, including the exclusion of functional groups that are known to always be much weaker than the current group under consideration and also ignoring groups that are in close proximity to a previously assigned group. For example, if two amines are only separated by two bonds, then the second group to be protonated is much less basic and can therefore be ignored in subsequent comparisons. These considerations do not affect the quality of the predictions but were considered to improve the performance of the Web-tool developed during this project.

3. RESULTS

This section summarizes models that were trained on compounds containing various combinations of the ionizable groups mentioned. Although the methodology can be extended to include compounds with very many ionizable groups, these initial results are limited to compounds with only one more acidic or basic group in addition to the ionizable group of interest. For most of the functional groups studied, compounds containing two identical ionizable groups were also included in the training sets. For these compounds, the assignment of pK_a values to each group was arbitrary, but, as previously described, the values had to be adjusted by an appropriate statistical factor prior to training. The compounds used to train each of the models are described in more detail throughout this section.

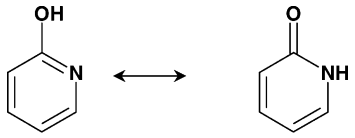
The models generated using either the semiempirical properties (SP) or information-based descriptors (IB) and the combination of these descriptors (SP and IB) are described. These are summarized by the number of compounds (N_{cpds}), number of latent variables (N_{LV}), correlation coefficient (R^2), root mean squared error of estimates (RMSE), and 7-fold cross-validated correlation coefficient (Q^2). Figures 2−8 show the observed and predicted pK_a values for the optimal models derived using the combination of descriptors. These plots highlight the neutral (○), positively charged (red square), and negatively charged (blue square) compounds used in each of the training sets.

3.1. Alcohols Model. The alcohols model was trained using compounds containing aliphatic alcohol (ROH), phenol (1a), 3/5-hydroxypyridine (1b), and 5-hydroxypyrimidine (1c) groups. The majority of compounds were aromatic alcohols which, since the negative charge of the resultant anions is stabilized over the delocalized system, are typically more acidic than aliphatic alcohols.



1a **1b** **1c**

Unsubstituted 2/4/6-hydroxypyridine (1d) and 2/4/6-hydroxypyrimidine groups were excluded from the training set due to the potential formation of pyrid-2/4/6-one (1e) and pyrimid-2/4/6-one tautomers, respectively. These structures tended to be significant outliers, with the hydroxyl group being much less acidic than predicted by the models.

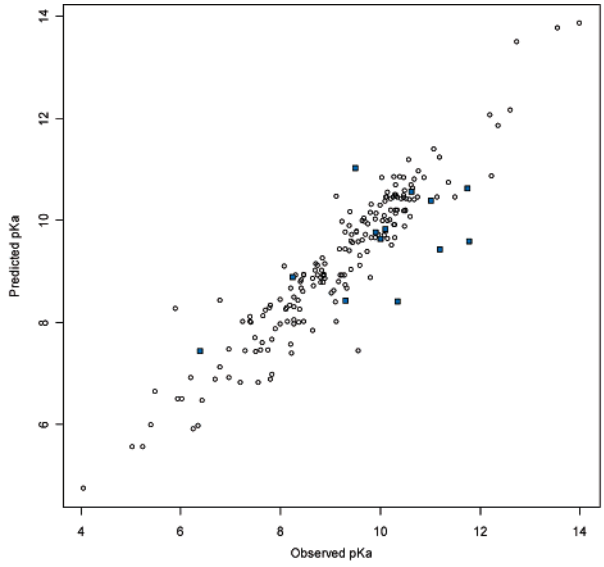


1d **1e**

Compounds containing an additional weakly basic aniline, pyrimidine, and pyridine group were also included in the training set. However, compounds with strongly basic amine groups were excluded since the pK_a of these groups were often difficult to distinguish from the pK_a of the weakly or moderately acidic hydroxyl groups. Compounds containing a stronger carboxylic acid or two identical alcohol groups were also included. The resultant predictive models are summarized in Figure 2.

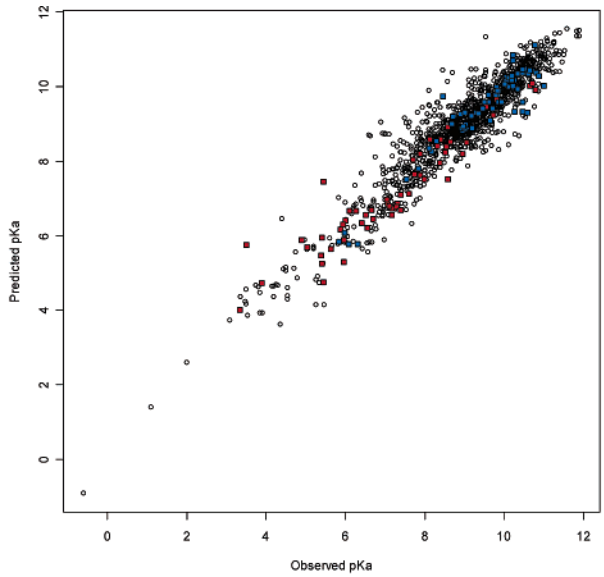
3.2. Amines Model. The amines model was trained using compounds containing primary (RNH_2), secondary (R_2NH), and tertiary (R_3N) amine groups. This is a particularly important model since amine groups commonly occur in druglike compounds and are typically ionized at physiological pH-levels.

Compounds containing an additional weakly basic aniline, amide, pyridine, or pyrimidine group were also included in the training set. Both compounds with strongly acidic carboxylic acid groups and weakly acidic alcohol groups were included, with the former functional groups being appropriately deprotonated prior to training. The resultant predictive models are summarized in Figure 3. The information-based descriptors also included substructural flags indicating the presence of ionizable groups within a primary, secondary or tertiary amine, a three- to six-membered ring, and/or within a morpholine, piperazine, or quinuclidine group.



N_{CPDS}	descriptors	N_{LV}	R^2	RMSE	Q^2
202	SP	3	0.76	0.77	0.74
	IB	2	0.82	0.66	0.74
	SP and IB	2	0.87	0.58	0.80

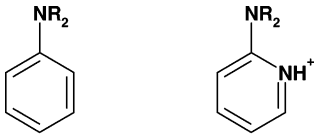
Figure 2. Summary of predictive models derived using the alcohols training set with predicted and observed pK_a shown for the combined (SP and IB) descriptor model.



N_{CPDS}	descriptors	N_{LV}	R^2	RMSE	Q^2
1403	SP	3	0.56	0.99	0.55
	IB	4	0.88	0.52	0.82
	SP and IB	4	0.89	0.49	0.84

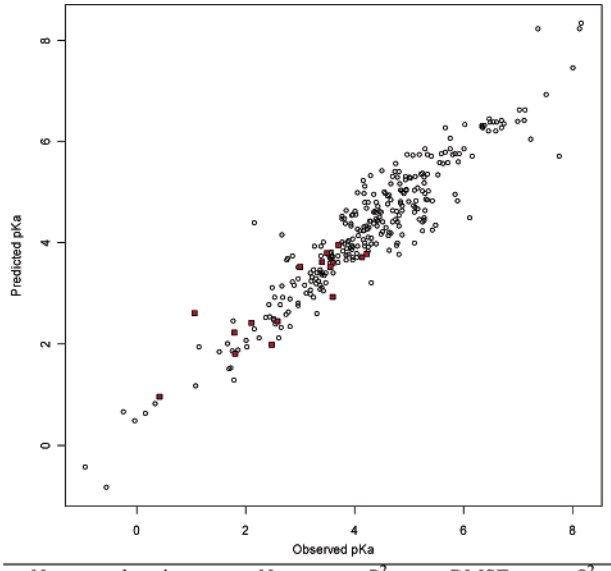
Figure 3. Summary of predictive models derived using the amines training set with predicted and observed pK_a shown for the combined (SP and IB) descriptor model.

3.3. Anilines Model. The anilines model was trained using compounds containing aniline (3a) and aminopyridine (3b) groups. Anilines are usually less basic than amines, with the aromatic stabilization of the group lost on protonation of the planar nitrogen. The planar nitrogen was also assumed to be less basic than the aromatic nitrogen in aminopyridines. The latter nitrogen atom was therefore protonated prior to the training of the anilines model. However, for the weakly basic aminopyridine group there were limited training data available, and, indeed, no data were available for analogous aminopyrimidine groups.



3a **3b**

Compounds containing an additional weakly acidic alcohol group or strongly basic amine group were also included in the training set, with the latter being protonated prior to training. However, compounds with weakly basic pyridine and pyrimidine groups were excluded since the pK_a of these groups was often difficult to distinguish from the pK_a of the weakly or moderately basic aniline groups. The resultant predictive models are summarized in Figure 4. The information-based descriptors also included substructural flags indicating the presence of ionizable groups within a primary, secondary, or tertiary amine and/or a three- to six-membered ring.

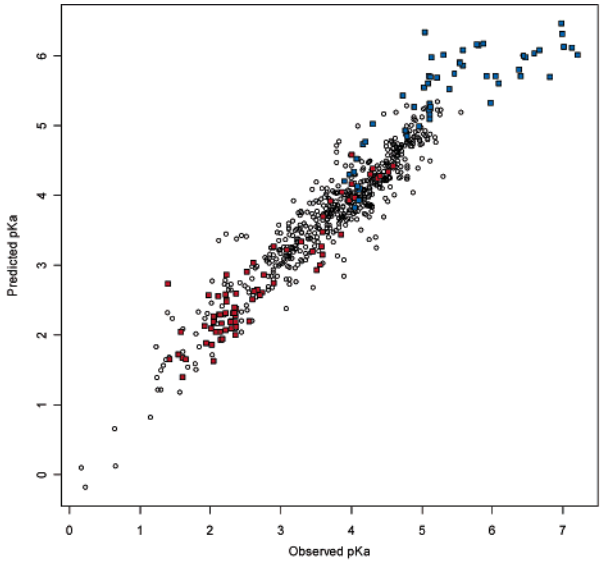


N_{CPDS}	descriptors	N_{LV}	R^2	RMSE	Q^2
311	SP	2	0.73	0.79	0.70
	IB	3	0.87	0.54	0.72
	SP and IB	3	0.90	0.49	0.78

Figure 4. Summary of predictive models derived using the anilines training set with predicted and observed pK_a shown for the combined (SP and IB) descriptor model.

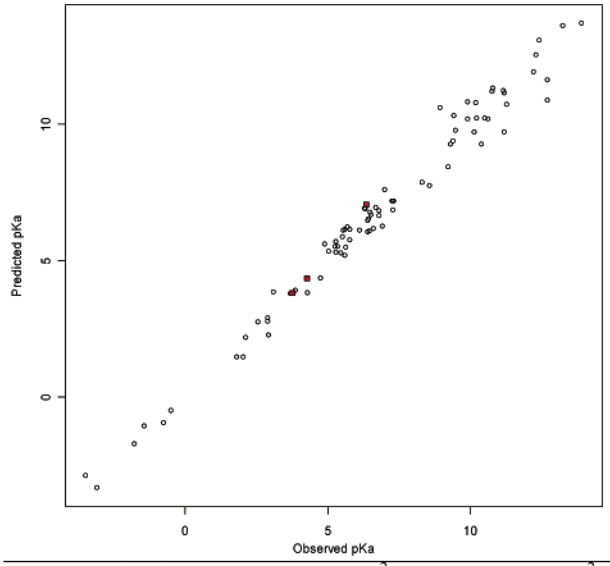
3.4. Carboxylic Acids Model. The carboxylic acids model was trained using compounds containing aliphatic and aromatic carboxylic acids (RCO_2H). These groups are reasonably acidic and therefore usually ionized at physiological pH-levels.

Compounds containing an additional basic amine or aniline group, both of which were protonated prior to training, or a less acidic alcohol group were also included in the data set. However, compounds containing weaker basic groups, including pyrimidines and pyridines, were excluded since the pK_a of these groups were difficult to distinguish from the pK_a of the strongly acidic carboxylic acids. The resultant predictive models are summarized in Figure 5. The semiem-



N_{CPDS}	descriptors	N_{LV}	R^2	RMSE	Q^2
681	SP	4	0.79	0.49	0.78
	IB	5	0.89	0.36	0.82
	SP and IB	4	0.90	0.34	0.86

Figure 5. Summary of predictive models derived using the carboxylic acids training set with predicted and observed pK_a shown for the combined (SP and IB) descriptor model.

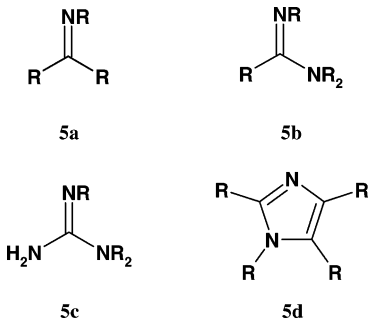


N_{CPDS}	descriptors	N_{LV}	R^2	RMSE	Q^2
84	SP	3	0.89	1.28	0.88
	IB	2	0.86	1.44	0.74
	SP and IB	4	0.98	0.55	0.88

Figure 6. Summary of predictive models derived using the imines training set with predicted and observed pK_a shown for the combined (SP and IB) descriptor model.

pirical properties of both oxygen atoms in the carboxylic acid were included in the respective models.

3.5. Imines Model. The imines model was trained using compounds containing imine (5a), amidine (5b), guanidine (5c), and imidazole (5d) groups. In each case, protonation of the more basic sp^2-hybridized nitrogen was considered. The sp^2-hybridized nitrogen of imines is more electronegative and therefore tends to be less basic than amine groups. However, stabilization of the conjugate acid formed by guanidines and, to a slightly lesser degree, amidines makes these related groups strongly basic. However, aromatic imidazoles are relatively stable and therefore tend to be weakly basic. The sp^3-hybridized nitrogen in the latter three groups is considerably less basic than the imine nitrogen, with either no stabilization of the ionized form occurring or aromatic stabilization being lost in the case of the imidazoles.



5a **5b** **5c** **5d**

Unsubstituted amidine, guanidine, and imidazole groups often form multiple tautomer forms. Training was therefore restricted to compounds where the resultant tautomers were equivalent to one another. This removed many outliers from the training set that otherwise resulted from either tautomer effects or the selection of the least prominent tautomer prior to training. Protonation of the compounds that were included in the training set, however, led to the formation of principal resonance structures that were equivalent. For these structures, an appropriate statistical factor therefore had to be subtracted prior to training, since each of the equivalent nitrogen atoms that were formed had an equal chance of losing a proton.

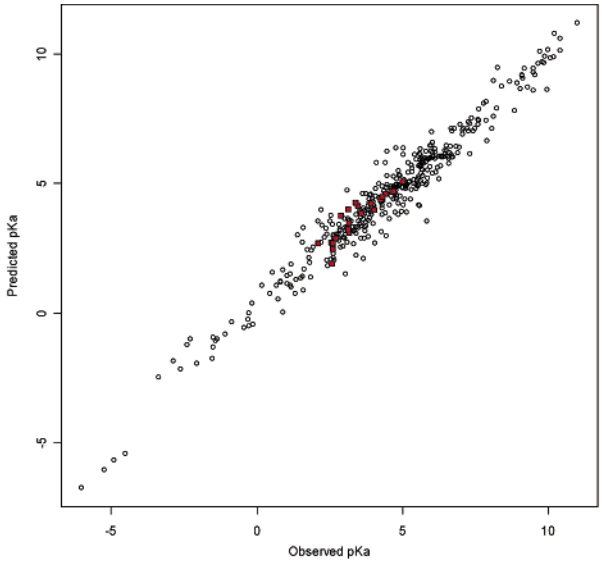
Due to the restriction placed on the training set compounds, the current models are limited to compounds containing just the single ionizable groups of interest. The resultant predictive models are summarized in Figure 6.

3.6. Pyridines Model. The pyridines model was trained using compounds containing pyridine (6a) and aminopyridine (6b) groups, with the more basic pyridine nitrogen considered to undergo protonation prior to the planar nitrogen in the latter case.

6a **6b**

Compounds containing additional weakly acidic aliphatic alcohol groups or basic amine groups were also included, with the latter being protonated prior to training. Aniline, phenol, carboxylic acid, and pyrimidine groups were not included in the current results, since the pK_a of these groups are often difficult to distinguish from the pK_a of the moderately basic pyridine groups. The resultant predictive models are summarized in Figure 7.

3.7. Pyrimidines Model. The pyrimidines model was trained using compounds containing pyrimidine (7a) and aminopyrimidine (7b) groups. Similar to the pyridines, the



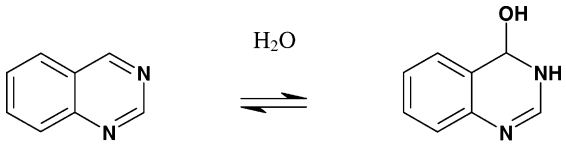
N_{CPDS}	descriptors	N_{LV}	R^2	RMSE	Q^2
397	SP	2	0.72	1.41	0.71
	IB	3	0.91	0.80	0.81
	SP and IB	4	0.95	0.58	0.86

Figure 7. Summary of predictive models derived using the pyridines training set with predicted and observed pK_a shown for the combined (SP and IB) descriptor model.

pyrimidine nitrogen was assumed to be more basic than the planar nitrogen in the latter group.

7a **7b**

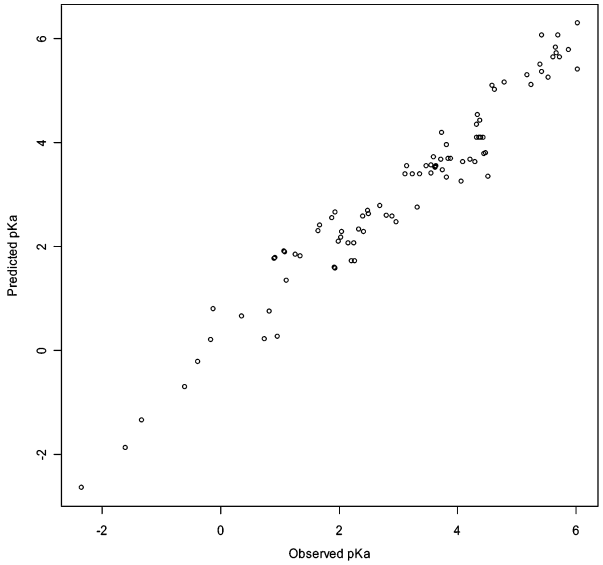
Since there are two nitrogens that could undergo protonation, only symmetric groups were considered in this initial study, i.e., where the protonated of either nitrogen was arbitrary. However, as with the imines model, an appropriate statistical factor was taken into account prior to training. Compounds containing quinazoline (7c) groups were also excluded because of this restriction. These compounds also have the potential to form hydrated species[20] (7d), and exclusion of such compounds further reduced the number of outliers observed for this model.



7c **7d**

These restrictions resulted in a relatively small training set of compounds, which was exacerbated by the pyrimidines being relatively weak bases. The current training set was also limited to compounds containing a single ionizable group. The resultant predictive models are summarized in Figure 8.

3.8. External Test Set. The models resulted in a combined RMSE of 0.81 for the external test set of 350 compounds (Figure 9). These compounds included 14 structures that contained a more basic or acidic (charged) group in addition



N_{CPDS}	descriptors	N_{LV}	R^2	RMSE	Q^2
91	SP	2	0.82	0.81	0.81
	IB	4	0.96	0.39	0.85
	SP and IB	3	0.95	0.43	0.87

Figure 8. Summary of predictive models derived using the pyrimidines training set with predicted and observed pK_a shown for the combined (SP and IB) descriptor model.

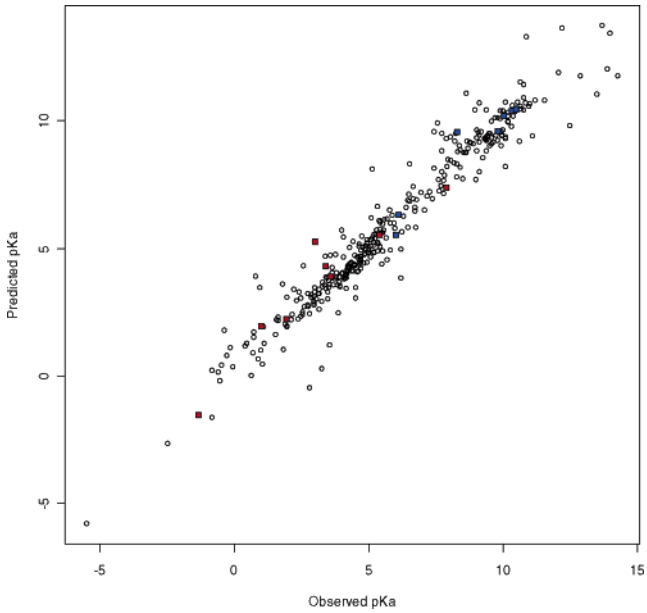


Figure 9. Predictions made for the validation set using both the SP and IB descriptors.

to the group undergoing ionization; many of these are also available in the literature (see Figure 10a−l). The worst prediction was obtained for the imidazole moiety of compound (10f) with an error, ΔpK_a, of 2.27. However, since the imines training set only contained three such charged compounds, this is not too unexpected.

3.9. Further Ionizable Groups. Predictive models have also been successfully derived for other ionizable groups including, hydroxamic acids, oxazoles, oximes, sulfonamides, and thiazoles using further experimental data available in-house. Unfortunately, reliable models are still not available for some important groups, such as phosphonic acids and

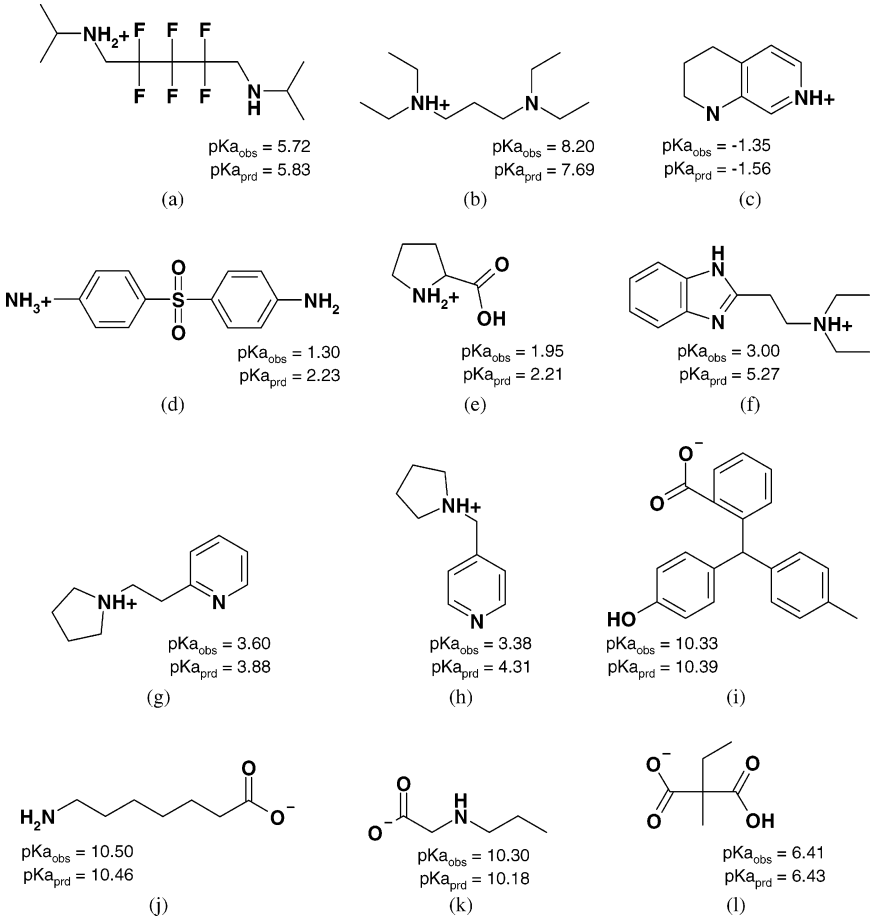


Figure 10. Predicted pK_a values for charged species in the test set with observed pK_a-values available in the literature (a,[23] b,[24] c,[25] d,[26] e,[27] f,[28] g and h,[29] i,[30] j,[31] k,[32] l[33]).

acidic tetrazoles. However, we are in the process of rectifying this in collaboration with the physical chemistry group at Novartis.

4. DISCUSSION

The semiempirical properties and information-based descriptors explored in this study modeled ionization constants effectively for a variety of functional groups. The close agreement between the R^2 and Q^2 of the models derived using the semiempirical properties suggest that these descriptors were unlikely to overfit the training data. In addition, replacing the absolute SE value with a series of relative values successfully modeled the effects of ionized groups upon the ionization of a further group of interest within multiprotic structures. For the 324 observations in the training set that involved a charged species all of the associated models provided reasonable predictions. However, the typically high RMSE of the models based on the semiempirical properties suggests that despite being highly extrapolative many outliers still persist. This is possibly since the properties are good at modeling electronic effects, but their effectiveness is greatly reduced when these effects are not dominant. This is particularly evident for the amines model, with an R^2 of 0.56, where the aliphatic nitrogen is insulated by adjacent sp^3-hybridized carbons. In contrast, for the anilines model, with an R^2 of 0.73, the basicity of the planar nitrogen is highly dependent on the stabilizing electronic (mesomeric and inductive) effects of substituents attached to the aromatic ring. Models based on the information-based descriptors, in comparison, provided a much better fit to the training compounds, with typically lower RMSE values, demonstrating their capability to encode many more of the structural phenomena affecting pK_a. This can be attributed to their ability to exploit the information resources available, although the greater discrepancy between the R^2 and Q^2 values of the resultant models suggests that some overfitting to the training data may have occurred, particularly for the smaller training sets. Overall, the two classes of descriptor complement one another well, with the combined models being highly predictive and providing a good fit to the training data.

The quality of the models was further exemplified by their ability to reproduce the correct ionization order of ionizable groups in all of the multiprotic structures studied. This was successfully achieved using the algorithm introduced in this paper, where a consensus of predictions is used to identify the most basic group in a structure during the stepwise assignment of ionization constants. This required the models to be applicable to the 'hypothetical' ionization states compared at each step and was certainly aided by the semiempirical properties, which allowed highly extrapolative models to be derived. In contrast, when using only the information-based descriptors, the models often provided unreliable predictions for compounds that were dissimilar to the training set compounds. This was particularly problematic for weak acids and bases, where the models tended to overestimate their respective acidity or basicity.

The collation of reliable training data for pK_a prediction was particularly tedious and problematic. Much effort was required to assign pK_a values to the correct groups and

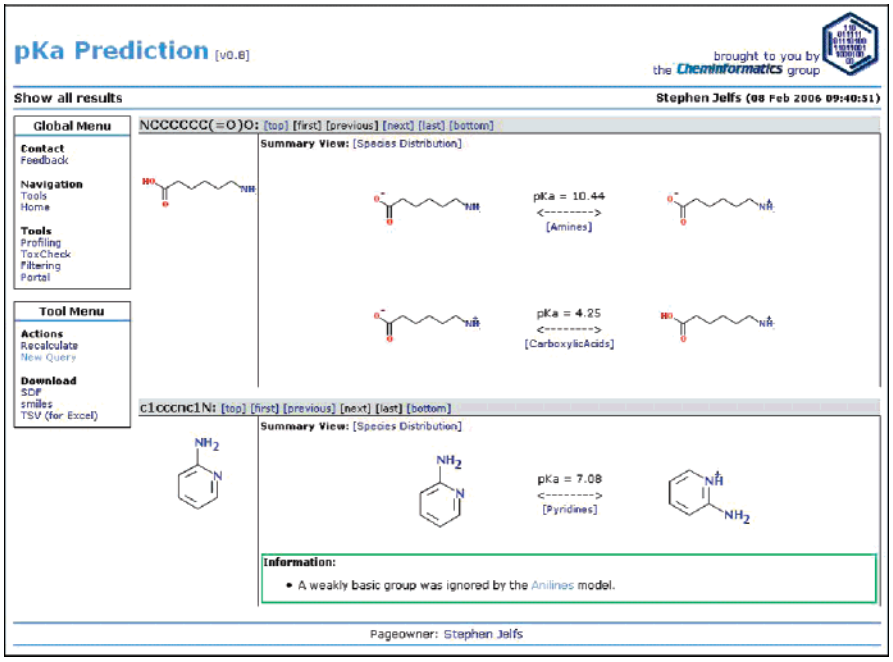


Figure 11. Web-tool for pK_a prediction.

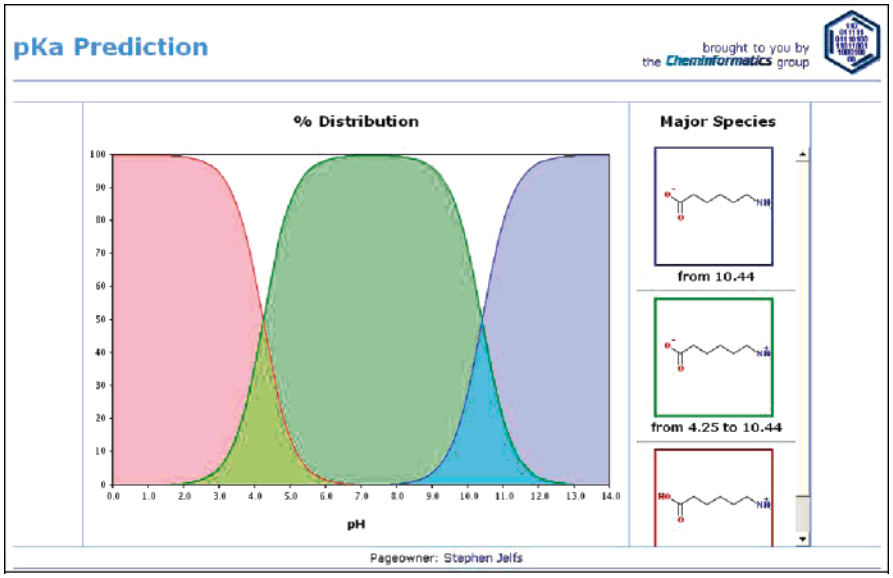
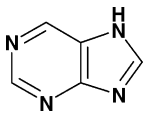


Figure 12. pH/ionization state distribution.

identify any stronger acidic and basic groups that were present. This procedure would certainly benefit from experimental methods to identify site-specific pK_a values, for example, based on structure elucidation techniques such as NMR. However, such time-consuming procedures are probably not feasible in the near-term.

The methods developed during the study appear to be applicable to a wide range of functional groups in mono- and multiprotic structures. However, the availability of suitable training data severely limits our ability to generate models for all conceivable functional groups and functional group combinations. The determination of the applicability domain of each model also requires some consideration. For instance, predictions for purine[21] could be made using the models developed in this study: the pyrimidine model providing a good estimation for the second ionization constant (observed = 2.3/predicted = 2.5) and the imine model providing a, not so good, prediction, for the first ionization constant (observed = 9.0/predicted = 5.7). The applicability of the latter model may therefore be questionable, and the generation of a more targeted model for purines may be required. We are currently working hard to increase the number of functional groups modeled and also to define rules for the application of these models to more complex functionality. Unfortunately, as previously stated, the availability of suitable training data is a severe hindrance—particularly when trying to implement this approach using public domain data.



5. WEB APPLICATION

The combination of descriptors and the predictive algorithm provided not only reasonable predictions but also a

pragmatic approach for pK_a prediction. This allowed a Web-based tool for pK_a prediction to be developed for use by the medicinal chemists at Novartis (see Figure 11). Importantly, the tool provides predictions for all of the ionizable groups in a structure with no prior knowledge of the ionization order of these groups being required. Appropriate warnings, information, and advice is also provided to the users when tautomers, potentially hydrated species, and other problematic groups are encountered. The interpretation of the results is aided graphically using plots showing the distribution of major species formed by compounds at varying pH-levels (see Figure 12). Finally, a Wiki[22] is provided which summarizes the descriptors, training compounds, and predictive ability of the models used by the Web-tool. This feature provides an informal forum for the exchange of ideas between the users and development team and should provide feedback for the continued development of the tool.

6. CONCLUSIONS

The semiempirical properties and information-based descriptors have been shown to provide consistently good models for the estimation of pK_a. The series of semiempirical properties used in this study also allows the derivation of pK_a for multiprotic compounds, overcoming limitations highlighted in previous publications. The addition of the information-based descriptors provided models with an excellent fit to the training data and allowed the large amounts of available training data to be fully exploited. The prediction algorithm successfully combined the multiple descriptor models to provide reasonable predictions for multiprotic compounds. This approach has also allowed the development of a Web-tool for pK_a prediction at Novartis. Further applications are currently being investigated, including the enumeration of chemical species prior to protein docking studies and logD prediction for druglike compounds.

ACKNOWLEDGMENT

The authors would like to thank Professor Peter Bladon for providing the modified version of Mopac. Insightful discussions with Drs. Bernhard Rohde, Peter Gedeck, Joerg Muehlbacher, and Nathan Brown were also appreciated. We would also like to acknowledge the medicinal chemists involved in testing the Web-tool, Dr. Barnard Faller and Frederique Loeuillet for providing experimental data, and Wolfgang Zipfel for his technical support.

Supporting Information Available: Lists of the substructural patterns used to identify the ionization groups included, the regression coefficients, and associated parameter settings for each of the predictive models. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Montana, J. G.; Manallack, D. T.; Gancia, E. Estimation of pK_a Using Semiempirical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids. *Quant. Struct.-Act. Relat.* **2002**, *21*, 457−472.

(2) 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.

(3) 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.

(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) *BioLoom, Version 2004*; BioByte Corp.: 201 W. 4th St., #204, Claremont, CA 91711-4707.

(6) Comer, J.; Tam, K. Lipophilicity Profiles: Theory and Measurement. In *Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical and Computational Strategies, E*; Testa, B., van de Waterbeemd, H., Folkers, G., Guy, R., Eds.; VHCA: Zurich, 2001; pp 275−304.

(7) 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.* **2002**, *75*, 883−392.

(8) *ACD/pK_a*; Advanced Chemistry Development Inc.: 110 Yonge St., 14th Floor, Toronto, Ontario, Canada M5C 1T4.

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

(10) Gross, K. C.; Seybold, P. G.; Peralta-Inga, Z.; Murray, J. S.; Politzer, P. Comparison of Quantum Chemical Parameters and Hammett Constants in Correlating pK_a Values of Substituted Anilines. *J. Org. Chem.* **2001**, *66*, 6919−6925.

(11) Bremser, W. HOSE - A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355−365.

(12) Ertl, P.; Selzer, P.; Mühlbacher, J. Web-based Cheminformatics Tools Deployed via Corporate Intranets. *Drug Discovery Today: BIOSILICO* **2004**, *2*, 201−207.

(13) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109−116.

(14) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(15) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.

(16) Stewart, J. J. P. *Mopac program package;* Quantum Chemistry Program Exchange no. 455.

(17) Dewar, M. J. S.; Zoebish, E. G.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular Models. AM1: a New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(18) Fukui, K.; Yonezawa, T.; Nagata, C. Theory of Substitution in Conjugated Molecules. *Bull. Chem. Soc. Jpn.* **1954**, *27*, 423−427.

(19) Clark, R. D.; Cramer, R. D.; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982−1012.

(20) Hilal, S. H.; Bornander, L. L.; Carreira, L. A. Hydration Equilibrium Constants of Aldehydes, Ketones and Quinazolines. *QSAR Comb. Sci.* **2005**, *24*, 631−638.

(21) Albert, A.; Barlin, G. B. 605. Ionization constants of heterocyclic substances. Part V. Mercapto-derivatives of diazines and benzodiazines. *J. Chem. Soc.* **1962**, 3129−3141.

(22) http://www.jspwiki.org/ (accessed Aug 2006).

(23) Marks, B. S.; Schweiker, G. C. Fluorine-containing Secondary Diamines. *J. Am. Chem. Soc.* **1958**, *80*, 5789−5792.

(24) Gero, A. Regularities in the Basicity of Some Tertiary Ethylenediamines, Trimethylenediamines and 2-Hydroxytrimethylenediamines. *J. Am. Chem. Soc.* **1954**, *76*, 5158−5159.

(25) Armarego, W. L. F. 813. Ionization and ultraviolet spectra of indolizines. *J. Chem. Soc.* **1964**, 4226−4233.

(26) Bell, P. H.; Roblin, R. O., Jr. Studies in Chemotherapy. VII. A Theory of the Relation of Structure to Activity of Sulfanilamide Type Compounds. *J. Am. Chem. Soc.* **1942**, *64*, 2905−2917.

(27) Smith, P. K.; Gorham, A. T.; Smith, E. R. B. Thermodynamic Properties of Solutions of Amino Acids and Related Substances. VII. The Ionization of some Hydroxyamino Acids and Proline in Aqueous Solution from One to Fifty Degrees. *J. Biol. Chem.* **1942**, *144*, 737−745.

(28) Edward, J. T.; Meacock, S. C. R. 385. Hydrolysis of amides and related compounds. Part III. Methyl benzimidate in aqueous acids. *J. Chem. Soc.* **1957**, 2009−2012.

(29) Barlow, R. B.; Hamilton, J. T. Effects of some isomers and analogues of nicotine on junctional transmission. *Br. J. Pharmacol. Chemother.* **1962**, *18*, 510−542.

(30) Schultz, O. E.; Fedders, S.; Holm, W. D.; Schulze, V. Relationships between structure and laxative action of triarylmethane derivates. *Arzneim.-Forsch.* **1974**, *24*, 1933−41.

(31) Tsai, R.-S.; Testa, B.; El Tayar, N.; Carrupt, P.-A. Structure−lipophilicity relationships of zwitterionic amino acids. *J. Chem. Soc., Perkin Trans. 2.* **1991**, *11*, 1797−1802.

(32) Basolo, F.; Chen, Y. T. Steric Effects and the Stability of Complex Compounds. III. The Chelating Tendencies of N-Alkylglycines and N-Dialkylglycines with Copper(II) and Nickel(II) Ions. *J. Am. Chem. Soc.* **1954**, *76*, 953−955.

(33) Gane, R.; Ingold, C. K. CCXCIV. Electrometric titration curves of dibasic acids. Part IV. *J. Chem. Soc.* **1931**, 2153−2169.

CI600285N