Novel Methods for the Prediction of logP, pK_a, and logD

Li Xing*,† and Robert C. Glen‡

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

Received September 7, 2001

Novel methods for predicting logP, pK_a, and logD values have been developed using data sets (592 molecules for logP and 1029 for pK_a) containing a wide range of molecular structures. An equation with three molecular properties (polarizability and partial atomic charges on nitrogen and oxygen) correlates highly with logP ($r^2 = 0.89$). The pK_as are estimated for both acids and bases using a novel tree structured fingerprint describing the ionizing centers. The new models have been compared with existing models and also experimental measurements on test sets of common organic compounds and pharmaceutical molecules.

INTRODUCTION

The bioavailability of a drug and its access to the therapeutic target are important considerations in rational drug design. Before the drug can elicit an effect, for example if it is orally administered, it usually has to pass through a series of barriers (e.g. biological membranes) either by passive diffusion and/or carrier-mediated uptake. Depending on the route of the administration of the drug and the location of the target site, the pH of the environments that the compound is exposed to may vary considerably. Some examples of physiological pH values are as follows: stomach 2.0, kidneys 4.2, small intestine (food 5.0; fasted 6.8), duodenal mucus 5.5, plasma 7.4. In this context, the affinity of the drug molecule for the target of interest and its ability to partition into a lipophilic environment at different pH values has to be quantified for a proper prediction of its ability to interact with the biological target and hence to be efficacious.

For many years the 1-octanol/water partition coefficient (logP) has been used as a measure of lipophilicity. Pioneering work by Hansch and Leo[1] has led to the use of logP in quantitative structure−activity relationship (QSAR) methods e.g. as a general description of cell permeability. LogP has since become a standard property determined for potential drug molecules e.g. Lipinski's "rule of 5".[2]

LogP refers to the neutral state of molecules. In the presence of a basic or acidic group the ionization of a molecule provides an additional factor to consider, since the partition then becomes pH dependent. The pH dependent distribution coefficient, logD, is related to logP through the ionization constant, pK_a. Many drug molecules contain ionizable groups and hence partition across cell membranes, through pores and via active transport mechanisms in a pK_a dependent fashion.

Since it is not always convenient or practical to perform experimental measurements, it is useful to develop easy-to-use and accurate models by which logP, pK_a, and ultimately logD values can be rapidly predicted. There is also an increasing need for reliable estimates of these physicochemical parameters for new compounds not yet synthesized, particularly for drug discovery. As an example of the use of logD, the partition coefficient (together with other physicochemical parameters) was used in an analysis of the effective permeability in the human jejunum (in vivo).[3] Although statistically good models were achieved using pH-independent partition coefficient logP, the pH-dependent distribution coefficient logD convincingly yielded the best models.

There have been a number of attempts at predicting logP using different algorithms. Fragment methods estimate logP using the additive contributions of functional groups and fragments as well as their interactions with each other. Both the Hansch/Leo and Rekker approach entail correction factors though the former is based on the principles of constructionism and the latter on reductionist principles.[4,5] Although accurate for molecules that have parametrized fragments, these are often not always available for the molecules of interest. Comparative evaluation of different fragmental approaches revealed that the goal of reliable logP calculation for many diverse structural types has not yet been fully realized.[6,7] A number of methods based on atomic and group contributions to molecular logP have been developed.[8−10] These approaches employ multiple regression equations to establish models based on a training set. Molecular orbital methods and neural network analysis have also recently been applied.[11,12] Another route to logP is the direct calculation of the free energy change for transferring a solute from aqueous to organic solution from a thermodynamic treatment.[13,14] These methods are usually generalizable to other two-phase systems.

The calculation of pK_a's has been attempted by a number of methods. In 1981 Perrin et al.[15] published a book on pK_a prediction which is widely used. Depending on the nature of the chemical structures, different algorithms were developed to calculate the pK_a's. These include the acid-strengthening and/or base-weakening factors of the substituted aliphatic acids and bases; the Taft equation; the Hammett equation for phenols, aromatic carboxylic acid and aromatic amines; the Hammett and Taft equations for the heteroaromatic acids and bases; and the extension of the Hammett and Taft equations to heterocycles. Fragment

* Corresponding author phone: (636)737-5466; fax: (636)737-7425; 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/ci010315d CCC: $22.00
Published on Web 04/04/2002

methods have proved very useful and are available as commercial systems.[16] Ab initio quantum mechanics calculations have been used extensively[17] as well as semiempirical quantum mechanics.[18] A number of methods have also been developed for prediction of pK_as of amino acid residues in proteins[19] where the environmental effects are particularly important and difficult to estimate.

The methods that we have developed here are based on an informatics approach in which a description of the system is parametrized by use of experimental data. A complication is that in practice not only neutral molecules but also ion pairs may partition. The charged species may pair with a reagent ion or even, in certain cases, itself. This leads to difficulties in the experimental determination as well as in prediction. Both the logP and the logD values may be affected if one or more of the charged species partitions. This effect is not considered here; however, work is in progress to include this phenomenon in the present algorithms.

METHODS

1. logP Model. There is mounting evidence that molecular size and hydrogen bonding ability can account for a major part of the variance in the partitioning of an organic molecule into different phases. A simplification we have taken here is to describe molecular size and dispersion interactions with solvent using molecular polarizability, which is estimated here by an empirical method described by Glen.[20] This avoids complicated and often ambiguous three-dimensional determinations of polarizability using e.g. more expensive quantum mechanics approaches. The algorithm is based on Slater's rules[21] for the calculation of effective atomic nuclear shielding constants. The calculated molecular polarizabilities of a series of organic molecules correlated well ($r^2 = 0.96$) with experimental measurements.[20]

Oxygen and nitrogen atoms are generally involved in hydrogen-bonding interactions with solvent. To take this into account, the partial atomic charges on these heteroatoms were included in the calculation of logP. The empirical Gasteiger-Huckel method implemented in Sybyl[22] was used for rapid charge estimation. Additional properties e.g. sulfur or halogen charges (or charges on other atoms), molecular dipole moment, ovality, etc. did not increase the reliability of the models generated and indeed led to much more complex models. It was thus decided to use simply the three parameters described: polarizability and partial charges on oxygen and nitrogen atoms.

Data Set. Measured logP values in 1-octanol/water for 449 molecules are available from the *CRC Handbook*,[23] and all refer to a nominal temperature of 25 °C.

The data set was further expanded to include 157 logP* values[24] (of which 12 had already been entered into the database). All the duplicates displayed identical logP values between *CRC Handbook*[23] and logP* except for thiophenol, whose logP is 3.5 according to ref 23 but 2.52 given by logP*. The value from logP* was chosen for thiophenol.

Two pairs of configurational isomers were identified in the training set: 2-butene (trans: 2.31; cis: 2.33) and 1,2-dichloroethylene (trans: 1.93; cis: 1.86). Since the current method does not take into account stereochemistry, only one entry was adopted from a *trans*/*cis* pair with the average logP of the isomers assigned to it.

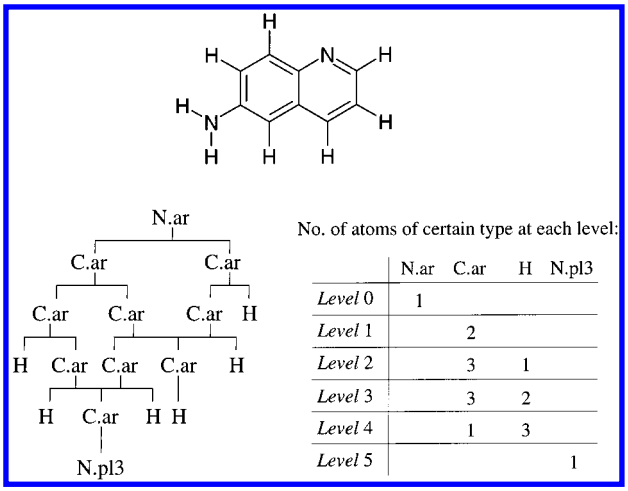


Figure 1. Construction of the hierarchical tree from one example, 6-amino quinoline.

After removing duplicate structures the logP data set consists of 592 distinct molecules in total.

2. pK_a Models. Our goal is to develop an accurate, simple and fast method, which does not involve 3D conformational determinations or rather more expensive quantum mechanics or dynamics calculations. It is also convenient not to need explicit fragment values, but rather having fragments as a precomputed and general metalayer embedded in the algorithm. Based on the hypothesis that the ionization state of a particular group is dependent upon its subenvironments constituted by its neighboring atoms and bonds, a hierarchical tree was constructed from the ionizing atom 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 and so forth.

Taking 6-aminoquinoline as an example (Figure 1), the root atom is an aromatic nitrogen (level zero), with two aromatic carbons connected to it, which compose the first level. Bonded to the 1-carbon of ring fusion at the first level are two aromatic carbons, while to the 2-carbon are an aromatic carbon and a hydrogen. These atoms then form the second level. The hierarchical tree up to five levels is demonstrated in Figure 1. It is noted that although both 4- and 6-carbons could be traversed by two different routes depicted by their parent atoms, they only contribute once to the level in which they reside. As shown in the right panel of Figure 1, the atoms of the same type for each level were then binned together (summed). After evaluating many atomic representations, the Sybyl force-field atom types were found to be the most useful in describing the environment of an ionizing center in this algorithm. These include 24 atom types in their different hybridization states, covering most organic molecules of biological interest (Table 1).

Based on this description, a string is formed which composes of the accumulated number of atoms of each type at each level, originating from the root. Each string is made up of the occurrences and positions of each atom type in the neighborhood of the center and thus fingerprints the characteristics of the ionizing center.

In addition to the primary atom types, certain chemical groups were treated explicitly as incorporation of atoms into

Table 1. Description of Atom Types and Their Hybridization States in Sybyl Forcefield

code	definition	code	definition
C.3	carbon sp^3	O.3	oxygen sp^3
C.2	carbon sp^2	O.2	oxygen sp^2
C.1	carbon sp	O.co2	oxygen in carboxylic and phosphoric acid
C.ar	carbon aromatic	S.3	sulfur sp^3
C.cat	carboncation (+)	S.2	sulfur sp^2
N.3	nitrogen sp^3	S.o	sulfoxide sulfur
N.2	nitrogen sp^2	s.o2	sulfone sulfur
N.1	nitrogen sp	H	hydrogen
N.ar	nitrogen aromatic	F	fluorine
N.am	nitrogen amide	Cl	chlorine
N.pl3	nitrogen trigonal planar	Br	bromine
N.4	nitrogen sp^3 positively charged	I	iodine

molecules was not adequately described (in particular, when delocalized π-electron systems were involved). These included nitro, nitroso, cyano, carbonyl, sulfonyl, sulfinyl, hydroxyl, and amino groups. The amino group was further divided into two subgroups depending on the hybridization state of the nitrogen atom, i.e., sp^2 or sp^3. This differentiated, for example, aniline-type nitrogens from alkyl nitrogens. The rationale was supported by the fact that both the cross-validated q^2 and the correlation coefficient r^2 significantly improved upon the inclusion of these functional groups in addition to the individual atoms.

Therefore at each level there are 33 (24 atom types and nine group types) atom/group type bins. The tree generation always starts with an atom type at the point of ionization. The maximum number of bins for five levels would be $[(5 \times 33) + 24] = 189$. Some particular atom/group types obviously may not occur at certain levels in the training sample (e.g. for bases there are no oxygen or sulfur atoms at the first level, the level next to the protonated nitrogen) hence their contributions are simply left out.

Using the strings described above for each of the pK_as in the training set, the implementation of the partial least squares (PLS) algorithm in Sybyl[22] with cross-validation[25] was used to create a series of predictive models.

Data Set. The pK_a data were abstracted from Lange's Handbook of Chemistry.[26] They are measured at 25 °C in water. After removing duplicate entries the data set was carefully reviewed, and five molecules were identified as having unexpected or conflicting values (reported pK_a in parantheses): 2,6-di-*tert*-butylpyridine (3.58), 2-aminofluorene (10.34), 2-ethoxyethylamine (6.26), *N*-methylethylamine (4.23), and *N*,*N*,*N*′,*N*′-tetramethylethylenediamine (6.35). Those of dubious pK_a values (typically with other literature values in conflict) were then removed, resulting in the training sets of 384 bases and 645 acids. Their pK_a values cover broad ranges, from −6.94 to 12.48 for bases and from −2.8 to 14.22 for acids, respectively.

3. Calculation of logD. LogD may be calculated from the predicted logP and pK_a of the singly ionized species at certain pH using the equations[4]

$$\log D_{(pH)} = \log P - \log[1 + 10^{(pH - pK_a)}]$$

for acids and

$$\log D_{(pH)} = \log P - \log[1 + 10^{(pK_a - pH)}]$$

for bases.

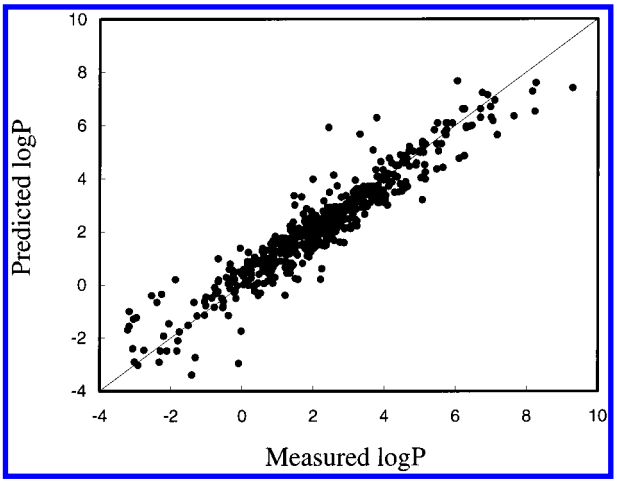


Figure 2. Predicted vs measured logP values for the training set of 592 molecules.

RESULTS AND DISCUSSION

LogP. a. The Model. Multiple linear regression suggested that the sum of the squared charges on nitrogen and oxygen atoms contribute more effectively to the model than the sum of the absolute charges. This nonlinear character of logP models have previously been reported.[27,28] The equation of the least-squares fit of 592 logP values with three adjustable parameters was

$$\text{LogP} = 0.287(0.066) + 0.190(0.004) * \alpha - 14.862(0.545) * q^2\_\text{N} - 8.445(0.216) * q^2\_\text{O}$$

where α is the molecular polarizability and q^2_N and q^2_O are the total squared partial charges on nitrogen and oxygen atoms, respectively. This calculation resulted in good agreement with experiment, with an r^2 value of 0.893 and a standard error (SE) of 0.653:

$$R^2 = 0.893,\ r = 0.945,\ \text{SE} = 0.653,\ F = 1631,\ n = 592$$

The cross-validated q^2 was assessed by randomly dividing the data set into 10 groups and each group being predicted by the model derived from the remaining nine groups. This process was then repeated 10 times resulting in an averaged q^2 of 0.877. The individual q^2s received from each random grouping only fluctuate negligibly around the average, indicating that the model is stable over the changing compositions of training and test sets.

Calculated against observed logP is plotted for the training set (Figure 2). The slope and intercept of the best fit are 1.00 and 0.00.

Molecular polarizability is more significant in the equation than the atomic charges, as implied by the regression analysis. The bigger the polarizability, the more hydrophobic the molecule is predicted to be—molecules which require a bigger cavity are predisposed to move into the 1-octanol layer. This indicates that the most important factor of the partitioning is the relative energy required in the creation of a cavity in water or 1-octanol. Another theoretical approach also reported that in the systems involving water as the polar phase, the hydrophobic effect is always the driving force that governs the distribution process.[14] It is probably simplistic to argue this effect only in terms of loss of entropy for the system when the molecule is in the water layer as

dispersion energy may also play a role, which is perhaps why polarizability appears to be such a useful parameter here.

The two remaining parameters are derived from computed charge densities of nitrogen and oxygen atoms of the molecule, which is a crude measure of their ability to form hydrogen bonds with the solvent molecules. It is noted that the coefficients for q^2_N and q^2_O have the same sign and similar orders of magnitude, implying the similar interaction they characterize, specifically the electrostatic interactions between the solute and the solvent molecules. The difference in the values of the corresponding regression coefficients could derive from the different energetics of hydrogen bonding for nitrogen and oxygen atoms.

While molecular polarizability increases logP, the charge densities correlate negatively with logP. This is in agreement with other prediction models[29] as well as with the general principle that molecules containing more highly charged groups are hydrophilic.

It is reasonable to envisage that other descriptors that encode hydrogen-bonding propensity (and are available in a fragment or atom-based lookup table) may relate to the partitioning thus could replace the computed atomic charges and speed the algorithm. One of these descriptors is the polar surface area (PSA), and we investigated combining it with polarizability to model logP. To be consistent with the atomic charge calculations only the nitrogen and the oxygen atoms are considered as "polar atoms". The Connolly surface generated by MOLCAD[30] was used to estimate the solvent exposed areas of oxygen, nitrogen, and hydrogen atoms bonded to oxygen and nitrogen. In order for fast calculation Concord conformations were used. Though this is not the most accurate it correlates with the other more time-consuming methods.[31] The model displayed r^2 of 0.688 and standard deviation of 1.11, disappointingly not as good as the atomic charge models. The reason could be that PSA captures mainly the accessibility rather than the strength of polar interactions. Adding PSA into the polarizability and the atomic charge model again did not improve the model, implying that its contribution is probably already absorbed in the other descriptors: polarizability and the atomic charges.

As logP is calculated solely from molecular properties, there are no additive constants or correction factors employed. The performance of this property model is respectable given its simplicity. Here only three readily obtainable property parameters serve as the descriptors, thus it can be easily implemented and does not produce ambiguous results due to different fragmentations or different interpretations of complex correction rules.

On the other hand, since the method only matches atomic composition and does not consider 3D structures, long-range interactions or intramolecular hydrogen bonds are not accounted for. The method also does not make stereochemical distinctions at present, which results in diasterioisomers and cis/trans isomers having the same calculated logP. These issues will be addressed in future developments of the method.

b. Predictions. The predictive power of the model was evaluated by comparison with experimental data taken from the literature. Forty drug molecules of complex and diverse structural patterns composed of five pharmacological classes were used as an external and realistic test set:[7]

Table 2. Model Predicted, ClogP Calculated, and the Measured logP Values for 40 Drug Molecules[a]

name	ClogP	Δ	model	δ	measured
Class I Antiarrhythmics					
carocainide	2.184	0.80	0.086	−1.29	1.38
disopyramide	1.704	−0.88	3.797	1.22	2.58
ethmozine	3.145	0.17	3.172	0.19	2.98
lidocaine	1.954	−0.46	3.020	0.61	2.41
mexiletine	2.569	0.42	2.483	0.33	2.15
nicainoprol	1.402	−0.23	2.215	0.59	1.63
procainamide	1.423	0.54	0.617	−0.26	0.88
propafenone	3.486	−1.14	3.875	−0.76	4.63
quinidine	2.785	−0.66	3.414	−0.03	3.44
Class II β-Blockers					
acebutolol	1.702	−0.01	1.289	−0.42	1.71
alprenolol	2.652	−0.48	3.062	−0.07	3.13
atenolol	−0.109	−0.89	0.317	−0.46	0.78
bunitrolol	1.736	0.14	1.079	−0.52	1.60
bupranolol	3.100	0.39	3.161	0.45	2.71
metipranolol	2.545	−0.12	2.510	−0.15	2.66
metoprolol	1.196	−0.78	2.109	0.13	1.98
oxprenolol	1.892	−0.40	2.360	0.07	2.29
penbutolol	4.039	−0.11	4.400	0.25	4.15
pindolol	1.671	−0.08	1.483	−0.27	1.75
propranolol	2.750	−0.42	3.194	0.02	3.17
Phenothiazines					
alimemazine	5.369	0.56	5.394	0.58	4.81
chlorpromazine	5.693	0.35	5.277	−0.06	5.34
fluphenazine	4.907	0.12	4.772	−0.02	4.79
levomepronzine	5.090	0.23	5.010	0.15	4.86
perazine	5.394	1.39	5.048	1.05	4.00
perphenazine	4.485	0.02	4.655	0.19	4.47
promazine	4.900	0.26	4.950	0.31	4.64
promethazine	4.644	−0.11	5.005	0.26	4.75
thiethylperazine	6.178	0.77	6.475	1.07	5.41
thioridazine	6.227	0.24	7.134	1.14	5.99
trifluorperazine	6.272	1.20	5.492	0.42	5.07
trifluopromazine	5.667	0.32	5.393	0.04	5.35
Benzamides					
alizapride	3.083	1.29	1.549	−0.24	1.79
amisulpride	0.990	−0.11	0.076	−1.02	1.10
bromopride	2.407	−0.21	0.782	−1.84	2.62
sulpiride	1.114	0.69	0.006	−0.41	0.42
sultopride	1.927	0.87	1.608	0.55	1.06
tiapride	1.282	0.26	2.309	1.29	1.02
veralipride	0.620	−0.85	0.006	−1.46	1.47
Class III Antiarrhythmics					
sotalol	0.226	−0.36	0.763	0.17	0.59

[a] Calculated using version 4.0 of the ClogP algorithm and version 18 of its associated fragment database.

I. Class I antiarrhythmics
II. β-blockers
III. Potassium channel openers
IV. Neuroleptics
V. Class III antiarrhythmics

To provide a quantitative evaluation of the diversity of the test set, the mean Tanimoto index for this set of molecules was computed to be 0.75 using Tripos Unity[32] derived fingerprints, with 70% of the total population below the (close neighbor) value of 0.85. The experimental logP values range from 0.42 to 5.99. The measured and model-predicted values together with the ClogP calculations for comparison were summarized in Table 2 and further depicted in Figure 3.

The ClogP values are associated with an error code of zero, with the exception of ethmozine, which has an error

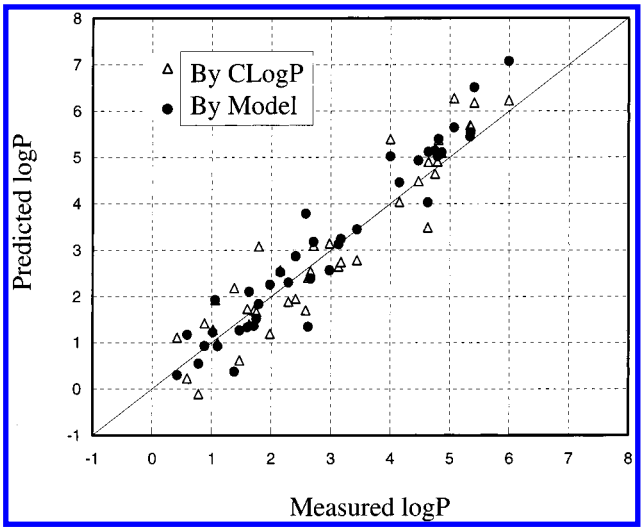


Figure 3. LogP values for 40 drug molecules predicted by the model and by the ClogP program vs the experimental values.

Table 3. Summary of Statistics of the logP Prediction on 40 Drug Molecules

					deviations	
	predicted r^2	SE	$\|\Delta\| > 0.5$	$\|\Delta\| > 1.0$	−	+
model	0.89	0.66	15	8	17	23
CLogP	0.89	0.61	15	4	18	22

code of 10, indicating a valid estimate for a difficult structure.

The predicted values from this model agree well with the experiment. The largest error in the prediction of logP is for bromopride whose predicted logP is 0.78, compared to the experimental value of 2.62. This is slightly larger than the largest error incurred by ClogP, i.e., 1.39 log units for perazine. Obviously the fragment that drives up the lipophilicity, the bromide attached to the phenyl ring in bromopride, was not sufficiently well explained by this model (possibly an underestimate of the polarizability of bromine).

Breaking down the external test set into therapeutic subsets makes it possible to identify classes of compounds which are particularly well modeled, giving some idea about the strengths and weaknesses of this approach in selection of drug-like molecules. In terms of standard error values, β-blockers stand out as the best modeled class. By far the worst class of molecules is Class I antiarrhythmics, with prediction error of about one log unit. Among this class, molecules containing the sulfonamide chemical group are mostly poorly predicted. The major reason for the failure on sulfonamide-containing molecules is its rare representation, with only two occurrences in the training set. The uncertainties in the atomic charge estimations may affect the model. When comparing the charges on sulfonamide oxygens calculated from the Gasteiger-Huckel method with those from AM1-derived Mulliken population analysis,[33] significant differences were noted. Interestingly, AM1-Mulliken charges together with polarizability yielded poorer models (r^2 = 0.643) than the empirical Gasteiger-Huckel charges.

The performance of the model and ClogP predictions on 40 drug molecules are summarized in Table 3. Overall, ClogP performs better than the present model on this set of drug molecules in terms of a slightly lower predictive error. ClogP also improves upon this method having fewer errors greater than one log unit. Of interest is that during the course of our study the ClogP version 4.0 was received whose predictive accuracy was greatly improved compared to the previous version 3.54. On the same set of drug molecules, the latter yielded 19 occurrences with prediction errors greater than 0.5 log unit, and eight greater than 1.0 log unit, compared to 15 and four, respectively, in the newer version. The underlying reason may be that the data set was expanded to include molecules similar to this set of drug molecules.

Table 4. Summary of Statistics of the pK_a Models for Acids and Bases

	n	PC	q^2	r^2	SE	*F*	*P*
acids	645	6	0.85	0.93	0.76	1384	0.0
bases	384	6	0.83	0.93	0.86	816	0.0

The method delivers reasonably accurate predictions of logP for a fairly wide range of complex molecules. Moreover, it is gratifying to see the accuracy obtained for molecules wholly unlike those used for training, such as the set of phenothiazines. The method does of course give predictions for the widest range of molecules given the simplicity and general utility of the algorithm and offers significant opportunities to be further developed and refined.

pK_a. a. The Model. The ionization models were developed using a combination of descriptors mapped onto the molecular tree constructed around the ionizable center using partial least squares (PLS) with cross-validation.

Various descriptors were mapped onto the hierarchical tree and evaluated for the prediction of pK_a. These included atomic charges, specifically Gasteiger, Gasteiger-Huckel, and Mulliken charges from the AM1 Hamiltonian; atomic polarizabilities; and the Sybyl atom and bond types. Also descriptor types were combined in different additive schemes to assess the significance of their contribution to the prediction of pK_a. However none of these trials gave models as good as those derived from the force-field atom/group types in terms of their statistical significance. It is interesting that adding the bond types did not improve the model significantly, albeit that the number of independent variables almost doubled. The reason is likely that the bond types in most cases are implicitly accounted for by the force field atom. Also bond types are more generalized than atom types with only four major types available, specifically single, double, triple, and aromatic types. Hence including the bond types would only serve to add more redundancy.

The creation of the connectivity tree could be as exhaustive as terminating at the leaf atoms on every branch for each organic molecule, yet five levels beyond the initial zeroth level was found to be sufficient. Indeed the correlation was seen to plateau beyond the fifth level. This implies that the ionization state of a specific group is mostly determined by its closest neighboring atoms/groups, with the effect attenuating as the chemical entities become more distanced from the ionizing center. For remote atoms, the effect becomes sufficiently negligible that in this algorithm their effects can be ignored. Of course conformational effects may bring distant atoms together and influence the result, and this is not taken into account here.

The statistics for the pK_a models are summarized in Table 4, with the q^2 obtained by averaging 10 cross-validation experiments as described in the logP section. The quality of the models was measured using primarily two statistical

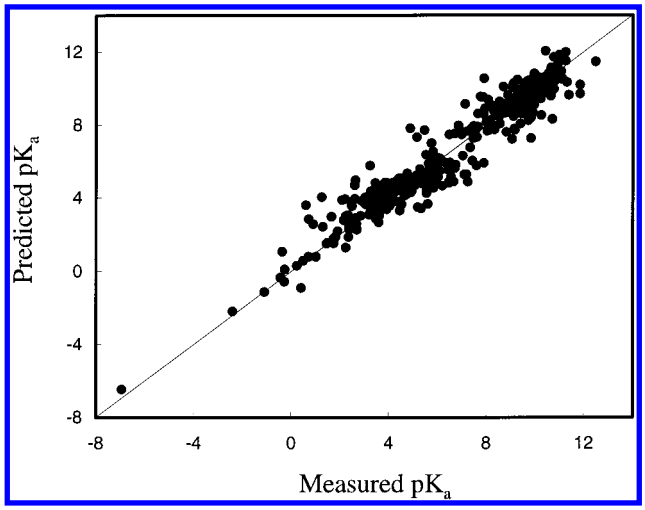


Figure 4. Predicted vs measured pK_a values for 384 molecules in the training set for bases.

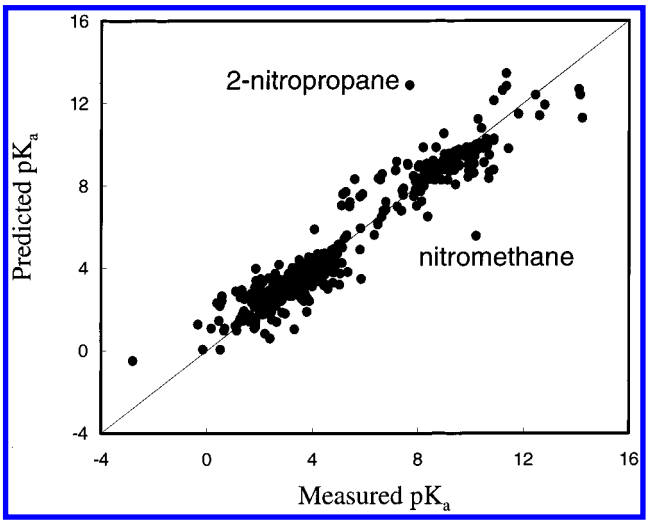


Figure 5. Predicted vs measured pK_a values for 645 molecules in the training set for acids.

parameters, r^2 and q^2. The q^2s for both the bases (0.82) and acids (0.85) models are well above 0.5, the threshold generally agreed as being statistically useful and significant. This indicates that both models are predictive (as does inspection of the measured/predictive plots), that is, their ability to extrapolate beyond the training set. The r^2s are prominently high, 0.92 for the bases and 0.93 for the acids, respectively. This indicates the models ability to interpolate within the range of pK_as in the training set.

The predicted vs the experimentally observed pK_a values for the training sets for bases and acids are plotted in Figures 4 and 5. The data points are distributed evenly around the diagonal in both figures, implying consistent error behavior of the residual values. For the bases, most of the data points fall into the pK_a range between 0 and 12, with a few extreme compounds being fitted well in line with the experiment values. The slope and intercept of the linear regression are optimal, 1.00 and 0.00, respectively. For the acids, there are at least two distinguishable clusters in Figure 5, with the one in the lower left regime primarily consisting of carboxylic and sulfonic acids, and the one of higher pK_a's composed mostly of alcohols and phenols (suggesting that perhaps they should be modeled separately). The slope and intercept of the linear regression is 1.00 and 0.00, respectively.

Another way to evaluate the models is to examine the fitted residues. More than 80% of the bases in the training sample are predicted with an accuracy of within one log unit of their measurements, and 96% are within two log units of the accuracy. The largest error for bases occurs for pyrazine, whose measured pK_a is 0.6 while predicted to be 3.6, a discrepancy of 3 log units. The replacement of carbon by nitrogen at the para position drives the pK_a from 5.17 for pyridine down to 0.6 for pyrazine, which is obviously not sufficiently accounted for by this model. An explicit treatment of the resonance effects could possibly compensate for the error and will be a future development of the models. Nitromethane and 2-nitropropane are the two most outstanding outliers in the acid set, the former with the predicted and measured values of 10.21 and 5.57 and the latter of 12.87 and 7.68, respectively. Most (86%) of the acids in the training sample are predicted within one log unit of accuracy, and 98% are predicted within two log units of accuracy.

As pointed out previously, although each level of the hierarchical tree can have as many as 33 different atom/group types, some of them may be absent depending on the particular molecular structure. For the training samples the algorithm used 93 columns that were filled in the connectivity table for the base data set and 97 for the acid data set. Therefore there are sufficient number of rows to help avoid the danger of overfitting, and the results from cross-validation certainly support that the models are not chance correlations.

b. Predictions. We used Perrin's examples of different chemical classes as an external test set.[15] After removing the compounds already existing in the training set, 25 novel molecules remain. The model predicted pK_a values, in comparison with Perrin's predictions, and the experimental measurements are listed in Table 5.

Quite a few molecules in this test set contain multiple ionizable centers, which could be either acidic or basic or could be mixed. The exact pK_as being considered were specified in the "Comment" field if confusion may arise as to the precise center being considered. It is, for example, easier to tell acids from bases and vice versa. But for molecules consisting of multiple basic or acidic groups, the relative basicity or acidity needs to be ranked in order to relate correctly the titration curves at different pH levels to the corresponding ionizing centers.

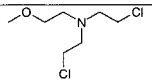
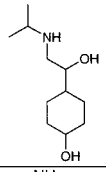
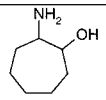
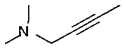
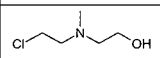
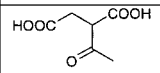
While evaluating the relative strengths of the multiple ionizing centers, it is worth noting that these models consistently predicted *the correct order* of the ionizing strengths for all the test cases explored. For example, in "furan-2,4-dicarboxylic acid", the first deprotonating center is the carboxylic acid ortho to the furan oxygen, and 4-carboxylic acid will ionize afterward according to experimental measurements. The pK_as for the 2- and 4-carboxylic acids were predicted separately, each time assuming the other acid is still protonated, and the values are 2.37 and 3.58, respectively. Therefore the 2-carboxylic acid is more acidic than that on the 4-position, which tallies with the experimental result. Similarly, an example for the bases is "4-aminopyridazine". It is probably not so difficult to speculate that the ring nitrogens are more basic than the amino group, but which one of the two pyridazine nitrogens will be protonated first, hence correspond to the first

Table 5. pK_a Predictions on 25 Organic Molecules

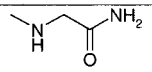
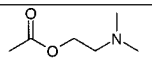
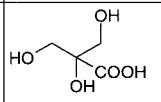
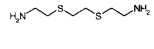
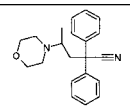
(a) Prediction of pK_a values of substituted aliphatic acids and bases:

Acid-strengthening: $-\Delta pK_a = 0.06 + 0.63\ \sigma^*$

Base-weakening: $-\Delta pK_a = 0.28 + 0.87\ \sigma^*$

Name	Structure	Perrin	Model	Measured	Comment
Bis(2-chloroethyl)(2-methoxyethyl)amine		5.10	5.33	5.45	
1-(4'-hydroxycyclohexyl)-2-(isopropylamino)ethanol		9.99	11.32	10.23	pK_a for amino
2-aminocycloheptanol		9.67	10.19	9.25	pK_a for amino
N,N-dimethyl-2-butyn-1-amine		~ 8.1	10.44	8.28	
5-chloro-3-methyl-3-azapentanol		7.1	8.16	7.48	pK_a for amino
2-acetylbutanedioic acid		3.15	3.19	2.86	pK_a1: 1-COOH; pK_a2: 4-COOH Predicted pK_a1 for 4-COOH: 3.68

(b) The Taft equation: $pK = pK^0 - \rho^* \Sigma(\sigma^*)$

Name	Structure	Perrin	Model	Measured	Comment
2-(methylamino)acetamide		8.43	9.07	8.31	pK_a for amino
2-(dimethylamino)ethyl acetate		8.26	8.38	8.35	pK_a for amino
2,3-dihydroxy-2-hydroxymethyl propanoic acid		3.01	3.42	3.29	pK_a for COOH
1,8-diamino-3,6-dithiaoctane		9.06	9.62	9.47	
4-morpholino-2,2-diphenylpentanenitrile		6.38	6.51	6.05	

(c) Prediction of pK_a values for phenols, aromatic carboxylic acids and aromatic amines

The Hammett equation: $pK_a = pK_a^0 - \rho(\Sigma\sigma)$

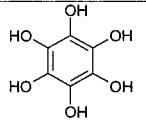
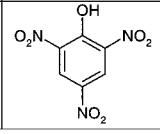
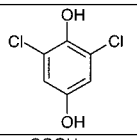
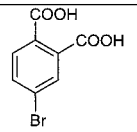
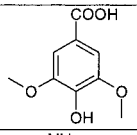
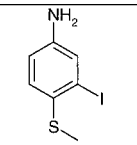
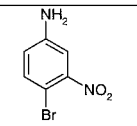
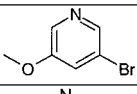
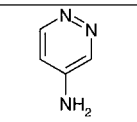
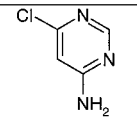
Name	Structure	Perrin	Model	Measured	Comment
Benzenehexol		8.31	8.60	9.0	
Picric acid		0.91	2.89	0.33	

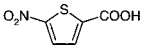
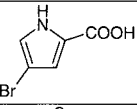
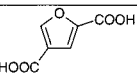
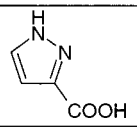
Table 5. (Continued)

Name	Structure	Perrin	Model	Measured	Comment
2,6-dichloro-1,4-benzenediol		6.82	6.50	7.30	pK_a1: 1-OH; pK_a2: 4-OH Predicted pK_a1 for 4-OH: 7.85
4-bromo-1,2-benzenedicarboxylic acid		2.86	2.74	2.5	pK_a1: 2-COOH; pK_a2: 1-COOH Predicted pK_a1 for 1-COOH: 3.48
4-hydroxy-3,5-dimethoxybenzoic acid		4.36	4.30	4.34	pK_a1: COOH; pK_a2: 4-OH Predicted pK_a1 for 4-OH: 7.13
3-iodo-4-methylthioaniline		3.34	3.44	3.44	
4-bromo-3-nitroaniline		1.82	1.80	1.80	

(d) Prediction of pK_a values of heteroaromatic acids and bases by the Hammett and Taft equations.

Name	Structure	Perrin	Model	Measured	Comment
3-bromo-5-methoxypyridine		2.30	2.50	2.60	
4-aminopyridazine		5.31	6.66	6.65	pK_a1: 1-N; pK_a2: 2-N Predicted pK_a1 for 2-N: 3.64
4-amino-6-chloropyrimidine		1.41	3.52	2.10	pK_a1: 3- N; pK_a2: 1- N Predicted pK_a2: 1.30

(e) Extension of the Hammett and Taft equations to heterocycles.

Name	Structure	Perrin	Model	Measured	Comment
4-nitrothiophen-2-carboxylic acid		2.70	2.44	2.68	
4-bromopyrrol-2-carboxylic acid		4.05	3.62	4.06	pK_a for COOH
Furan-2,4-dicarboxylic acid		2.77	2.37	2.63	pK_a1: 2-COOH; pK_a2: 4-COOH Predicted pK_a1 for 4-COOH: 3.58
Pyrazole-3-carboxylic acid		3.98	3.86	3.74	pK_a for COOH

measured pK_a? The basicity of the two nitrogens were predicted separately, each time with the other nitrogen being deprotonated, or in other words, with the molecule kept in its neutral state. The resulting values are 6.66 and 3.64 for

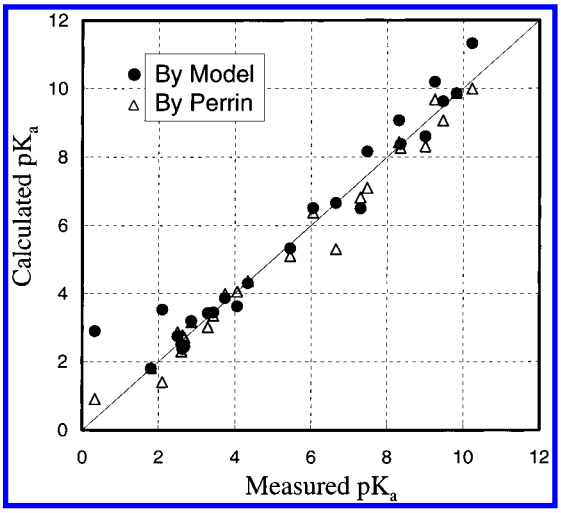


Figure 6. The pK_a values for 25 organic molecules (including acids and bases) predicted by the models and by Perrin vs the experimental values.

Table 6. Summary of Statistics of pK_a Predictions on 25 Organic Molecules

	predicted r^2	SE	$\|\Delta\| > 0.5$	$\|\Delta\| > 1.0$	deviations −	deviations +
model	0.95	0.70	7	3	8	17
Perrin	0.98	0.41	4	1	13	12

the 1- and 2-nitrogens, respectively. This infers that the 1-nitrogen is more basic, which is again in agreement with the experimental observations.

The predicted versus measured pK_a's by both the present and Perrin's models are also plotted in Figure 6. The statistics are summarized in Table 6.

The tree-structured fingerprint model is comparable in performance to Perrin's algorithm, which performs slightly better on average on the test set. It is encouraging that a large number of diverse molecules have low predicted errors using the model. There is scope for improvement of the algorithm, in particular toward developing specific models for major different classes, e.g. aniline, pyridine, phenol, etc. which are expected to result in significant improvements in prediction.

CONCLUSIONS

We have investigated novel methods for the prediction of logP (based on polarizability and atom charge) and pK_a (based on a novel tree structured fingerprint). The preliminary investigations of these methods reported here imply that they may be developed into useful methods for the estimation of logP, pK_a, and logD. The methods are fast and use simple algorithms which offer significant development potential and insights into the mechanisms of partitioning and acid/base properties.

ACKNOWLEDGMENT

We thank Dr. Micheal Lawless, Dr. Richard Cramer, and Dr. Robert Clark for helpful discussions. We are also grateful to Dr. Robert Clark and the anonymous reviewer for their suggestions on the manuscript.

REFERENCES AND NOTES

(1) Leo, A.; Hansch, C.; Elkins, D. Partition Coefficients and Their Uses. *Chem. Rev.* **1971**, *71*, 525.
(2) Linpinski, C. A.; Lombardo, F.; Dominy, B. W.; Freeny, P. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3−25.
(3) Winiwarter, S.; Bonham, N. M.; Ax, F.; Hallberg, F.; Hallberg, A.; Lennernas, H.; Karlen, A. Correlation of Human Jejunal Permeability (in vivo) of Drugs with Experimentally and Theoretically Derived Parameters. A Multivariate Data Analysis Approach. *J. Med. Chem.* **1998**, *41*, 4939−4949.
(4) Hansch, C.; Leo, A. J. *Substituent constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.
(5) Rekker, R. F.; Mannhold, R. *Calculation of Drug Lipophilicity*; VCH Publishers: Weinheim, 1992.
(6) Rekker, R. F.; ter Laak A. M. On the Reliability of Calculated logP−Values: Rekker, Hansch/Leo and Suzuki Approach. *Quant. Struct.-Act. Relat.* **1993**, *12*, 152−157.
(7) Mannhold, R.; Rekker, R. F.; Sonntag, C.; Ter Laak, A. M.; Dross, K.; Polymeropoulos, E. E. Comparative Evaluation of the Predictive Power of Calculation Procedures for Molecular Lipophilicity. *J. Pharm. Sci.* **1995**, *84*, 1410−1419.
(8) Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer Automated logP Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752−781.
(9) Buchwald, P.; Bodor, N. Octanol−Water Partition: Searching For Predictive Models. *Curr. Med. Chem.* **1998**, *5*, 353−380.
(10) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(11) Duprat, A. F.; Huynh, T.; Dreyfus, G. Toward a Principled Methodology for Neural Network Design and Performance Evaluation in QSAR. Application to the Prediction of logP. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 586−594.
(12) Beck, B.; Breindl, A.; Clark, T. QM/NN QSPR Models with Error Estimation: Vapor Pressure and LogP. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1046−1051.
(13) Reynolds, C. H. Estimating Lipophilicity Using the GB/SA Continuum Solvation Model: A Direct Methord for Computing Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 738−742.
(14) Ruelle, P. Universal Model Based on the Mobile Order and Disorder Theory for Predicting Liphphilicity and Partition Coefficients in All Mutually Immiscible Two-Phase Liquid Systems. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 681−700.
(15) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. *pK$_a$ Prediction for Organic Acids and Bases*; Chapman and Hall: New York, 1981.
(16) 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.
(17) 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.
(18) Citra, M. J. Estimating the pK_a of Phenols, Carboxylic Acids and Alcohols from Semiempirical Quantum Chemical Methods. *Chemosphere* **1999**, *38*, 191−206.
(19) Warwicker, J. Simplified Methods for pK_a and Acid pH-Dependent Stability Estimation in Proteins: Removing Dielectric and Counterion Boundaries. *Protein Sci.* **1999**, *8*, 418−425.
(20) Glen, R. C. A Fast Empirical Method for the Calculation of Molecular Polarizability. *J. Computer-Aided Mol. Des.* **1994**, *8*, 457−466.
(21) Slater, J. C. Atomic Shielding Constants. *Phys. Rev.* **1930**, *36*, 57−64.
(22) Sybyl is a product of Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144; www.tripos.com.
(23) *CRC Handbook of Chemical and Physics*, 72nd ed.; Lide, D. R., Ed.; CRC Press: Boston, 1991.
(24) Pomona Medchem database (MASTERFILE); Pomona College: Claremont, CA 91711.
(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: 1985; Tables 5−8.
(27) Klopman, G.; Iroff, L. D. Calculation of partition coefficients by the charge density method. *J. Comput. Chem.* **1981**, *2*, 157−160.
(28) Bodor, N.; Gabanyi, Z.; Wong, C.-K. A New Method for the Estimation of Partition Coefficient. *J. Am. Chem. Soc.* **1989**, *111*, 3783−3786.
(29) Platts, J. A.; Abraham, M. H.; Butina, D.; Hersey, A. Estimation of Molecular Linear Free Energy Relationship Descriptors by a Group Contribution Approach. 2. Prediction of Partition Coefficeints. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 71−80.
(30) Brickmann, J.; Goetze, T.; Heiden, W.; Moeckel, G.; Reiling, S.; Vollhardt, H.; Zachmann, C.-D. Interactive Visualization of Molecular Scenarios with MOLCAD/SYBYL. In *Data Visualization in Molecular*

Science − Tools for Insight and Innovation; Bowie, J. E., Ed.; Addison-Wesley Publishing Company Inc.: Reading, MA, 1995; pp 83−97.
(31) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *43*, 3714−3717.
(32) Unity is a product of Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144; www.tripos.com.
(33) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *103*, 3902−3909.

CI010315D