Predicting pK_a by Molecular Tree Structured Fingerprints and PLS

Li Xing,*,† Robert C. Glen,‡ and Robert D. Clark

Tripos, Inc., 1699 S. Hanley Road, St. Louis, Missouri 63144

Received November 5, 2002

This is the second phase of the pK_a predictor published earlier (*J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796−805). The algorithm has been extended by treating specific chemical classes separately and generating tree-structured molecular descriptors tailored to each individual class. A training set consisting of 625 acids and 412 bases covers the major areas of chemical space involved in protonation and deprotonation. The models obtained demonstrate excellent statistics (SE = 0.41 for acids and 0.30 for bases) and yielded accurate predictions on an external test set. The quality and statistical performance of pK_a prediction has been improved considerably over the initial implementation of the method.

INTRODUCTION

The bioavailability of a drug is an important consideration in rational drug design. Before a drug can elicit any effect it usually has to pass through at least one biological membrane by passive diffusion or by carrier-mediated uptake. Depending on the route of administration and the location of the target site, the pH of intervening aqueous environments may vary considerably. For any acidic or basic drug, the extent to which it partitions into a lipophilic environment and, hence, the ease with which it diffuses into the next aqueous compartment depends on its degree of ionization at the prevailing pH. Similarly, the affinity of a drug molecule for the target of interest or for active transport carriers may be critically dependent on the degree of dissociation at the locally prevailing pH. Indeed, strong electrostatic and hydrogen bonding interactions are often key contributors to overall free energies of binding.[1] It follows that the ability to accurately predict pK_a from the structure is critically important in designing drug candidates which are efficacious, particularly if oral bioavailability is to be “designed-in” at an early stage.

A second key determinant of aqueous solubility and intestinal absorption, both of which contribute significantly to bioavailability, is characterized by the octanol/water partition coefficient *P*. However, logP refers to the neutral state of molecules. For acids and bases, the ionization state must also be considered, since their partition is dependent on the pH of the aqueous phase. The pH-dependent distribution coefficient *D* is, then, a function of both *P* and the dissociation constant, pK_a. For simple monofunctional acids, for example:

$$\mathrm{logD} = \mathrm{logP} - \log(1 + K_a/[\mathrm{H}^+]) \tag{1}$$

$$\approx \mathrm{logP} - (\mathrm{pH} - \mathrm{p}K_a) = \mathrm{logP} - \Delta\mathrm{pH} \tag{2}$$

where the approximation holds at pH values well above the pK_a of the acid in question. At pH values well below the pK_a, dissociation of the acid will be negligible and eq 1 reduces to

$$\mathrm{logD} \approx \mathrm{logP} \tag{3}$$

Besides its critical role in determining pharmacological behavior, pK_a can also be important in determining chemical reactivity. If too strong a base is used during sulfonylation, for example, the intermediate sulfonamide may deprotonate, leading to competing disulfonylation (R. D. Clark, unpublished observation).

There is a growing need for reliable estimates of physicochemical parameters for compounds yet to be synthesized, particularly in new drug discovery and combinatorial library design applications. Nor is it always convenient or practical to perform experimental measurements on existing compounds. Hence it is useful to develop broadly applicable and accurate models for predicting pK_a and, hence, logD from molecular structures. Several robust programs already exist for estimating logP (e.g. the popular ClogP program[2]), but a method for getting accurate pK_a predictions across a diverse set of structures has proven more elusive. In fact, a study of logD prediction using the PrologD[3] program indicates that errors in predicted logD were largely due to poor pK_a predictions.[4]

In principle, dissociation constants can be calculated directly using quantum mechanics. This is an attractive approach because it should be both accurate and general. The main practical difficulties lie in how to take effects of solvent molecules fully into account and how to properly describe the modifications in a solute molecule due to interaction with the solvent. This is a particularly difficult task in the most relevant medium—water. Ab initio methods are universally applicable but are impractical for large systems, especially for high throughput virtual screening applications.[5−8] Semiempirical quantum mechanics methods have also been applied, and their results compared with those from ab initio and density functional theory methodologies.[9,10]

pK_a values can also be calculated using formalisms from statistical thermodynamics. In particular, considerable work

* Corresponding author phone: (636)247-5466; fax: (636)247-7607; e-mail: li.xing@pharmacia.com.

† Current address: Pharmacia, 700 Chesterfield Parkway North, BB4I, Chesterfield, MO 63198.

‡ Current address: The Unilever Center for Molecular Informatics, The University Chemical Laboratory, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, U.K.

10.1021/ci020386s CCC: $25.00
Published on Web 03/13/2003

has been done using numerical solutions of the Poisson–Boltzmann equation.[11–14] Recently a method using the Debye–Huckel formalism was developed for predicting pK_a's of amino acid residues in proteins. Environmental effects are particularly important in this context, which makes it difficult to estimate pK_a's accurately.[15]

Comparative molecular field analysis (CoMFA), a widely used 3D quantitative structure–activity relationship tool, has been used to model pK_a values for small sets of structures (between 30 and 50 molecules) drawn from specific chemical series.[16–18] This application, however, involved establishment of consistent molecular conformations and well-defined alignment rules. Hence, it is not readily automated or generally applicable across structural series.

In 1981 Perrin et al. published a book on pK_a prediction which is widely used.[19] Depending on the nature of the chemical structures, different algorithms were recommended for estimating pK_a.

Fragment-based methods have proven very useful and are available in commercial programs,[20,21] but they are limited in scope. Because every prediction is based on a congeneric parent structure, pK_a's can only be reliably predicted for compounds very similar to those in the training set, making it difficult or impossible to get good estimates for novel structures. A further disadvantage is the need to derive a very large number of fragment constants and correction factors, a process which is complicated and potentially ambiguous. Although this is probably the most widely used method, the accuracy and extensibility of the predictions obtained have not been gratifying.

Here we describe a method based more on an informatics approach. The first phase of the work has been detailed elsewhere.[22] It is parametrized using experimental data for a large training set—625 acid and 412 bases spanning a very broad range of chemical classes. Acids covered include aliphatic and aromatic carboxylic acids; sulfonic and sulfinic acids; phenols, thiophenols, alcohols, and thiols; and acidic carbon and nitrogen centers. The base data set includes alkyl, allylic, and cyclic amines; pyridines, imidazoles, and other heteroaromatic bases; and anilines. The predictive errors are less (usually much less) than half a log unit (decade) for the majority of the data set (>89%), a precision that has not been achievable for such an extensive and diverse data set by other methods.

METHODS

Our goal was to develop a method which is accurate, simple, and fast but which does not require knowledge of 3D structure or quantum mechanical or molecular dynamics calculations.[23] Rather than use explicit fragment values, we rely on identifying generic substructures for basic and acidic centers and a few key fragments (e.g., nitro groups) and then characterizing all other atoms by atom type—in particular, by their SYBYL atom type.[24] Because the atom typing takes neighboring atoms into account, fragment effects are accounted for implicitly. Based on the principle that the degree of ionization of a particular group is dependent upon its subenvironment as defined by the neighboring atoms and bonds, a connection tree is constructed from the acidic or basic center outward. This contains the atoms directly connected to the root atom at the first level, those bonded to the first level at the second level, and so on. A count vector is then constructed based on the total number of atoms of each type at each level, originating from the root. Each vector is made up of the occurrences and positions of each atom type in the neighborhood of the ionized center, thereby fully characterizing the ionizing center.

Atom typing alone, however, gives insufficiently predictive models in many instances. Hence certain chemical groups have to be treated explicitly, especially those involving delocalized π-electron systems—i.e., nitro, nitroso, cyano, carbonyl, carboxylate, sulfone, sulfonate, sulfoxide, sulfinate, hydroxyl, and sulfhydryl groups. The carboxylate, sulfonate, and sulfinate groups are particularly important. Because their protonation state will affect dissociation at other positions, e.g. suppressing ionization of less acidic centers and enhancing basicity elsewhere in a molecule, they were parametrized as the anions. This is critical for accurately predicting secondary pK_a's for compounds with several acidic centers and for amphoteric species such as amino acids.

Using the atom and group definitions discussed above, 33 counts (22 atom types and 11 group types) are determined at each connection level. For the ionizing center one or more of the descriptive variables were used depending on the chemical class it belongs to. The maximum number of descriptor elements for the five levels used here is therefore $5 \times 33 = 165$. Atom and group types not found at particular levels in the training set (e.g., there are no compounds in which sulfur is bonded to the basic nitrogens in alkylamines, i.e., at the first level) were omitted from the descriptor.

These count vectors for each type of acidic and basic centers were then combined with the corresponding pK_a values and used as input to create and cross-validate models via the partial least squares projection onto latent structures (PLS) facility in SYBYL 6.7.[24,25]

The PLS models generated by this principle are of generic form as follows

$$\mathrm{p}K_a = \mathrm{p}K_c^{\,0} + \Sigma a_i x_i + \Sigma g_j y_j + \Sigma q_k z_k \qquad (4)$$

where pK_c^0 is the base dissociation constant for the specific subclass into which each ionizing center falls. The summations represent corrections to this base value for effects of individual atoms and groups elsewhere in the molecule. The x_i and y_i terms indicate the number of occurrences of the corresponding atom type and group, respectively, at each level, whereas z_k's correspond to class-specific indicator variables. The coefficients a_i, g_j, and q_k were obtained from PLS analyses applied across all structural subclasses within each class for the corresponding atom types and groups at each connectivity level; indicator variable coefficients q_k were obtained the same way. The summations cover all possible atom types and groups up to five levels away from the ionizing center as well as the relevant class-specific indicator variables.

DATA SETS

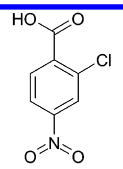
The main source of the pK_a data for the training set was Lange's Handbook of Chemistry.[26] A number of these values or their substructural assignments, however, were recognized as dubious as the data were being transcribed for analysis. Other compilations of dissociation constants[27,28] were used to resolve the ambiguities. These latter sources also served to substantially augment the training set for several structural classes underrepresented in Lange's Handbook. Data for (benz)imidazole pK_a's, for example, were mostly compiled from the latter, because too few examples were provided in the former compilation. In the process of model building outliers were identified, which frequently turned out to be either misassignment of pK_a values or suspicious molecular structures. Suitable corrections were made where possible, but in some cases the corresponding data had to be omitted from the training set. In other cases, outliers served to point up a need to split one class into two or more subclasses based on the substructure in which the acidic or (more often) basic center is embedded. Such modifications were kept to a minimum and only applied where reasonable physical chemical rationale existed. The resulting training set consisted of 625 acids and 412 bases.

The quality of the PLS models generated were assessed using several regression statistics: r^2 as a measure of fit; its cross-validated counterpart, q^2, as a measure of predictivity; and the standard error as a measure of precision. The latter is useful for comparing results between classes, since it is independent of the pK_a range of the training set. Calculation of q^2 was carried out for the complete data sets for acids and bases but not for the individual chemical classes. For fast evaluation and comparison purposes correlation coefficient r^2 was computed using 10 principal components (the maximum acceptable number) for each class of compounds, unless otherwise noted.

RESULTS FOR ACIDS

The acids were divided into four classes: aromatic acids; aliphatic acids and alcohols; phenols and thiophenols; acidic carbons and acidic nitrogens.

(1) Aromatic Carboxylic, Sulfonic, and Sulfinic Acids. Aromatic systems are distinguished from the others by resonance effects due to electron delocalization. Substitution at the *ortho*, *meta*, and *para* positions have distinct electronic effects both in nature and in magnitude. To take this effect fully into account the molecular connectivity tree was grown from three base nodes, one for each substitution position. For example 2-chloro-4-nitrobenzoic acid maps onto its connectivity tree as described in the following table:



2-chloro-4-nitrobenzoic acid (pKa: 1.96)

	level 0			level 1				
				ortho		meta		
	ortho C.ar	meta C.ar	para C.ar	H	Cl	H	COO	para NO2
count	2	2	1	1	1	1	1	1

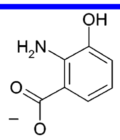
The carbon atoms in the benzene ring map onto the zeroth level, which allows coverage of aromatic systems incorporating heteroatoms (e.g. pyridine). Each of the three subclasses of acid in this group was assigned a separate indicator variable to capture the underlying subclass acidity pK_c^0, but the frequencies for each atom type or group at each level (out to five bonds removed from the atoms in the zeroth level) were shared across subclasses. The subsequent regression was forced through the origin by setting the CENTERING configuration in the SYBYL PLS module to YES.

The model for 143 molecules including aromatic carboxylic, sulfonic, and sufinic acids is summarized by

$$r^2 = 0.94,\ \mathrm{SE} = 0.258,\ F = 198,\ k = 93,\ N = 143$$

where k is the number of input descriptor elements and N is the number of compounds used to derive the model.

(2) Phenols and Thiophenols. There are only six thiophenols in the data set, so these are folded into the phenol subclass. The construction of molecular trees for this class of molecule follows rules similar to those described above, except that the ionizing centers now become OH or SH groups connected to an aromatic ring. 2-Amino-3-hydroxybenzoic acid is a simple example. At the pH level of around 10 when the phenol titrates, the benzoic acid is deprotonated, and the anilino nitrogen is not protonated. This is the overall protonation state used for the tree construction.



2-amino-3-hydroxybenzoic acid (pKa3: 10.12)

	level 0			level 1				
				ortho		meta		
	ortho C.ar	meta C.ar	para C.ar	H	N.pl3	H	COO	para H
count	2	2	1	1	1	1	1	1

	level 2			level 3		
	ortho	meta	para	ortho	meta	para
	H					
count	2					

In this case, the class model gives

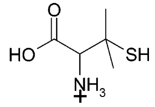
$$r^2 = 0.99,\ \mathrm{SE} = 0.195,\ F = 802,\ k = 91,\ N = 120$$

An attempt was made to pool the aromatic acids with the aromatic alcohols to yield a single model. The combined data set of 263 acids exhibited an unacceptable increase in standard error:

$$r^2 = 0.97,\ \mathrm{SE} = 0.536,\ F = 825,\ k = 190,\ N = 263$$

Were the interactive effects the same within each class, we would expect the combined model to be comparable in quality to the separate ones. In fact, however, the standard error for the mixed set (0.536) was more than double that of either class evaluated separately (0.258 for acids and 0.195 for alcohols). This implies that the effect of the substitutions around the aromatic ring, although broadly similar in nature (the combined model being still statistically significant), differ substantially between (thio)phenols and other aromatic acids.

(3) Aliphatic and Alicylcic Carboxylic, Sulfonic, and Sulfinic Acids. The deprotonation centers in this class are specified in SYBYL line notation (SLN)[29] as C(=O)OH (carboxylic acids), S(=O)(=O)OH (sulfonic acids), and S(=O)OH (sunfinic acids). Specific atom types and groups at successively greater topological distances from the acidic group were mapped onto a connectivity tree for each structure. 2-Amino-3-mercapto-3-methylbutanoic acid, for example, exhibits three pK_a's: pK_{a1} for the carboxylic acid is 1.80, pK_{a2} for the thiol is 7.9, and pK_{a3} for the amino group is 10.5. As the carboxylate titrates at low pH, the amino is protonated and the sulfhydryl group is neutral. The molecule is evaluated in this protonation state so that appropriate structural features can be captured in the molecular tree. The table that describes results of the mapping is given by



2-amino-3-mercapto-3-methylbutanoic acid (pK_{a1}: 1.8)

	level 0	level 1	level 2			level 3			level 4
	COOH	C.3	C.3	N.4	H	C.3	H	SH	H
count	1	1	1	1	1	2	3	1	6

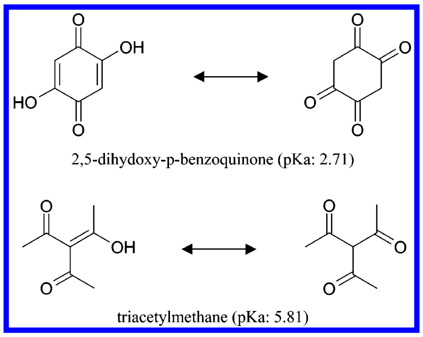
The PLS model obtained for aliphatic and alicyclic acids is statistically characterized by

$$r^2 = 0.95,\ \mathrm{SE} = 0.248,\ F = 541,\ k = 87,\ N = 306$$

(4) Aliphatic and Alicyclic Alcohols and Thiols. Aliphatic alcohols are generally not appreciably acidic in the absence of proximal strong electron withdrawing groups. Thiols are considerably more acidic than alcohols. Hence there is little experimental data available for this class of acids, but data for five alcohols and 18 thiols were available. Consolidation of this structural class into the aliphatic acids did not significantly compromise the statistics for the model obtained:

$$r^2 = 0.98,\ \mathrm{SE} = 0.287,\ F = 1466,\ k = 94,\ N = 329$$

(5) Acidic Nitrogens and Carbons. When strong electron withdrawing groups (e.g. nitro, nitrile, carbonyl, etc.) are attached to nitrogen or carbon the proton on the nitrogen or carbon atom may become appreciably acidic. Multiple tautomers usually coexist for these compounds, which can make it difficult to determine which class it belongs to. Two examples are given below:



2,5-dihydoxy-p-benzoquinone (pKa: 2.71)

triacetylmethane (pKa: 5.81)

The enol tautomers shown at the left for each equilibrium belong to the hydroxy class, while those on the right describe acidic carbon centers. At least for the representatives in hand, the pK_a's observed for these compounds indicate that they are better handled as acidic carbons rather than as acidic alcohols.

Geometric constraints can alter pK_a dramatically when the acidic carbon atom is part of a ring, e.g. a pK_a of 5.3 for 1,3-cyclohexanedione versus ~10 for noncyclic analogues. This necessitates inclusion of an indicator variable when the acidic carbon is in a ring. There are 21 examples available from this structural class, with pK_a values ranging from 1 to almost 12. The carbon centers are substituted by nitro, nitrile, and carbonyl groups, often multiple instances and/or in combination. Because of the relative scarcity of data and a lack of consensus for pK_a measurements made by different groups, the statistical models obtained for this class are not as good as for the other classes. Five principal components were used for these smaller sets of data. For the acidic carbon centers

$$r^2 = 0.92,\ \mathrm{SE} = 1.010,\ F = 34,\ k = 23,\ N = 21\ (5\mathrm{PC})$$

Twelve molecules contain nitrogen as acidic centers, and the following parameters summarizes the model for them:

$$r^2 = 0.98,\ \mathrm{SE} = 0.326,\ F = 66,\ k = 22,\ N = 12\ (5\mathrm{PC})$$

After combining the acidic carbons with the acidic nitrogen centers the model retained statistical significance (note the increase in the F statistic). In particular, it showed substantial improvement over that obtained for acidic carbons alone:

$$r^2 = 0.93,\ \mathrm{SE} = 0.791,\ F = 71,\ k = 32,\ N = 33\ (5\mathrm{PC})$$

(5) Complete Set for Acids. The complete acid set contains 625 structurally diverse molecules representing each of the chemical classes discussed separately above. The PLS model used 290 individual connectivity descriptors (all atom types, groups and separations) and yielded a correlation coefficient and standard error of

$$r^2 = 0.98,\ \mathrm{SE} = 0.405,\ F = 2766,\ k = 290,\\ N = 625;\ q^2 = 0.86,\ \mathrm{SE_x} = 1.04$$

where q^2 and $\mathrm{SE_x}$ denote the cross-validated correlation coefficient and cross-validated standard error, respectively.

The biggest predictive errors are for the molecules containing an acidic carbon or nitrogen center. This reflects the complex nature of the equilibrium involved as well as under-representation in the training set. If these 33 compounds were excluded from the training set, the model improved remarkably to a correlation coefficient of 0.99 and a standard error of 0.310.

The cross-validated q^2 was obtained by randomly dividing the data set into 10 groups, with each group then predicted by the model derived from compounds in the other nine groups. The initial partitioning was repeated five times, and the statistics obtained were averaged across all 50 trials. The average cross-validated q^2 and r^2 values obtained are plotted in Figure 1 as functions of the number of principal components included in the PLS model. The un-cross-validated correlation coefficient r^2 starts at 0.63 for one component and plateaus at 0.98 around 10 components. The q^2 slowly increases to 0.9 at eight components, with no fall off apparent up to 10 components.

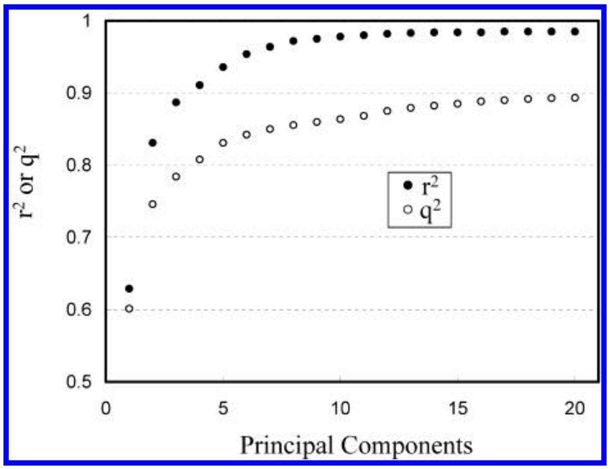


Figure 1. Plot of q^2 and r^2 for 625 molecules in the acids data set.

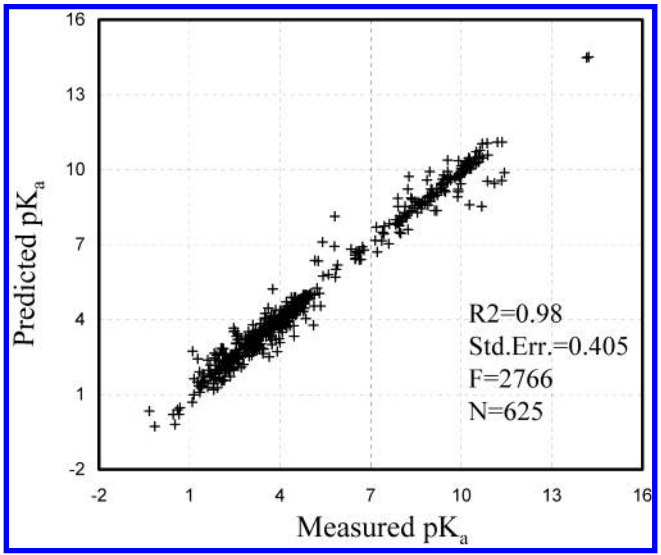


Figure 2. Predicted vs measured pK_a values for the acids data set.

The pK_a values predicted by the aggregate model are shown as a function of the values obtained experimentally in Figure 2. The regression line has an intercept of 0.000 and a slope of 1.000, indicating that the model is not appreciably biased. Errors are distributed evenly along the regression line, so no systematic errors as a function of pK_a are evident.

The distribution of residuals is plotted in Figure 3. The two highest peaks fall between −0.5 and 0.5, with counts of 273 and 260 respectively (they are truncated at 50, otherwise the much shorter bars at larger residual ranges would be nearly invisible). These 533 molecules predicted within half a log unit constitute more than 85% of the data set. A total of 602 compounds (more than 96% percent) are predicted correctly within 1 log unit. Most poorly predicted molecules (more than 1.5 log units in error) involve acidic carbon or nitrogen—they constitute seven of the eight molecules giving errors in that range. Although important pharmacologically, these are the most poorly represented and the most poorly characterized class considered here, so it is no great surprise that they are also the most poorly predicted.

RESULTS FOR BASES

We were able to get accurate predictions for nitrogenous bases by separating them into four classes: pyridines, anilines, imidazoles, and alkylamines.

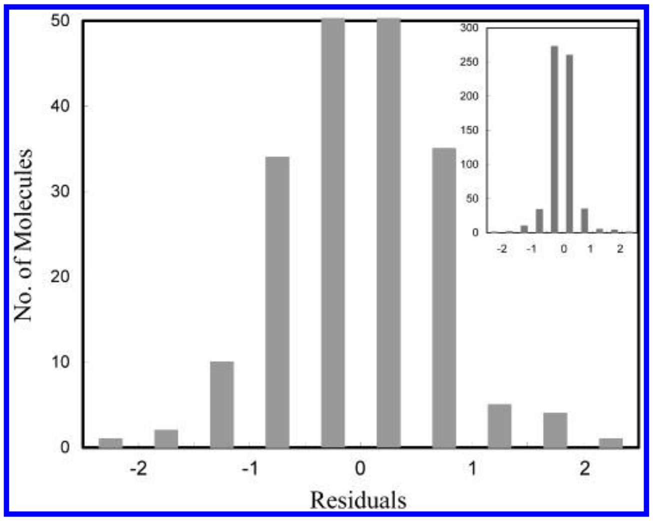
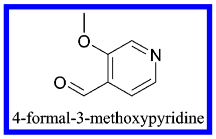


Figure 3. Distribution of residuals for acids data set. The two highest peaks are 273 and 260, respectively, but are truncated at 50. The upper-right panel is the full-scale plot.

An attempt to consolidate the pyridines and the anilines into a single model failed to yield good predictions for either class. Careful examination of outliers showed that similar substituents at analogous positions around the respective aromatic rings influence the pK_a of the ionizing center to quite different degrees. In general, the elevation or suppression of pK_a resulting from electron withdrawal or electron donation was far greater for pyridines than for the corresponding anilines. This may be due to the fact that protonated lone pair has much more *s* character for pyridines, whereas *p* character dominates in anilines.

(1) Pyridines. The pyridine class was generalized to include all six-membered aromatic rings containing at least one nitrogen. Hence pyridazines, pyrimidines, and pyrazines are all included in this class.

Here, the molecular tree is rooted at the pyridine nitrogen. Because of resonance effects, positions for substituents around the aromatic ring are particularly pivotal. Depending on the connecting points to the ring the tree was separated into three branches, specifically the ortho, meta, and para subtrees. The connectivity tree for 4-formal-3-methoxypyridine is given below:



4-formal-3-methoxypyridine

	level 0			level 1			
	ortho	meta	para	ortho	meta		para
	C.ar	C.ar	C.ar	H	H	O.3	C=O
count	2	2	1	2	1	1	1

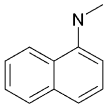
	level 2			level 3		
		meta	para		meta	
	ortho	C.3	H	ortho	H	para
count		1	1		3	

The empty cells or the nodes that were not included in the table were filled by zeros. Note the presence of the aldehyde group (represented by “C=O”) at level 1 for the para position.

There are 98 pyridine-type (pyridazine, pyrimidine, pyrazine, quinolines, isoquinolines, etc.) compounds in the training set. The statistics for the model derived from this class are

$$r^2 = 0.99, \mathrm{SE} = 0.239, F = 827, k = 84, N = 98$$

(2) Anilines. Any aromatic six-membered ring bearing a protonable nitrogen was placed in the aniline class. Thus the method covers aromatic systems containing heteroatoms as well as simple anilines. The atoms in the aromatic ring and the two exocyclic atoms directly bonded to the aniline nitrogen comprise the zeroth level. This ensures that the molecular connectivity trees for anilines correspond directly to those for pyridines at any given level. Exocyclic substitutions are separated from the π system as appropriate. As for pyridines, the ring atoms and ring substitutents are treated separately for the *ortho*, *meta*, and *para* positions because of their differential resonance effects. These rules are exemplified by the case of *N*-methyl-1-naphthylamine:



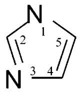
	level 0						level 1					
	exocyc.			ortho	meta	para	exocyc.	ortho		meta		para
	H	C.3	C.ar	C.ar	C.ar	C.ar	H	H	C.ar	H	C.ar	H
count	1	1	1	2	2	1	3	1	1	1	1	1

	level 2						level 3			
		ortho		meta						
	exocyc.	H	C.ar	H	C.ar	para	exocyc.	ortho H	meta H	para
count		1	1	1	1			1	1	

The aniline subset consists of 99 molecules, giving model statistics of

$$r^2 = 0.98, \mathrm{SE} = 0.309, F = 374, k = 88, N = 99$$

(3) Imidazoles. Imidazoles and benzimidazoles are best treated as a separate class in this method. One complication is that the two nitrogens are equivalent when the imidazole is not substituted. The proton is shared roughly equally between the two nitrogens. When the imidazole is substituted at either nitrogen, however, the two nitrogen atoms are no longer equivalent. Our best results for this class were obtained by mapping substituents into three different zeroth level nodes: one for those at the C2 carbon, another for those at N1 or N3, and a third for those at the C4 or C5 carbons. This is consistent with the similarity between the chemical characteristics of C4 and C5 and their differences from those of C2.



Imidazole

Unlike pyridine or aniline, where heteroatoms could appear in the aromatic ring, the atom and bond arrangements for imidazoles are fixed. Other azoles are not appreciably basic, so there is no need to map the ring itself onto the connectivity tree. Instead the complete imidazole ring becomes the ionizing center, filling the zeroth level. The first level takes the proximal ring substituents into consideration, which is consistent with the treatment of pyridines and anilines.

The 49 imidazoles and benzimidazoles produced a model with excellent statistics:

$$r^2 = 0.99, \mathrm{SE} = 0.158, F = 566, k = 50, N = 49$$

(4) Alkylamines. Nitrogenous bases not belonging to any of the aforementioned three classes were assigned to this class. Most were simple primary, secondary, or tertiary aliphatic amines. The protonatable nitrogen itself defines the zeroth level, with the atoms/groups directly connected to it taken as the first level. Atoms bonded to those in the first level define the second level and so on.

Some cyclic bases needed to be assigned separate pK^0's to obtain accurate predictions. These include morpholines, piperazines, and pyrrolidines. Piperidines, on the other hand, did not need to be treated as a separate subclass.

The six ring atoms in morpholine and piperazine were defined with respect to the protonated nitrogen of morpholine and piperazine, respectively, and hence were omitted from the corresponding connectivity trees. Ring substituents were combined with the rest of the alkylamines because they share the same mechanism of field effect through σ bonds. Care was taken to ensure that mappings for different positions around each ring were analogous to those applied to the noncyclic amines. Atoms or groups directly bonded to geminal carbons were assigned to the second level (the ring carbon itself is in the first level but its effect is subsumed in the corresponding indicator variable), and substituents on the vicinal carbons were assigned to the third level. Substituents further removed from the basic center were treated analogously.

Guanidino and amidino groups are also special cases, because the positive charge is shared among the nitrogen and carbon atoms in the substructure. Therefore the ionizing centers were treated as a unit and comprise the zeroth level. The atoms directly bonded to the three nitrogens in guanidino or to the two nitrogens and the carbon in amidino contributes to the first level. Those connected to the first level compose the second level and so on.

2,3-Pyrrolines behave in a different way from other vinylamines because of the unique geometric constraints inherent in the basic nitrogen being part of a five-membered ring. The nitrogen is directly bonded to an sp^2 carbon and so shares some of the sp^2 character, which makes it less basic. This effect is not as strong as for vinylamines because of the fixed geometry at the basic nitrogen forced by ring closure. This necessitated introduction of a specific indicator variable for 2,3-pyrrolines.

The pK_a's of this class are quite sensitive to additional substitution at the nitrogen atom. When the nitrogen bears an exocyclic alkyl group, it generally becomes more basic. Most of this effect can be attributed to the first exocyclic heavy atom, while further extension of the substitution has little effect. This was evident in the negligible change in statistics obtained when N-substituent connectivity was taken fully into account (r^2 = 0.95, SE = 0.342) vs the statistics obtained when only the first substituent atom was considered (r^2 = 0.95, SE = 0.341). Hence the exocyclic branch off

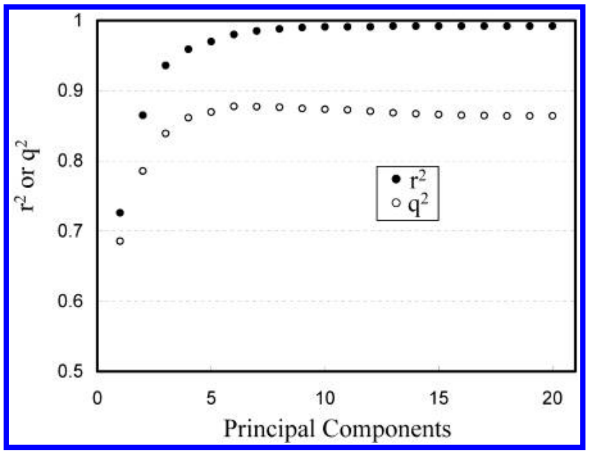
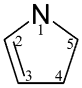


Figure 4. Plot of q^2 and r^2 for 412 molecules in the base data set.

the nitrogen was only accounted for at the first level, and the connectivity was not mapped any farther out. Substituents elsewhere in the molecule were mapped normally.



2,3-pyrroline

The 166 alkylamines in the data set gave a PLS model with the following statistics:

$$r^2 = 0.95, \text{SE} = 0.341, F = 303, k = 65, N = 166$$

(5) Complete Set for Bases. The data set contained 412 bases, and 290 columns went into the final PLS model. The statistics for the entire base set are

$$r^2 = 0.99, \text{SE} = 0.298, F = 43064, k = 290, N = 412; q^2 = 0.87, \text{SE}_x = 1.12$$

where q^2 and SE_x are cross-validated correlation coefficient and cross-validated standard error, respectively.

The internal predictive power of the model was evaluated by the same leave-some-out approach described above for acids. Ten cross-validation groups were used, and five different random ten-way partitions were evaluated. The average q^2 values obtained are plotted in Figure 4 as a function of the number of principal components included in the models, as is the correlation coefficient r^2. The q^2 is maximal at 6−7 components and then drops slightly as more principal components are included. The correlation coefficient (r^2) has a steeper initial slope than q^2 and levels off at a value of 0.99 near 10 components.

Figure 5 shows the predicted pK_a's as a function of the experimentally determined values for the model generated using 10 principal components. The slope of the regression line is 1.000, and the intercept is very near zero (−0.001). The ideal slope and the intercept indicate that the underlying relationship is linear and that the pK_a predictions are unbiased with respect to chemical class and to position in the pK_a range.

The distribution of residuals is depicted in Figure 6. The biggest error is observed for cyclopropylamine, for which the predicted pK_a is 10.6 and the experimental value is 9.1, a discrepancy of 1.5 log units. This probably reflects the unusual hybridization state of the cyclopropyl carbon, which is not adequately accounted for by the atom typing used here. This can, of course, be compensated for as a special case, and it may well be worth doing so if a data set is encountered which includes a variety of cyclopropylamine derivatives. However, we want to minimize the number of such special case corrections so as to maintain the generality of the model. In addition, the predicted and the measured pK_a's are relatively high, so that the protonation state will not be ambiguous in physiological settings (pH 6−9). The next most poorly predicted molecule is 1,4,5,6-tetrahydro-1,2-dimethylpyridine, whose pK_a (11.38) was underpredicted by 1.2 log units. Here the experimental value seems suspiciously high. Therefore no special treatment was applied to this molecule. The third molecule which displays comparatively high error is lysine, with an overprediction of experimental pK_a (9.06) by 1.09 log units. The rest of the errors are all less than one log unit, and for most the predicted pK_a falls within 0.5 log unit of the tabulated value (375 molecules or 91%).

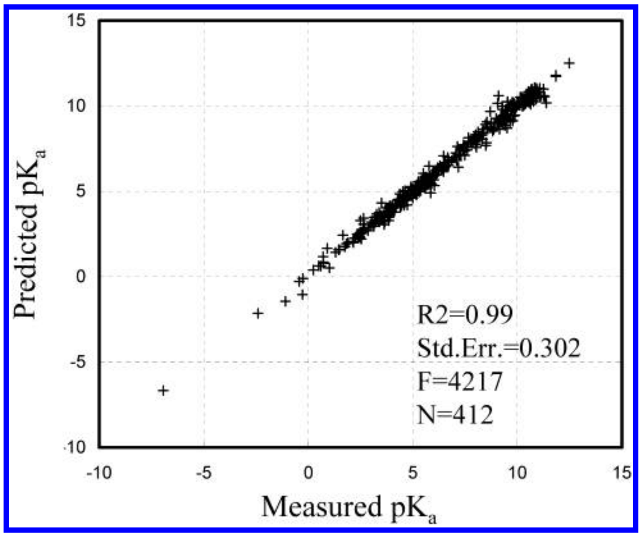


Figure 5. Predicted vs measured pK_a values for base data set.

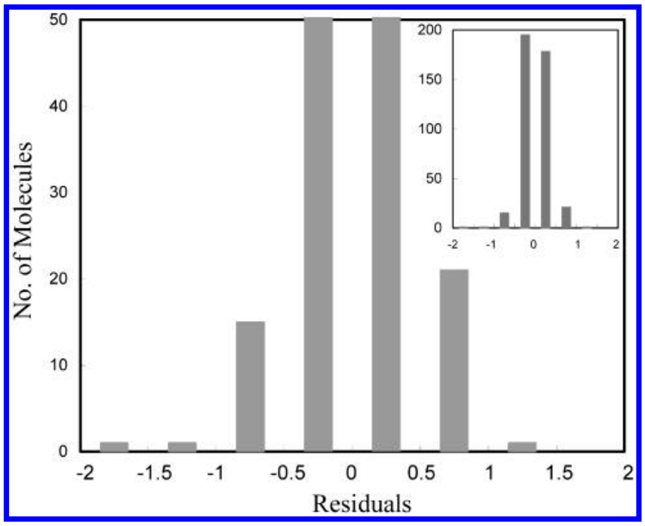


Figure 6. Distribution of residuals for bases data set. The two highest peaks are 195 and 178, respectively, but are truncated at 50. The upper-right panel is the full-scale plot.

PREDICTIONS

For comparison purposes, predictions for the same external test set used in the previous study[22] were generated using

Table 1. pK_a Predictions on Perrin's Data Set[a]

name	Perrin	model	measured
bis(2-chloroethyl)(2-methoxyethyl)amine	5.10	(6.91)	5.45
1-(4′-hydroxycyclohexyl)−2-(isopropylamino)ethanol	9.99	10.03	10.23
2-aminocycloheptanol	9.67	9.84	9.25
N,*N*-dimethyl-2-butyn-1-amine	∼8.1	(10.17)	8.28
5-chloro-3-methyl-3-azapentanol	7.1	(9.52)	7.48
2-acetylbutanedioic acid	3.15	2.35	2.86
2-(methylamino)acetamide	8.43	8.11	8.31
2-(dimethylamino)ethyl acetate	8.26	8.60	8.35
2,3-dihydroxy-2-hydroxymethylpropanoic acid	3.01	3.85	3.29
1,8-diamino-3,6-dithiaoctane	9.06	9.26	9.47
4-morpholino-2,2-diphenylpentanenitrile	6.38	(7.45)	6.05
benzenehexol	8.31	9.28	9.0
picric acid	0.91	1.18	0.33
2,6-dichloro-1,4-benzenediol	6.82	7.60	7.30
4-bromo-1,2-benzenedicarboxylic acid	2.86	3.26	2.5
4-hydroxy-3,5-dimethoxybenzoic acid	4.36	4.54	4.34
3-iodo-4-methylthioaniline	3.34	3.29	3.44
4-bromo-3-nitroaniline	1.82	1.78	1.80
3-bromo-5-methoxypyridine	2.30	3.19	2.60
4-aminopyridazine	5.31	6.76	6.65
4-amino-6-chloropyrimidine	1.41	1.68	2.10
4-nitrothiophen-2-carboxylic acid	2.70	2.58	2.68
4-bromopyrrol-2-carboxylic acid	4.05	4.22	4.06
furan-2,4-dicarboxylic acid	2.77	2.22	2.63
pyrazole-3-carboxlylic acid	3.98	3.77	3.74

[a] Predictions in parentheses are partial due to missing parameters.

the enhanced models described here. These compounds are collected from Perrin's book[19] and are external to the training set. Table 1 summarizes the pK_a values predicted by the new models, together with Perrin's predictions and the experimental results.

Most molecules in the test set contain multiple acidic/basic centers that could become ionized at different pH levels. Therefore it proved necessary to rank the relative strength of the different ionizable sites within each molecule. This was accomplished by making pK_a predictions for each potential site while assuming that the rest of the molecule is neutral. By comparing the values the most acidic/basic center was identified and assumed to correspond to the experimental measurement. It is gratifying to see that the ionizing strengths are consistently predicted in the correct order for test molecules, extending the successful performance of the initial program.[22]

Partial prediction was obtained for four molecules in the test set (in parentheses in Table 1) because of missing atomic parameters. This did not occur, of course, when different chemical classes were combined using the more generic atom typing rules used earlier.[22] But categorization of compounds and calibrating atom and group contributions to pK_a's separately for specific families of chemical structures inevitably leads to some molecules falling "outside of the box". This can be interpreted as a necessary tradeoff between generality and precision. Further expanding the training set will undoubtedly reduce the number of missing parameters and will be a focus of future development.

One may also question, however, whether failing to make predictions for such unusual compounds is a particularly bad thing. No predictions were made for (2-chloroethyl)(2-methoxyethyl)amine and 5-chloro-3-methyl-3-azapentanol because the training set contained no alkylamine examples of Cl vicinal to the basic nitrogen. In fact, the stability of the free amines is dubious given the ability of nitrogen to act as a nucleophile, thereby cyclizing at the β-carbon by displacement of chloride ion. This would explain the absence of such fragments in the training set. The pK_a of *N*,*N*-dimethyl-2-butyn-1-amine was predicted (10.17) to be higher than its real value (8.28) because the contribution of acetylene was neglected, but this was still closer to the mark than the prediction for the original implementation (10.44[22]). Given the strong electron withdrawing effect of the sp^1 hybridized carbons, especially positioned in the close neighborhood of the protonating center, it is reasonable to expect that the prediction would come closer to the actual pK_a had such acetylenic substitution been adequately parametrized. 4-Morpholino-2,2-diphenylpentanenitrile was also overpredicted (7.45), evidently because due account was not taken of the somewhat surprising electron-withdrawing strength of the cyano group. Hence, given the nature of the missing parameters, we can conclude that reasonably good predictions can be obtained even for molecules containing undercharacterized fragments.

Our enhanced models outperform Perrin's methods overall on this test set of molecules in terms of both higher predicted r^2 (0.99) and a lower predicted standard error (0.40).[22] The largest error was found for picric acid, the pK_a of which was overestimated by 0.85 units. This compares favorably with Perrin's largest error, which is an underestimate of 1.34 units for 4-amino-6-chloropyrimidine. Furthermore, the results shown here represent significant improvements over the initial version of the method.[22] In addition, the predictions display a more balanced error distribution, i.e., are less biased.

DISCUSSION

The pK_a prediction model described here was developed using a count vector descriptor drawn from a molecular connectivity tree rooted at the ionizable center, with parameters derived by application of partial least squares projection onto latent structures (PLS).

Similar treatments of chemical structures have been reported for other QSAR/QSPR applications.[30,31] In the description of multilevel neighborhoods of atoms the elemental atom types were specified, and some elements of the same or neighboring groups were combined.[29] Such descriptors were used to predict boiling point and mutagenicity. Another algorithm combined atom types with bond types to more fully account for hybridization states.[30] Euclidean distances, as opposed to the topological distances used here, have also been incorporated into molecular features count vectors. Such descriptors have performed well in several cases when compared to CoMFA and EVA analyses.

We also evaluated related descriptors based on connectivity trees but using properties other than atom type (augmented with special substituent groups) for prediction of pK_a. These included atomic charges calculated by different methods (specifically Gasteiger, Gasteiger-Huckel, and Mulliken charges from the AM1 Hamiltonian) and atomic polarizabilities. We also tried augmenting atom types with bond type information. The atomic charges and polarizabilities were binned in several different ways, and combinations of two or more descriptor types were also considered. None of these alternative approaches produced models as good as those derived from SYBYL atom types augmented by the special group definitions and indicator variables described above. Including bond types failed to improve model statistics significantly probably because the bond types are in large part implicitly accounted for in the atom typing e.g. the bond between two aromatic carbons is usually aromatic—except in the case of biaryls. Hence including the bond types roughly doubled the number of independent variables involved in the models while adding mostly redundant information.

The connectivity trees used can, of course, be extended to include every atom in the molecule, but considering only five connectivity levels produced quite satisfactory models. In fact, the quality of the models obtained plateaued at that level. This is consistent with the intuitively appealing notion that the ionization state of a specific group is determined mostly by nearby atoms, with substituent influence falling off rapidly as the number of intervening bonds increases. Effects of more remote atoms and groups are negligible, especially when the intervening bonds are aliphatic. Of course, conformational constraints may bring topologically distant atoms into close enough proximity to influence pK_a. Such an effect is clearly evident for maleic (pK_{a1}: 1.91; pK_{a2}: 6.33) and fumaric acids (pK_{a1}: 3.1; pK_{a2}: 4.6), for example. It cannot be properly accounted for by the methodology described here.

In this second phase tremendous efforts were made to expand the training set, especially with respect to those chemical classes poorly represented in the original work; numerous corrections were applied to the data; and the methodology was modified to better capture the different contributing factors that are intrinsic to separate chemical classes. It is gratifying to see that remarkable improvements compared to the previous models have been achieved.

The number of principal components was limited to one-fifth to one-sixth the number of contributing observations to minimize the risk of overfitting the data. Extra subclasses and indicator variables were carefully scrutinized and evaluated before being added; only independent variables appearing to be physicochemically relevant and which contributed significantly to the PLS models were included. Note that the "true" external predictive errors are expected to fall somewhere between the standard errors for each full model and the SE_x found for the corresponding reduced cross-validation models. The latter are, after all, based on 20% fewer observations. The accuracy of the predictions obtained for the structurally diverse molecules making up the external test set argues against any suspicion of chance correlation or overfitting. It should be noted, however, that several of the chemical classes represented in the training set (i.e., aliphatic alcohols and thiols, acidic carbons and nitrogens, and imidazoles, etc.) are not present in the external test set, so predictivities for those aspects of the models are less well characterized. Nevertheless the models produced cover all of the common acid (carboxylic acids, benzoic acids, phenols, etc.) and base (alkylamines, anilines, pyridines, etc.) classes.

CONCLUSIONS

We report here on the second phase of development for pK_a prediction models based on molecular tree structured fingerprints. The methodology has been extended to cover a broad range of chemical classes. Compared to the first phase, significant improvements and enhancements have been achieved in terms of wider representation of chemical spaces, better fit to the training set, and more accurate prediction on the test set. The method is expected to serve as a useful tool to quickly estimate pK_a values based on molecular connectivity once a suitable "wrapper" is created to automatically account for interactions between multiple potentially ionizable sites present in the same molecule, thereby converting atomic pK_a's into more pharmacologically relevant molecular ones.

REFERENCES AND NOTES

(1) Oprea, T. I.; Marshall, G. R. Receptor-Based Prediction of Binding Affinities. In *3D QSAR in Drug Design, Vol. 2*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordercht, 1998; pp 35−61.

(2) *CLOGP, LogP Calculation Algorithm;* Pomona College Medicinal Chemistry Project: Claremont, CA 91711, 1989.

(3) Csizmadia, F.; Szegezdi, J.; Tsantili-Kakoulidou, A.; Panderi, I.; Darvas, F. Prediction of Distribution Coefficient from Structure. 1. Estimation Method. *J. Pharm. Sci.* **1997**, *86*, 865−871.

(4) Tsantili-Kakoulidou, A.; Panderi, I.; Csizmadia, F.; Darvas, F. Prediction of Distribution Coefficient from Structure 2. Validation of Prolog D, an Expert System. *J. Pharm. Sci.* **1997**, *86*, 1173−1179.

(5) Shapley, W. A.; Bacskay, G. B.; Warr, G. G. Ab initio Quantum Chemical Studies of the pK_a's of Hydroxybenzoic Acids in Aqueous Solution with Special Reference to the Hydrophobicity of Hydroxybenzoates and Their Binding to Surfactants. *J. Phys. Chem. B* **1998**, *102*, 1938−1944.

(6) Schueuermann, 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*, 6707−6712.

(7) da Silva, C. O.; da 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.

(8) Tran, N. L.; Colvin, M. E. The Prediction of Biochemical Acid Dissociation Constants Using First Principles Quantum Chemical Simulations. *Theochem* **2000**, *532*, 127−137.

(9) Citra, M. J. Estimating the pK_a of Phenols, Carboxylic Acids and Alcohols from Semiempirical Quantum Chemical Methods. *Chemosphere* **1999**, *38*, 191−206.

(10) Chen, I.-J.; MacKerell, A. D. Computation of the Influence of Chemical Substitution on the pK_a of Pyridine using Semiempirical and ab Initio Methods. *Theor. Chem. Acc.* **2000**, *103*, 483−494.

(11) Bashford, D. Karplus, M. pK_a's of Ionizable Groups in Proteins: Atomic Detail from a Continuum Electrostatic Model. *Biochemistry* **1990**, *29*, 10219−10225.

(12) Oberoi, H.; Allewell, N. M. Multigrid Solution of the Nonlinear Poisson−Boltzmann Equation and Calculation of Titration Curves. *Biophys. J.* **1993**, *65*, 48−55.

(13) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. Prediction of pH-dependent Properties of Proteins. *J. Mol. Biol.* **1994**, *238*, 415−436.

(14) Sham, Y. Y.; Chu, Z. T.; Warshel, A. Consistent Calculations of pK_a's of Ionizable Residues in Proteins: Semi-Microscopic and Microscopic Approaches. *J. Phys. Chem. B* **1997**, *101*, 4458−4472.

(15) Warwicker, J. Simplified Methods for pK_a and Acid pH-Dependent Stability Estimation in Proteins: Removing Dielectric and Counterion Boundaries. *Prot. Sci.* **1999**, *8*, 418−425.

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

(17) 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.

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

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

(20) CompuDrug NA, Inc. p$K_a$lc version 3.1, 1996.

(21) ACD Inc. ACD/pK_a Version 1.0, 1997.

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

(23) Tajkhorshid, E.; Paizs, B.; Suhai. Role of Isomerization Barriers in the pK_a Control of the Retinal Schiff Base: A Density Functional Study. *J. Phys. Chem. B* **1999**, *103*, 4518−4527.

(24) SYBYL is distributed by Tripos, Inc., St. Louis MO 63144; http://www.tripos.com.

(25) Wold, S.; Sjostrom, M. In *Chemometrics: Theory and Application*; Kowalski, B. R., Ed.; American Chemical Society: Washington, DC, 1977; p 243.

(26) *Lange's Handbook of Chemistry*; 13th ed.; Dean, J. A., Ed.; McGraw-Hill, Inc.: 1985; Tables 5−8.

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

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

(29) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71−79.

(30) Filimonov, D.; Poroikov, V.; Borodina, Y.; Gloriozova, T. Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 666−670.

(31) Baumann, K. An Alignment-Independent Versatile Structure Descriptor for QSAR and QSPR Based on the Distribution of Molecular Features. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 26−35.

CI020386S