pK_a Prediction of Monoprotic Small Molecules the SMARTS Way

Adam C. Lee, Jing-yu Yu, and Gordon M. Crippen*

Department of Medicinal Chemistry, College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109

Received June 3, 2008

Realizing favorable absorption, distribution, metabolism, elimination, and toxicity profiles is a necessity due to the high attrition rate of lead compounds in drug development today. The ability to accurately predict bioavailability can help save time and money during the screening and optimization processes. As several robust programs already exist for predicting log*P*, we have turned our attention to the fast and robust prediction of pK_a for small molecules. Using curated data from the Beilstein Database and Lange's Handbook of Chemistry, we have created a decision tree based on a novel set of SMARTS strings that can accurately predict the pK_a for monoprotic compounds with R^2 of 0.94 and root mean squared error of 0.68. Leave-some-out (10%) cross-validation achieved Q^2 of 0.91 and root mean squared error of 0.80.

INTRODUCTION

Intense focus is being placed on the quick and accurate prediction of physicochemical properties, driven in particular by the pharmaceutical industry and the need to identify lead compounds with favorable absorption, distribution, metabolism, elimination, and toxicity (ADMET). It is well-known that pK_a, and in particular the ionic state of a molecule at physiological pH, affects pharmacokinetics and pharmacodynamics. Bioavailability measures, often characterized by the octanol/water partition coefficient log*P* and Lipinski's rule-of-five,[1] now include pK_a so as to determine the pH-dependent distribution coefficient, log*D*.

Aqueous solubility, lipophilicity, and amphiphilicity all contribute to intestinal absorption and membrane permeability, and these are at least partially determined by pK_a.[2] In order to be absorbed, orally administered drugs must first dissolve in the gastrointestinal (GI) fluids. The ionic state of a molecule can be affected during passage through the GI tract, due to environmental pH in the stomach (pH 1−3), duodenum (pH 5−7), and jejunum and ileum (pH ∼8).[3] Furthermore, the majority of drugs administered orally are ionizable at physiological pH levels.[4,5] Significant dissolution enhancement has been observed when the buffer maintains the pH near or above the pK_a of the dissolving drug.[6] Therefore, adjusting the pK_a of a drug is of particular interest when dissolution has been found to be the rate-limiting step in the process of absorption, especially when dealing with drugs having poor water solublity.

Ionizable groups also affect the ability of a drug to interact with a target. It has been shown that pK_a influences the rate and site of metabolism of drugs by CYP1A2, a metabolic enzyme.[7] As strong electrostatic and hydrogen bonding interactions are key contributors to the overall free energies of binding,[8] pK_a can be critical for binding potency at the target. Moreover, based on a study of compounds targeting the human Ether-a-go-go Related Gene (hERG) potassium channel, selectivity can be influenced by controlling pK_a.[9]

Issues of toxicity are also related directly to a drug's pK_a at physiological pH, such as cardiovascular toxicity due to the lengthening of time between the start of the Q wave and the end of the T wave in the heart's electrical cycle or QT prolongation, resulting from the blockade of the hERG potassium channel.[10] Another common toxicity issue that can be affected by the ionic state of a drug molecule is the potential liability for phospholipidosis, an adverse drug reaction that occurs with many cationic amphiphilic drugs.[11] All in all, approximately 50% of all drug failures have been attributed to poor ADME, and in 2003 approximately $8 billion was lost due to the inability to predict the toxic nature of a substance during the early stages of drug development.[12−14]

The main advantage of *in silico* pK_a prediction is that physical samples are not needed. Even when one considers newer methods of high throughput pK_a screening, there are two limiting factors: the costs and time associated with obtaining or synthesizing the compounds of interest. Hence, there is a need for a quick, accurate, and robust model for pK_a prediction for large as well as small compound libraries.

Much research has been done in the area of pK_a prediction. Seminal publications[15,16] on the prediction of pK_a base their model on linear free energy relationships (LFER), applying the Hammett equation. LFER models are still commonly used and are also implemented in popular commercially available software packages, such as SPARC.[17−19] One of the most common techniques used in pK_a prediction is quantitative structure activity/property relationships (QSAR/QSPR) deriving their fit equations from partial least squares (PLS) or multiple linear regression (MLR).[5,18,20−25] Other methods include neural networks,[26] quantum mechanical continuum solvation models,[27−29] and anticonnectivity models.[30,31] It has often been the case that a model was based on a relatively small set of experimental data for a specific ionizable group, such as carboxylic acids.[18,22,25−27,29−31] Others have tackled the problem of chemical diversity by devising and combining multiple models, each applied to a relatively small set of compounds when compared to the complete set of experimental data.[23,26] Here the overall combination of models is

* Corresponding author e-mail: gcrippen@umich.edu.

10.1021/ci8001815 CCC: $40.75
Published on Web 10/01/2008

more robust at handling novel chemical structures, but the individual training sets may suffer from a lack of chemical diversity due to their small size. This may allow for a good fit on the training sets but has the potential disadvantage of leaving little freedom for cross-validation. Furthermore, there is a danger of cherry picking or manually selecting compounds that are well represented in the training set for the test set. In the Methods section we describe how our single model is derived and applied to a large and diverse training set of monoprotic molecules. We also describe how the training and test sets were randomly selected from diverse clusters of compounds in order to ensure that no cherry picking occurred and that chemical space was fairly represented by all training and test sets for both the final model and in cross-validation.

METHODS

All calculations were performed using the Molecular Operating Environment (MOE)[32] on a Dell Precision 380 workstation utilizing Red Hat Linux Enterprise version 4.0.

1. Preliminary Studies. *1.1. Data Acquisition.* Data were obtained from Lange's Handbook of Chemistry 15th Edition[33] and the Beilstein (2007/04) database via the Molecular Design Limited (MDL) CrossFire Commander.[34] In all cases possible we applied the following filters in our curation of monoprotic compounds where the ionizable site was an oxygen, nitrogen, or sulfur atom: titration performed in H_2O at 23−27 °C, ionic strength (the molar concentration of all ions present in a solution) less than or equal to 0.1 M, and pK_a range from −1.74 (H_3O^+) to 15.7 (H_2O). We also accepted multiprotic compounds, such as *o*-nitrophenol, where only one ionizable site had pK_a within our accepted pK_a range.

1.1.1. The Lange data set provided International Union of Pure and Applied Chemistry (IUPAC) nomenclature and pK_a values for 2162 compounds with up to six ionizable sites. We accepted 700 of the monoprotic compounds: 417 carboxylic acids and alcohols, 14 thioacids and thiols, and 269 compounds having ionizable nitrogens. IUPAC nomenclature for each of these compounds was translated into simplified molecular input line entry specification (SMILES) strings both manually and using ChemDraw Ultra 10.[35]

1.1.2. The Beilstein database provided us with 10334 unique substances. These substances were washed to remove any salts; thus only the major component with the largest number of bonded atoms was retained. After applying our filters, we accepted 1577 monoprotic compounds: 898 carboxylic acids and alcohols, 33 thioacids and thiols, and 642 molecules having ionizable nitrogens. The following relevant fields were downloaded in structure data format (SDF):[36] Beilstein registry number (BRN), molecular structure, dissociation exponent (DE) or pK_a, dissociation group (DE.GRP), dissociation temperature (DE.T), dissociation solvent (DE.S), dissociation method (DE.MET), dissociation type (DE.TYP), and dissociation comments (DE.COM).

1.1.3. Many published pK_a values for the same compound were often found to exist. This was partly due to the large overlap between Lange and Beilstein. Beilstein also contained multiple instances with literature references where more than one pK_a value was published for a single compound. The two data sets including their duplicate entries were merged into a single MOE database. In the cases where the pK_a values for a specific compound varied less than 0.5 units, the mean was accepted as the experimental pK_a. In all instances where the variation was greater than 0.5 units, the literature reference was checked to explain the discrepancy, which was often due to the use of a solvent other than H_2O. Only three compounds were accepted that had variations greater than 0.5. All had variations less than 0.7 units.

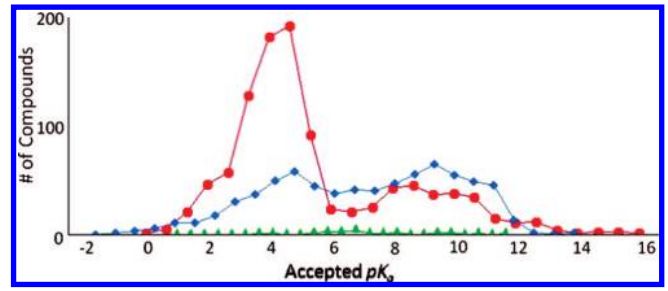

Figure 1. Each pK_a distribution is based on 25 bins for the monoprotic compounds having one of the three ionizable atom types considered here: oxygen (red ●), nitrogen (blue ◆), and sulfur (green ▲).

The resulting data set contained 1881 unique monoprotic compounds, involving 1088, 33, and 760 ionizable oxygen, sulfur, and nitrogen atoms, respectively. 2086 experimental pK_a values were found for the 558 compounds having multiple literature references. The RMSE for the duplicate experimental pK_a values from literature and our accepted experimental values after curation was 0.08. Figure 1 portrays the pK_a distributions of the ionizable atom types of interest.

1.2. Clustering Molecules To Obtain Equally Diverse Training and Test Sets. For cross-validation we separated our compounds into ten groups of equal size and having similar chemical diversity. To do this, we built a decision tree using the 166 Molecular ACCess System (MACCS) keys from MDL and required the leaf nodes to contain a minimum of ten compounds. The root of the tree consists of all the monoprotic compounds. Branching decisions are made by the MACCS key which most evenly (or ideally) divided the compounds at any given node and terminates when the node cannot be split into two children each containing at least ten compounds. See section 1.4 for a description of the ideal fingerprint.

Equations 1−4 describe how decisions are made at each node. $\mathbf{h}_k$ and $\hat{\mathbf{h}}_k$ are binary vectors with length equal to the number of compounds m at a node with $\mathbf{h}_k$ representing the hit profile and $\hat{\mathbf{h}}_k$ representing the inverse hit profile with respect to MACCS key k. $|\mathbf{h}_k|$ and $|\hat{\mathbf{h}}_k|$ are the number of compounds hit and missed, respectively. See eqs 1−3.

$$|\mathbf{h}_k| = \mathbf{h}_k \bullet \mathbf{h}_k = \sum_{j=1}^{m} h_{kj} \quad (1)$$

$$|\hat{\mathbf{h}}_k| = \hat{\mathbf{h}}_k \cdot \hat{\mathbf{h}}_k = \sum_{j=1}^{m} \hat{h}_{kj} \quad (2)$$

$$m = |\mathbf{h}_k| + |\hat{\mathbf{h}}_k| \quad (3)$$

Eq 4 defines the idealness score, I_k, for MACCS key k where $0.5 \leq I_k \leq 1.0$. A value of 0.5 indicates that the compounds at a particular node are evenly divided into children nodes, whereas a score of 1.0 indicates that one child node contains all of the compounds and the other contains none. The compounds at each leaf were randomly

Table 1. Example MOE SMARTS Strings

SMARTS string[a]	definition
[OH]aaa[#X]	non-carbon aliphatic atom, [#X], meta to a hydroxyl, [OH]
a[#G7]	any aromatic atom with a halogen substituent, [#G7]
a[C;!i]	any aromatic atom connected to a non-π-bonded carbon atom
[OH][A;r]=[A;r][A;!r]=O	hydroxyl covalently bonded to a nonaromatic ring atom, which shares a double bond with an adjacent aliphatic ring atom that is single bonded to a nonring atom sharing a double bond with oxygen
[OH]A[Ov2]	hydroxyl bonded to any aliphatic atom sharing a single bond to an oxygen that is explicitly bonded to two non-hydrogen atoms
A[N+]=O	aliphatic atom covalently bonded to the positively charged nitrogen of a nitroso group

[a] The strings described above were specifically selected to help readers unfamiliar with SMARTS notation better understand the SMARTS strings presented in Table 2.

divided into ten bins, each consisting of approximately 10% of the monoprotic compounds. Nine of the bins were used for training purposes, while the tenth was set aside as a test set for our final model.

$$I_k = \frac{\max[|\mathbf{h}_k|, |\hat{\mathbf{h}}_k|]}{m}, \text{ for } k = 1, \ldots, 166 \quad (4)$$

1.3. Analyzing the Molecular Diversity of the Training Set Using Principal Components Analysis (PCA) and the MACCS Keys. PCA was applied to assess the molecular diversity of the 1693 compounds. We formed a binary matrix consisting of 1693 rows (the training set compounds) and 166 columns (the MACCS keys) with every entry being either a '1' or a '0', corresponding to the respective presence or absence of each MACCS key. Of the 166 MACCS keys, 152 are found in our training set. PCA showed that 62 principal components accounted for 90% of the total system variance. Furthermore, 144 principal components were required to account for 100% rounded to the nearest thousandth of the total system variance. A nontrivial number of principal components are required to explain the vast majority of the system variance; hence, the training set has been shown to be widely diverse and applicable for the methodology implemented in developing our model.

1.4. Tailoring an Ideal Fingerprint with Predictive Power Using SMARTS Descriptors. **1.4.1.** An ideal fingerprint is one in which the descriptors are mutually independent and each evenly divides a data set of n compounds into those having vs not having the descriptor. Here we use MOE SMiles ARbitrary Target Specification (SMARTS) strings as descriptors, as described in Table 1. Note, while MOE SMARTS strings are based on the concepts of Daylight SMARTS strings[37] and are for the most part the same, there are some differences. For example MOE uses the SMARTS string [i] to denote any π bonding atom, but this is not used in Daylight SMARTS.

The ideal fingerprint has length n_d, where n_d is the minimum number of descriptors to uniquely identify all n compounds in the training set.

$$n_d = \lceil \log_2 n \rceil \quad (5)$$

Such a fingerprint uses the minimum number of descriptors possible to uniquely identify each compound in a specific data set, that is, the fingerprint tends to be ideal only for the data set on which it was based.

1.4.2. Each n_d-digit binary number representing the occurrence/absence of descriptors can be used not only to identify the respective unique compound but also may correlate with physicochemical properties.

Similarity measures, such as the Tanimoto score described in eq 6,[38] can then be applied to associate such physicochemical properties with structurally similar molecules. Let $\mathbf{f}_i$ and $\mathbf{f}_j$ be binary vectors of equal length where each vector component represents the presence or absence of some qualitative descriptor as 1 or 0, respectively. Then the Tanimoto similarity score $s_{i,j}$ is the ratio of intersection and the union of the two sets of qualitative descriptors. If no descriptors from vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ match, then $s_{i,j} = 0$. If exactly the same descriptors are present in $\mathbf{f}_i$ and $\mathbf{f}_j$, then $s_{i,j} = 1$. Finally, if some, but not all, of the descriptors in $\mathbf{f}_i$ match those in $\mathbf{f}_j$, then $s_{i,j}$ will be some rational value between 0 and 1. The closer $s_{i,j}$ is to 1, the more similar are the vectors $\mathbf{f}_i$ and $\mathbf{f}_j$. In our case, pairs of molecules with sufficiently high $s_{i,j}$, based on an ideal or close to ideal fingerprint, are expected to have similar pK_a values.

$$s_{i,j} = \frac{\mathbf{f_i} \bullet \mathbf{f_j}}{|\mathbf{f_i}| + |\mathbf{f_j}| - \mathbf{f_i} \bullet \mathbf{f_j}} \quad (6)$$

1.5. Concept Validation Using Molecular Similarity Measures. In order to validate our concept that pK_a can be predicted based on the structural similarity between two molecules using SMARTS strings as descriptors, we applied it to our data set of 403 monoprotic alcohols.

1.5.1. 202 monoprotic alcohols from five of the ten bins were selected as the training set, leaving the remaining 201 alcohols as the test set.

1.5.2. The 50 SMARTS string descriptors described by Table 2 were manually created and uniquely identified the molecules in the alcohol training set. An ideal fingerprint with 50 descriptors should be able to uniquely identify 2^{50} molecules. This is gross overfitting in our case, as $2^{50} - 202 = \sim 10^{15}$ binary profiles correspond to no molecule. In our case, the ideal set of descriptors would contain only 8 SMARTS strings, the minimum number of descriptors to uniquely identify all 202 alcohols.

1.5.3. Apply reverse stepwise regression. Using the 50 SMARTS string descriptors (fingerprint components) as indicator variables, reverse stepwise regression on pK_a was applied in an attempt to minimize the number of SMARTS strings, as it was extremely unlikely that a set of ideal SMARTS string descriptors existed. First the fingerprint of all molecules in the training set was taken.

1.5.3.1. Drop the descriptor that least affects the fit R^2 of the training set. The predicted pK_a for each compound is assigned based on the average pK_a of the compounds sharing the highest $s_{i,j}$. Then the Pearson correlation coefficient (R^2) is calculated using the experimental vs predicted pK_a values. Deleting any one of the SMARTS strings can decrease the R^2, so we drop the string that causes the smallest decrease. In the case of a tie, drop the least ideal SMARTS string.

Table 2. Hand Selected SMARTS String Fingerprint That Uniquely Identifies the Alcohol Training Set

	SMARTS string subsets[a]		
type	A	B	C
aromatic centers	[OH]a, [OH]aa[#X], [OH]aaa[#X], [OH]aaaa[#X]	[OH]a(a[#X])a[#X], [OH]a(a[#X])aa[#X]	[OH]a(aa[#X])aa[#X], [OH]a[#7+]
aromatic modifiers	a[#G7], a[Ov2], aC=C	aC=O, a[Sv2], a[#8+]	a[N+]=O, a[#7+](C)C, a[S+], a[S+2], aC=*Cc*[n+], aC=C[N+](=O), a[C;!i], a[P+]
aliphatic centers	[OH]A=A, [OH]A[O-], [OH]AAA=O, [OH][A;r](=[A;r])[A;r]=O, [OH][A;r]=[A;r][A]=O	[OH]C, [OH]A=S, [OH][A;r]=[A;r][A;!r]=O	[OH]A[#X], [OH]AA[#X], [OH]AA([#X])[#X], [OH]AAA[#X], [OH]AAAA[#X], [OH]A(A[#X])A[#X], [OH]A[Ov2], [OH]A[Sv2]
aliphatic modifiers	A[#G7], A[N+]=O, [#7+](~*)~*	[Sv2], [n+]	[S+], [S+2], C=C, C=Cc[n+], C=C[N+]=O, [P+]

[a] The SMARTS string subsets A+B+C make up a fingerprint that uniquely identifies the 202 monoprotic alcohols in the training set, $R^2 = 1.0$. Subsets A and A+B make predictive fingerprints which respectively yielded $R^2 = 0.75$ and 0.86 on the test set of monoprotic alcohols.

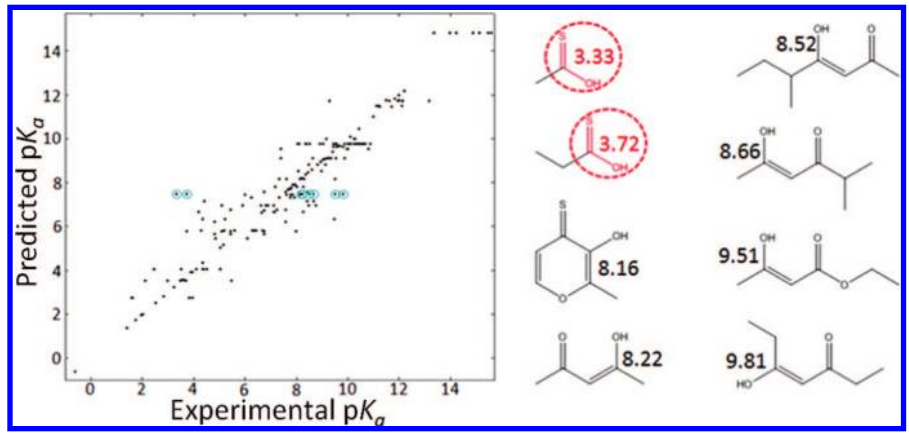

Figure 2. The scatter plot on the left depicts experimental vs predicted pK_a values for 202 monoprotic alcohols from the training set using the fingerprint defined by SMARTS string subset A, having an $R^2 = 0.90$. The molecules sharing the same fingerprint profile with largest experimental pK_a variance are identified by the points circled in blue (⊙). The molecular structure and pK_a of the highlighted points on the scatter plot are shown on the right. Highlighted and encircled in red, the fragment represented by the SMARTS string, [#8][i]=[#16], differentiates the molecules with low pK_a from those with high pK_a.

Retain the dropped SMARTS string in a pool for later re-evaluation steps.

1.5.3.2. Check pool containing dropped SMARTS strings for possible improvements to R^2 for the training set. If the discard pool contains more than one SMARTS string descriptor, then check whether re-adding any one descriptor from the discard pool except for the last dropped SMARTS string can increase the R^2 of the current model. If any such descriptors exist and their inclusion yields an R^2 greater than or equal to 0.90, re-add the SMARTS string that increases the R^2 of the model the most.

1.5.3.3. Check threshold values. Repeat the stepwise regression as described in steps 1.5.3.1−1.5.3.2 until the R^2 falls below 0.90.

1.5.3.4. Re-add the most recently dropped SMARTS string. In this manner, we obtained a best estimate for the most ideal predictive fingerprint that yields a correlation greater than or equal to 0.90 based on the training set.

1.5.4. Validation. Step 1.5.3 was repeated twice using threshold $R^2 = 0.95$ and 0.90 for the alcohols training set only. For all other investigations the threshold $R^2 = 0.90$ was used. The descriptors of the two predictive fingerprints that were identified are shown in Table 2. The first, having the lowest R^2 greater than 0.95, consisted of the 25 SMARTS string descriptors in subsets A and B. The second, having lowest R^2 greater than 0.90, used only the 15 SMARTS string descriptors in subset A. Applying the predictive fingerprints described by the combination of subsets A and B and of subset A alone to our test set provided concept validation and yielded $R^2 = 0.86$ and 0.75, respectively.

1.6. Expanding and Refining Predictive SMARTS Descriptors. Our goal is to identify very general SMARTS string descriptors, which are both as ideal as possible and differentiate groups of compounds by pK_a. Selecting a discriminating SMARTS string can often be a nontrivial task, so as not to make it too specific.

1.6.1. After applying the fingerprint consisting of only 15 SMARTS strings described in Table 2 (subset **A**), we examined the set of compounds where all members had the same fingerprint value but large experimental pK_a variance. In this case, the largest experimental pK_a variance is at the predicted pK_a of 6.48. Figure 2 demonstrates how outlier molecules were identified and the predictive fingerprint was modified.

1.6.2. We then generalized existing SMARTS descriptors by substituting more general atom and bond types in our existing descriptors. If no such modification could be identified, we added new descriptors until the R^2 became

Table 3. SMARTS String Refinement Process

group	training set size	no. of bins used	no. of SMARTS preregression	no. of SMARTS postregression[a]	no. of refined SMARTS[b]	fit R^2
[OH] and !C(=O)[OH]	202	5	50	15	21	0.96
[OH]	548	5	132	32	122	0.99
[#G6!H0]	564	5	139	35	123	0.99
[#7!H0]	606	8	145	59	130	0.96
all	1693	9	284	140	256	0.96
all (final)	1693	9			262[c]	0.95[c]

[a] Stepwise regression, reducing the number of SMARTS strings was performed until the fit R^2 was reduced to 0.90. [b] Refinement, as discussed in section 1.6, was performed to improve the fit R^2 to exceed 0.96. [c] The final refinements to the pool of SMARTS strings were made using the decision tree as discussed in step 2.2.

greater than 0.96 and allowed no predicted value to deviate by more than 1.0 pK_a units from the experimental value. Figure 2 depicts the thioacid moiety as the clear differentiating factor, separating the molecules into two smaller sets while minimizing the pK_a variance. When added to the fingerprint, the SMARTS string [#8][i]~[#16], representing an oxygen atom single bonded to a π bonding atom that shares some bond with a sulfur atom (inclusive of the thioacid), the R^2 rose from 0.90 to 0.92.

1.6.3. After the R^2 was improved to greater than 0.96 and deviation for all predicted values was less than 1.0, stepwise regression was performed again to identify the descriptors with least influence on the training set. These descriptors tend to be the most specific and are present in only a few molecules or almost all the molecules in the training set. Each set of molecules identified by such a descriptor was then analyzed with the intent of replacing the overly general or overly specific SMARTS string with a more ideal SMARTS descriptor that split the data set more evenly as well as successfully differentiated the set of molecules.

1.6.4. At this point the training set was broadened with 346 (~50%) of the monoprotic carboxylic acids, which were taken from the same bins as the monoprotic alcohols. The fingerprint was then modified to include the general structural differences between the alcohols and the carboxylic acids, namely the carbonyl group between the ionizable oxygen and the remainder of the molecules. This was done by adding SMARTS descriptors similar to those used in the fingerprint for the alcohol training set but inserting a carbonyl moiety in parentheses next to the ionizable oxygen, represented by [OH]. For example, the motif representing a non-carbon meta substitution of a phenol-like substance, [OH]aaa[#X], becomes [OH]C(=O)aaa[#X]. For our convenience, each SMARTS descriptor containing an ionizable atom places the ionizable moiety first in the string. Our process has led us to pay less attention to the general aromatic SMARTS 'a', representing any aromatic atom, and incorporate '[i]', representing a π bonded atom which matches both aromatic atoms as well as those with conjugated bonds. This generalized SMARTS string, [OH][i](=O)~[i]~[i]~[i][#X], will identify both meta substituted phenyl rings and some conjugated carboxylic acids. Following the logic from steps 1.4−1.6.4 the set of SMARTS strings was modified and refined. Next, 16 monoprotic molecules containing ionizable sulfur atoms were added to the ionizable oxygen training set and the SMARTS fingerprint was refined accordingly. The ionizable site of this entire class of monoprotic molecules can be described by the SMARTS string [#G6!H0], which identifies a group 6 atom from the periodic table bonded to at least one hydrogen.

1.6.5. Far more diversity is evidenced among the monoprotic molecules containing ionizable nitrogens which are identified by the SMARTS string [#7!H0]. To simplify matters, this set of compounds was divided into three groups (namely, primary, secondary, and tertiary amines) and treated in the same manner as the molecules identified by [#G6!H0].

1.6.5.1. Starting from scratch using molecules with ionizable nitrogens drawn from the same bins from which the alcohols, carboxylic acids, thiols, and thioacids were obtained, a new SMARTS string fingerprint was trained on a random selection from the primary ionizable nitrogen data set, containing both amines and amides. The secondary and tertiary ionizable nitrogen data sets were sequentially added and trained. It is of note that we were unable to derive sufficiently good predictive fingerprints for these groups of compounds using only 50% of the bins.

1.6.5.2. In order to achieve predictive fingerprints that surpassed $R^2 = 0.75$ for the test set of all ionizable nitrogens from the unused bins, we needed to expand our training set to include 80% of the ionizable nitrogens. Here the training set was expanded by including the ionizable nitrogen containing compounds from three of the unused bins of equally diverse compounds based on the MACCS keys.

1.7. Combining Fingerprints Trained for [#G6!H0] and [#7!H0] Monoprotic Molecules. It was obvious that there was some overlap between the descriptors (aliphatic and aromatic modifiers) from our predictive fingerprint trained on the compounds containing ionizable oxygen and sulfur atoms and the predictive fingerprint trained on the compounds containing ionizable nitrogen atoms. Based on the size of the combined fingerprints, it was likely that we were overfitting the data. However, this was expected as we were unable to identify any descriptor that was truly ideal.

1.8. Obtain Basis SMARTS String Fingerprint. Our final training set was formed using 90% (9 bins) of our monoprotic data, including the compounds from the aforementioned 8 bins. None of the curated SMARTS strings were derived from compounds in the tenth bin, our test set. We once again performed stepwise regression, identifying the most ideal basis set of descriptors, thus allowing for the manual discovery of other novel and very general descriptors based on the outliers. As before, the fingerprint was refined until R^2 exceeded 0.96 with all calculated values deviating less

than 1.0 from the experimental values. Details of the analysis and refinement of the SMARTS strings are shown in Table 3.

2. Training and Validating the Final Model. Figure 3 illustrates the generalized workflow used to develop the final model as described in steps 1.1−2.2.

2.1. Growing the Predictive Decision Tree. Our predictive decision tree is based on two factors: diversity and accuracy. Diversity is represented by the ideal nature of our SMARTS fingerprint. Using eq 4, we calculated a new diversity score, this time allowing a minimum of two molecules at each leaf node in the decision tree using our ideal SMARTS strings instead of the MACCS keys used when selecting our training and test sets. An element of accuracy is also included in making each branching decision. Eq 7 describes the accuracy score, A_k, which is intended to help minimize the pK_a range at each node in the tree. For both eqs 4 and 7, the index k now refers to the descriptors in our SMARTS fingerprint. Binary hit vectors $\mathbf{h}_k$ and $\hat{\mathbf{h}}_k$ are now based on the SMARTS fingerprint rather than the MACCS keys. m refers to the number of compounds at the parent node. r_k and $\hat{r}_k$ are the pK_a ranges at the children nodes for SMARTS string k. The values of the resulting score range from (0,1.0]. In this case, values less than 1.0 represent child nodes with smaller pK_a ranges than the parent node. Values closer to zero reflect child nodes with narrower pK_a ranges.

$$A_k = \frac{r_k|\mathbf{h}_k| + \hat{r}_k|\hat{\mathbf{h}}_k|}{\max_k[r_k, \hat{r}_k] \times m} \tag{7}$$

Eqs 4 and 7 have been normalized such that the lower values of the respective scores are desirable. Eq 8 describes how the final decision is made at each node and considers the ability of a particular SMARTS string to evenly split the substances and at the same time minimize the pK_a range of the children nodes. Weighting factors for I_k and A_k were systematically determined by spawning several decision trees and varying the weights. The final weighting scheme was the one resulting in the highest fit R^2 based on correlating the mean of all pK_a values of the leaf and the experimental pK_a values from our training set.

$$S = \min_k[I_k + 2.5 \times A_k] \tag{8}$$

2.2. Refining the Predictive Decision Tree To Establish the Final Model. New SMARTS strings were added based on outliers found when fitting the training set. The final SMARTS string fingerprint based on our training set consisted of 262 SMARTS strings, 139 of which were used in creating the decision tree, which gave calculated pK_as for the training set that fit the experimental values with R^2 greater than 0.95 and root mean squared error (RMSE) less than 0.65.

This method is a refinement of the one used in our previous prediction of generalized cytotoxicity where the MACCS keys were used to make decisions on a randomly partitioned training and test set.[39] Here manually derived SMARTS strings are used to make decisions, and the MACCS keys were used to uniformly partition the data into the training and test sets for cross-validation purposes.

3. Cross-Validation and pK_a Shuffling Studies. In order to perform proper cross-validation and shuffling studies, one should redo all steps in the method. Unfortunately, the method for obtaining the pool of SMARTS strings used in the decision tree is complex and immensely time-consuming. We have chosen to perform leave-some-out (10%) cross-validation and pK_a shuffling studies using the final pool of SMARTS strings. In the cross-validation procedure, the entire data set of 1881 molecules was used to form all possible combinations of nine of the diverse bins as training sets and the remaining bin in each situation as the test set.

pK_a shuffling studies were performed to evaluate the likelihood of overfitting the model. An attempt was made to repeat the entire procedure using only the alcohols. After shuffling the pK_a of training set compounds, reverse stepwise regression was performed as in step 1.5. Using the same 50 SMARTS strings, the R^2 for the training set was less than 0.46, so we were unable to continue the SMARTS string reduction procedure. This alone demonstrates that the 50 SMARTS strings being used could recognize bogus data rather than fit it. Consequently, the shuffling test was only applied to the final step of model development, leaving the laboriously selected SMARTS strings fixed.

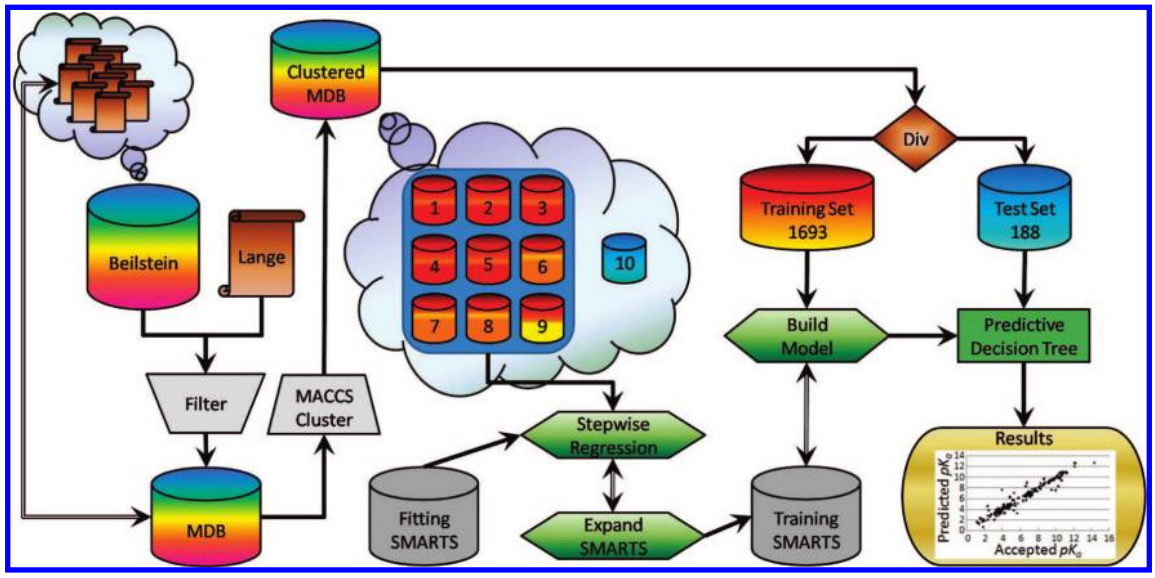

Figure 3. Workflow used to achieve the SMARTS pK_a decision tree. Clouds represent database contents, and the brown scrolls are literature references.

RESULTS AND DISCUSSION

The decision tree is 13 levels deep using 139 SMARTS strings with 1527 nodes including 741 terminal nodes containing two or more of the 1693 molecule training set. The decision pathway depends on the presence or absence of the SMARTS string descriptors from the pool and the equations provided in the methods section. It is clear that any SMARTS string can be used at any level in the decision tree and is often the case with general motifs, such as long carbon chains or linear strings of atoms. Examining the first five levels of the decision tree, it is readily evident at least to some degree that the main ionizable sites are separated into groups first. Starting at the root (node 1) the first decision is always whether the molecule contains a carboxylic or thioic acid (node 2) or not (node 3). This is immediately followed at node 2 by differentiating carboxylic (node 4) from thioic acids (node 5). At node 3, aliphatic charged conjugate acids (node 6) are differentiated from other amines and alcohols (node 7). Ortho, meta, and para substitutions and any combination thereof are decided in the middle levels of the tree along with other motifs relating the positions of substituents relative to the ionizable site in aliphatic molecules. The decision tree also tends to identify the more specific substituents closer to or at leaf nodes, such as the halogens and the methyl group as well as specialized branches and bonding configurations. As 1527 nodes exist in the decision tree and only 139 SMARTS strings are used, it is clear that the same SMARTS strings are being reused along different decision pathways. Therefore, it is the case that SMARTS strings determining the leaf nodes also occur in the middle of the decision tree but never in the same decision pathway.

Originally the test set contained 188 molecules, but two were dropped as their tautomeric forms lead to loss of the proton from a carbon atom, rather than the expected ionizable site. It is worth mentioning that the SMARTS pK_a prediction for these molecules had errors of only 0.2 and 0.91. A third molecule was dropped because of ambiguity in converting its name to a structure. The decision tree, including predictions and ranges for each node, is provided in the Supporting Information.

The decision tree, utilizing 139 SMARTS strings, was created from a pool of 262 SMARTS string descriptors. In order to deal with the primary concern of overfitting the model, we retrained the decision tree 100 times allowing the SMARTS descriptors to be selected from the original pool while randomly shuffling the pK_a values within the training set. The R^2 of the original predictive model was shown to lie over 28 standard deviations above the mean R^2 of the 100 models with randomized data, while the RMSE of the original model was over 422 standard deviations below the mean of the respective RMSEs obtained from the randomized data models. The statistics comparing the R^2 and RMSE of the final predictive model and the models with randomized pK_a data are shown in Table 4. Based on these results, it is clear that we have not overfit the model.

In order to validate that the decision tree model (SMARTS pK_a) was producing satisfactory predictions, a comprehensive literature survey was made. Furthermore, benchmarking on our test set was performed with SPARC, MARVIN, Advanced Chemistry Development (ACD)/Labs Online v. 8.03, and ADME Boxes. Details of the survey and benchmark are provided in Table 5, which includes descriptions and statistics for training sets as well as cross-validation and or external test sets when provided. In summary for the various methods and the respective training sets, the R^2 ranged from 0.81 to 0.99 and the RMSE was less than 1.0. R^2 for leave-one-out or leave-some-out (10% or 20%) cross-validation studies ranged from 0.78 to 0.92 with RMSE typically less than 1.0. Finally, R^2 for the external data sets ranged from 0.69 to 0.99. High values for the Pearson correlation coefficient and the lowest RMSE are expected on all the training set predictions (fits) with some falloff on predictions for cross-validation and external test sets.

Table 4. Overfitting Test: Accepted Model vs 100 Models with Randomized pK_a Data

model	statistic	R^2	RMSE
100 randomized	mean	0.4830	2.1382
	std. dev.	0.0162	0.0334
	mode	0.4909	2.1220
	min	0.4388	2.0721
	max	0.5145	2.2382
accepted		0.9548	0.6512

While SMARTS pK_a does not produce the highest fit scores, R^2 of 0.95 is very respectable. Two of the reasons our method does not outperform other methods are (a) the limited size of our complete data set and (b) the fact that the SMARTS strings were not inclusive of any increased molecular diversity found in the 185 compound test set. Furthermore, this method is not designed to produce a fit having $R^2 = 0.99$ due to the pK_a average taken at each terminal node. By decreasing the node size from two molecules to one, both the SMARTS fingerprint and the MACCS keys were able to produce a fit with $R^2 = 0.99$. Finally, the only reason the MACCS keys were unable to produce a fit of 1.0 with the node size reduced to one was because we did not eliminate stereochemistry or E/Z isomers from the data set. Neither our SMARTS strings nor the MACCS keys were able to distinguish between molecules having these structural differences. Altogether 152 of the 166 MACCS keys and 156 of the 262 SMARTS strings were used in this exercise. This is valuable information regarding the chemical diversity of the training set and the potential of our SMARTS strings to divide the compounds. In an ideal situation, only 11 MACCS keys or SMARTS descriptors would be required to build a tree having 2^{11} nodes which would uniquely identify each of the 1693 training compounds.

SMARTS pK_a also performed well in cross-validation with $Q^2 = 0.91$ and RMSE = 0.80. As the SMARTS strings were manually created (labor intensive) specific to the training set, we consistently retrained the cross-validation models with the same pool of SMARTS from step 2.2.

Finally, SMARTS pK_a performed exceptionally well on the external test set, which was randomly selected from the uniformly diverse MACCS descriptor space described in the Methods section. The predictive R^2 was 0.94 and RMSE = 0.68. In fact, we outperformed some respected online pK_a calculators based on three standards: robustness, overall accuracy, and fewest outliers.

Both SMARTS pK_a and ChemAxon's MARVIN pK_a calculator plug-in were able to provide predictions for all compounds, while the other methods missed some. Note

Table 5. Literature Survey of pK_a Calculators with Benchmarking Using SPARC, MARVIN, ACD/Labs Online v. 8.03, and ADME Boxes[g]

			training set			test set			external test set		
method	ref	class	*n*	R^2	RMSE	*n*	Q^2	RMSE	*n*	R^2	RMSE
QSPR/PLS	20	all subclasses							25	0.95[a]	0.78[a]
		acids	625	0.98	0.405	10%	0.86	1.04			
		bases	412	0.99	0.298	10%	0.87	1.12			
QSPR/PLS	23	33 subclasses	24617						39	0.80	0.90
		acidic nitrogen	421	0.97	0.41	20%	0.87	0.41			
		6 member N-heterocyclic bases	947	0.93	0.60	20%	0.85	0.86			
QSPR/PLS	24		49			49	0.86		23	0.77	
QSPR/MLR	21		15	0.97	0.12				3	0.99[a]	0.10[a]
QSPR/MLR	22	all subclasses									
		carboxylic acids	1122	0.81	0.42[b]	20%	0.81	0.43[b]			
		alcohols	288	0.82	0.76[b]	20%	0.81	0.78[b]			
QSPR/MLR	25	aromatic acids	74						33	0.99	0.27
QSPR/LFER	17	monoprotic oxy acids	135	0.99	0.455				14		0.471
continuum solvation	27	carboxylic acids							16	0.69	0.72
anticonnectivity	31		31			31	0.87	0.463			
neural network	26	12 classes	>16000						665	0.83	
(ChemSilico)		primary amine	1100	0.95		20%	0.92				
		tertiary amines	870	0.92		811	0.80				
		monoprotic acids	1640	0.95		1640	0.88				
		aromatic nitrogen	1480	0.92		1367	0.80				
		alcohols	1302	0.88		1302	0.85				
semiempirical/PLS	5	all			0.48				350		0.81
(Norvatis In-House)		alcohols	202	0.87	0.58		0.80				
		amines	1403	0.89	0.49		0.84				
		anilines	311	0.90	0.49		0.78				
		carboxylic acids	681	0.90	0.34		0.86				
		imines	84	0.98	0.55		0.88				
		pyridines	397	0.95	0.58		0.86				
		pyrimidines	91	0.95	0.43		0.87				
statistical thermo quantum solvation	28	bases	43	0.98	0.56[b]				58		0.66
(COSMO-RS)	29	acids	64	0.98	0.49[b]						
QSAR, LEFR (SPARC)	18		2500	0.99	0.36[b]				4338	0.99	0.37[b]
	19	Pfizer data set[c]							123	0.92	0.78[b]
	40	Pfizer internal data set[d]							537	0.80	1.05[b]
									185[c]	0.84	1.15
MARVIN	41,42					208[c]	0.98	0.38[b]	185[c]	0.88	1.03
ACD/Labs Online v. v8.03	43		>31000						185[c]	0.90	0.93
ADME Boxes	44								185[c]	0.93	0.69
SMARTS pK_a			1693	0.95	0.65	10%	0.91	0.80	185	0.94	0.68
	45,46								112	0.77	1.59
									107[e]	0.89[e]	1.04[e]
consensus[f]									185	0.96	0.60

[a] External set statistics were calculated from data presented in the referenced material. [b] Standard deviation. [c] It is unknown whether these molecules were used in the training set. [d] These molecules were unlikely to be found in the SPARC training set. [e] Results recorded after removing significant outliers from the referenced materials. [f] The consensus model used predictions from SPARC, MARVIN, ACD/Labs Online v. 8.03, and SMARTS pK_a. [g] In all training sets *n* refers to the number of pK_a measurements; in the test sets *n* refers to the number of pK_a measurements or percentage of the training set.

that SPARC was not designed to handle molecules containing selenium or silicon, which resulted in the missing prediction values. If one were to substitute oxygen for selenium and carbon for silicon, SPARC predicted the pK_a with error less than 1.0 in most of the instances. SMILES for all compounds and their predicted pK_a values from SMARTS pK_a, SPARC, MARVIN, ACD/Labs Online v. 8.03, and ADME Boxes are available in the Supporting Information. A consensus model, having $R^2 = 0.96$ and RMSE = 0.60, was derived using the mean of the three predictions with the smallest pK_a discrepancies from each of the aforementioned five methods. Statistics for all methods are provided in Table 5 and depicted in Figure 4. Table 6 summarizes the overall accuracy of the prediction methods as well as the consensus based on the number of compounds predicted within five ranges of error and missed predictions.

A second external test set consisting of 112 compounds satisfying our filters from more recent literature was identified.[45,46] Note, the primary pK_a for the majority of these molecules has been previously calculated in a comparison study which included all of our benchmarking toolkits.[45] Applying SMARTS pK_a to this data set resulted in an R^2 of 0.77 with RMSE of 1.59. The results are shown in Figure 5. Removing the largest five outliers improves the statistics to 0.89 and 1.04, respectively. It follows that the chemical space occupied by the outlier molecules was poorly represented in the training set.

The following is an example of a prediction made using our model. First, we selected a compound from the external test set, 4-(benzyloxy)benzoic acid. At the root of the decision tree the range of pK_a values of all compounds in the training set is 17.32. The predicted value is the mean pK_a of these compounds, 5.91 rounded to the nearest hundredth. The SMARTS string from our pool of SMARTS strings which optimally splits these compounds into two sets according to the scoring function defined in eq 8 is [#G6H]C(=O), identifying a Group 6 atom attached to both a

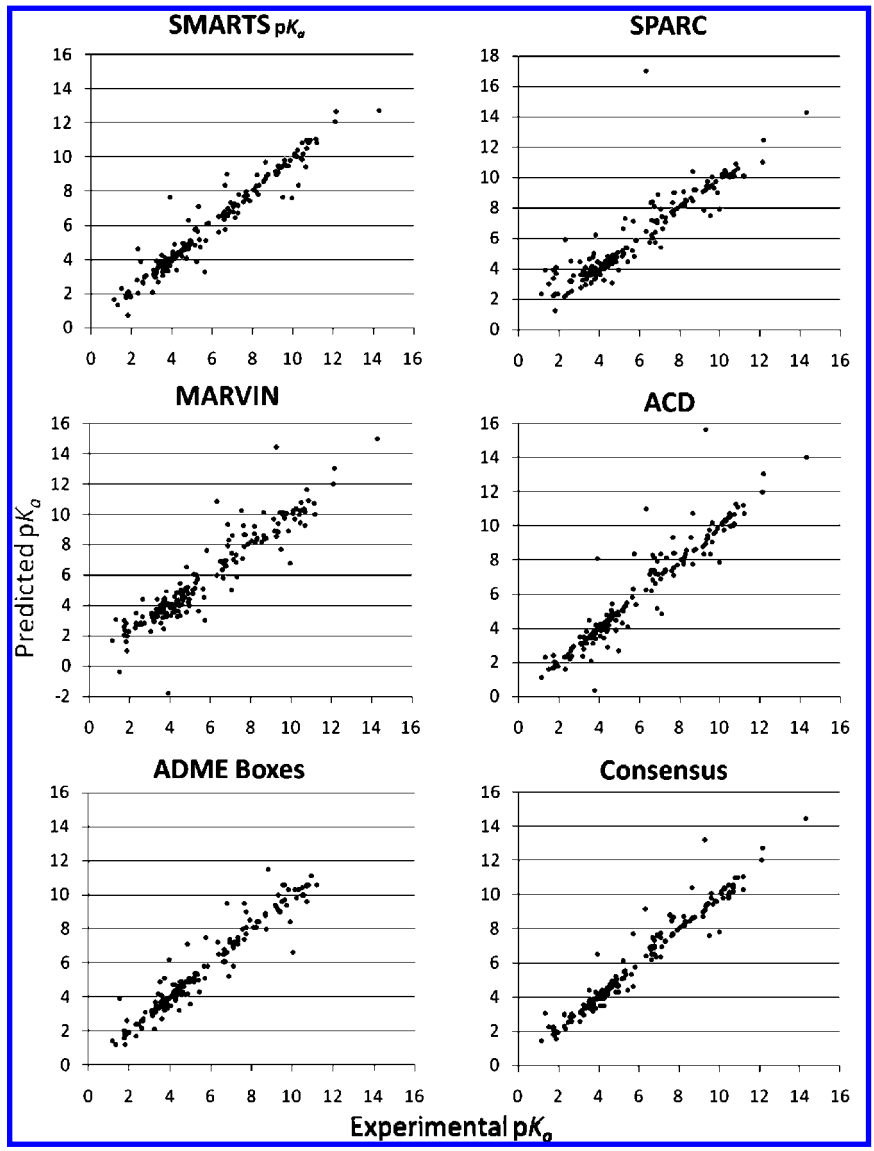

Figure 4. Scatterplots of the experimental vs predicted pK_a values for the 185 compounds test set with SMARTS pK_a, SPARC, MARVIN, and the ACD/Labs Online v. 8.03 and ADME Boxes.

Table 6. External Test Prediction Errors with Available online pK_a Calculators and Comparison to SMARTS pK_a[a]

pK_a error	SPARC	MARVIN	ACD	ADME Boxes	SMARTS pK_a	consensus
[0,1]	145	154	165	156	168	174
(1,2]	26	23	10	14	12	7
(2,3]	7	4	5	5	4	3
(3,4]	1	1	1	1	1	1
(4,∞)	1	3	3	0	0	0
miss	5	0	1	9	0	0

[a] For each method the number of compounds is given that had prediction errors in the given range.

lone hydrogen and a carbonyl. The SMARTS string positively identifies 4-(benzyloxy)benzoic acid, and another SMARTS string is selected from the pool which best splits this subset of compounds. The decisions leading to a final pK_a prediction for 4-(benzyloxy)benzoic acid are described in Table 7. Notice how the pK_a range at each child node remains the same or decreases, providing an overall estimate of accuracy based on the compounds sharing the terminal node.

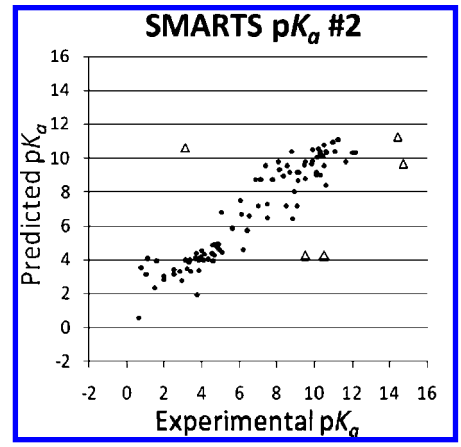

Figure 5. Scatterplot of the experimental vs predicted pK_a values for 112 compounds outside the training and first test sets. Outliers with predicted values differing more than 3 pK_a units are depicted as open triangles. Three of the outliers are secondary amines covalently bonded to either carbonyl or sulfonyl moieties. The others include triethenylamine and an aliphatic halogenated alcohol.

SMARTS pK_a is capable of addressing issues of resolution for both aromatic and aliphatic analogs as well as being able

Table 7. pK_a Prediction for 4-(Benzyloxy)Benzoic Acid Using the Final Model

node no.	SMARTS string	*is* identified[a]	*has* child[b]	pK_a of node	pK_a range
1	[#G6H]C(=O)	Yes	Yes	5.91	17.32
2	[OH][i](=O)*(~*)~*	Yes	Yes	3.19	5.96
4	a[#X]	Yes	Yes	3.53	5.88
8	*~*~*~*~*~*~*~*~*~*	Yes	Yes	3.22	3.99
16	[i][#G6v2]	Yes	Yes	3.17	3.99
32	[O][i]~[i]~[i]~[i]~[i]~[i]~[i]~A	No	Yes	3.64	2.92
65	[OH][i]~[i](~*)~*	Yes	Yes	4.06	1.51
130	[OH][i](=O)[i]~[i]~[i]~[i]~[i]-A	No	Yes	4.34	0.69
261	[CH3]	No	Yes	4.51	0.26
507			No	4.47	0.08

[a] Yes/No refers to the presence/absence of the SMARTS string in 4-(benzyloxy)benzoic acid. [b] Yes refers to nonterminal nodes having children, while no refers to the terminal or leaf node.

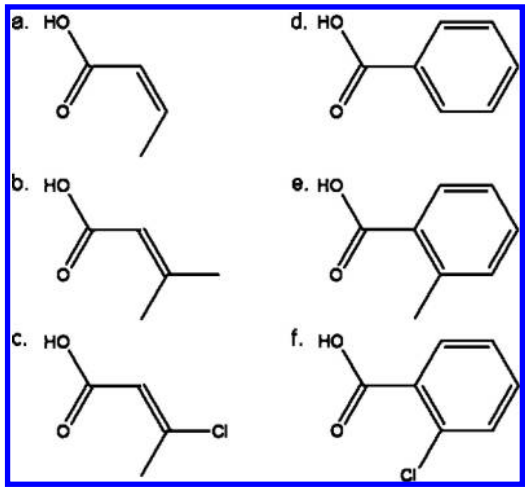

Figure 6. Close aromatic and aliphatic analogs with pK_a values (experimental:predicted) using the SMARTS pK_a method: a. but-2-enoic acid (4.68:4.56), b. 3-methylbut-2-enoic acid (5.12:5.14), c. 3-chlorobut-2-enoic acid (4.02:3.93), d. benzoic acid (4.21:3.83), e. 2-methylbenzoic acid (3.91:3.74), and f. 2-chlorobenzoic acid (2.94:2.84).

to discriminate between small differences in chemical structure. See examples shown in Figure 6. With over 700 terminal nodes in the decision tree, there is sufficient differentiating power for fine grain predictions over the pK_a range of interest. Most terminal nodes contain only two compounds, but in the cases where a decision pathway terminates at a node with more than two compounds, there appears to be lessened sensitivity to small structural differences. This must be taken with the proverbial "grain of salt," as the pK_a range of all terminal nodes is minimized by the method, such that structural modifications with less influence on the pK_a tend to group together. SMARTS pK_a is capable of high resolution caused by minor structural changes that can have a significant influence on the pK_a and tends to ignore structural changes that have minimal impact on the experimental pK_a.

The SMARTS pK_a method has two other advantages: it is fast and it provides a pK_a range for each prediction based on the maximum and minimum pK_a of the molecules at each terminal node. Speed benchmarks show that it would take less than 1 h to predict the pK_a for 1 million compounds. Also, when providing predictions for the external test set, 81% of all experimental values fell within the intervals of prediction (IoPs) determined by the terminal nodes. The width of an IoP is the maximum minus the minimum pK_a of the molecules at a terminal node, while the predicted value is the mean of the pK_as and falls somewhere between the maximum and minimum values. By extending the maximum and minimum values of each terminal node by only 0.3 pK_a units (i.e., increasing the pK_a range by 0.6), this confidence interval for an accurate prediction is increased from 81% to 100%. See Table 8 for a description of the test set compounds that fell inside the IoP widths.

Table 8. Analysis of Molecules Falling within IoPs

IoP width[a]	test set compounds	experimental values within IoP
(0,1]	130	99
(1,2]	34	39
(2,3]	14	14
(3,4]	3	3
(4,9)	4	4

[a] Interval of prediction width is in pK_a units. Extending the IoPs by 0.6 increases confidence from 81% to 100%.

CONCLUSIONS

As a measure of strength of acidity or basicity, pK_a is a major factor in chemical reactions and biological interaction of all compounds. It is relevant to physicochemical properties, such as aqueous solubility and log$D$ as well as ADME.[47] These facts, along with the immense and ever growing number of known and theorized chemical entities, make pK_a a major focal point in the drug discovery pipeline. Hence we need to continue creating new predictive models and improving the efficiency and accuracy of existing prediction methods.

Here we have presented a new predictive model for the pK_a of monoprotic compounds. Having obtained 1881 unique monoprotic compounds with their pK_a from Lange's 15th Handbook of Chemistry and the Beilstein Database, we used a novel set of SMARTS strings derived from a training set of 1693 monoprotic compounds to create a decision tree where the leaf nodes provide pK_a predictions and IoP, based on two or more training compounds identified by the respective leaf nodes. Leave-some-out (10%) cross-validation study shows a respectable Q^2 of 0.91 and RMSE of 0.80, while an external test set has $R^2 = 0.94$ and RMSE = 0.68. Based on an overall comparison to methods described in the literature and our own benchmark comparison, SMARTS pK_a outperforms previous models which have been trained on larger data sets as measured by the Pearson correlation coefficient, RMSE, and for having the fewest and least outstanding outliers.

One major difference between many prediction methods and SMARTS pK_a is that only one training set was used to derive our model, whereas other methods have combined many models specific to ionizable site types to produce their final prediction utility. It is far easier to overfit a model based on a small training set, especially when using qualitative descriptors. This leads to the major flaws inherent in any prediction method: the quality and amount of data available from which to train the model. By not breaking up the training set into subsets based on different ionizable site type, we were able to maintain the largest possible training set. Stepwise regression and the final decision tree led to the identification of 139 optimized SMARTS string descriptors. To prove that we had not overfit our model, we retrained the model 100 times, randomizing the pK_a prior to each retraining, and showed that the R^2 and RMSE of the randomized models were significantly worse than the non-randomized model.

Combining information from multiple data sets can increase the size and diversity of the training set reducing the prediction error.[48] Furthermore, with so much data being held proprietary, consensus models can lead to improved prediction accuracy and reduced overall errors. When considering the *in silico* evaluation of physicochemical properties of molecules in the pre- and postscreening stages of drug development, it is advisable to simultaneously examine multiple predictive models, as they are typically based on different training sets.

Issues of breadth and expanding the chemical space of our model are currently being addressed. We are now curating the data for over 10000 unique mono- and polyprotic molecules from Beilstein and Lange. The intent is to first expand the model to predict primary pK_a followed by a more comprehensive model capable of handling polyprotic molecules. Finally, we intend to combine SMARTS pK_a with MOE's Slog*P* to produce a utility for the prediction of log*D*.

ACKNOWLEDGMENT

We would like to acknowledge David R. Peck for the introduction to and setup of the MDL Crossfire Commander which was used to access the Beilstein database. Also special thanks to the University of Michigan's Undergraduate Research Opportunities Program students Seifu Chonde and Vincent Parr for their research efforts and assistance with the curation of data from Beilstein. We thank Dr. John Dearden for providing the test set used in his comparison of commercially available pK_a predictors. A.C.L. was supported by the Lyons fellowship of the College of Pharmacy, University of Michigan. This work was supported by the following NIH grant: P20HG003890.

Supporting Information Available: Decision tree including predictions and ranges for each node and SMILES for all compounds and the predicted pK_a values for all test compounds from SMARTS pK_a, SPARC, MARVIN, Advanced Chemistry Development (ACD)/Labs Online v. 8.03, and ADME Boxes. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational approaches to estimate solubility and permeability in drug discovery and developmental settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

(2) Avdeef, A. In *Absorption and Drug Development: Solubility, Permeability, and Charge State*; John Wiley & Sons: Hoboken, NJ, 2003; Chapter 1, pp 15−17.

(3) Hoener, B. A.; Benet, L. Z. In *Modern Pharmaceutics*; Banker, G. S., Rhodes, C.T., Ed.; Mercel Dekker Inc.: New York, 1990; pp 142−180.

(4) Wells, J. I. In *Pharmaceutical Preformulation*; Ellis Horwood Ltd.: New York, 1988; p 25.

(5) Jelfs, S.; Ertl, P.; Selzer, P. Estimation of pK_a for Druglike Compounds Using Semiempirical and Information-Based Descriptors. *J. Chem. Inf. Model.* **2007**, *47*, 450–459.

(6) Chakrabarti, S.; Southard, M. Control of Poorly Soluble Drug Dissolution in Conditions Simulating the Gastrointestinal Tract Flow. 1. Effect of Tablet Geometry in Buffered Medium. *J. Pharm. Sci.* **1996**, *85*, 313–319.

(7) Upthagrove, A. L.; Nelson, W. L. Importance of Amine pK_a and distribution coefficient in the metabolism of fluorinated propanolol analogs: metabolism by CYP1A2. *Drug Metab. Dispos.* **2001**, *29*, 1377–1388.

(8) Oprea, T. I.; Marshall, G. R. *Receptor-Based Prediction of Binding Affinities. In Perspectives in Drug Discovery and Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Great Britain, 1998; Vol. 9−11, pp 35−61.

(9) Alberati, D.; Hainzl, D.; Jolidon, S.; Krafft, E. A.; Kurt, A.; Maier, A.; Pinnard, E.; Thomas, A. W.; Zimmerli, D. Predicting and Tuning Physiochemical Properties in Lead Optimization: Amine Basicities. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4311–4315.

(10) Jamieson, C.; Moir, E. M.; Rankovic, Z.; Wishart, G. Medicinal Chemistry of hERG Optimizations: Highlights and Hang-Ups. *J. Med. Chem.* **2006**, *49*, 5029–5046.

(11) Fischer, H.; Kansy, M.; Bur, D. CAFCA: a Novel Tool for the Calculation of Amphiphilic Properties of Charged Drug Molecules. *Chima* **2000**, *54*, 640–645.

(12) Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery. 3. Modeling Blood-Brain Barrier Partitioning Using Simple Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2137–2152.

(13) Acton, G. *Toxicogenomics and Predictive Toxicology Market and Business Outlook*; Vivo Group: Concord, MA, 2004. http://www.vivogroup.com/reports.html, http://www.the-infoshop.com/study/cd25153_toxicogenomics.html (accessed Mar 14, 2008).

(14) Caldwell, G. W. Compound optimization in early- and late-phase drug discovery. Acceptable pharmacokinetic properties utilizing combined physicochemical, in vitro and in vivo screens. *Curr. Opin. Drug Discovery* **2000**, *3*, 30–41.

(15) Clark, J.; Perrin, D. D. Prediction of the Strengths of Organic Bases. *Q. Rev. Chem. Soc.* **1964**, *18*, 295–320.

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

(17) Dixon, S. L.; Jurs, P. C. Estimation of pK_a for Organic Oxyacids Using Calculated Atomic Charges. *J. Comput. Chem.* **1993**, *14*, 1460–1467.

(18) Hilal, S. H.; Karickhoff, S. W. A Rigorous Test for SPARC's Chemical Reactivity Models: Estimation of More Than 4300 Ionization pK_as. *Quant. Struct.-Act. Relat.* **1995**, *14*, 348–355.

(19) Lee, P. H.; Ayyampalayam, S. N.; Carreira, L. A.; Shalaeva, M.; Bhattachar, S.; Coselmon, R.; Poole, S.; Gifford, E.; Lombardo, F. In Silico Prediction of Ionization Constants of Drugs. *Mol. Pharm.* **2007**, *4*, 498–512.

(20) Xing, L.; Glen, R. C.; Clark, R. D. Predicting pK_a by Molecular Tree Structured Fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 870–879.

(21) Soriano, E.; Cerdán, S.; Ballestros, P. Computational determination of pK_a values. A comparison of different theoretical approaches and a novel procedure. *J. Mol. Struct. (Theochem)* **2004**, *684*, 121–128.

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

(23) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and Original pK_a Prediction Method Using GRID Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, *47*, 2172–2181.

(24) Gieleciak, R.; Polanski, J. Modeling Robust QSAR. 2. Iterative Variable Elimination Schemes for CoMSA: Application for Modeling Benzoic Acid pK_a Values. *J. Chem. Inf. Model.* **2007**, *47*, 547–556.

(25) Ghasemi, J.; Saaidpour, S; Brown, S. D. QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. *J. Mol. Struct.* **2007**, *805*, 27–32.

(26) http://www.chemsilico.com/CS_prpKa/PKAexp.html (accessed Mar 11, 2008).

(27) Schüürmann, 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.* **1998**, *102*, 6706–6712.

(28) Eckert, F.; Klamt, A. Accurate Prediction of Basicity in Aqueous Solution with COSMO-RS. *J. Comput. Chem.* **2006**, *27*, 11–19.
(29) Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M. E. First Principles Calculations of Aqueous pK_a Values for Organic and Inorganic Acids Using COSMO-RS Reveal an Inconsistency in the Slope of the pK_a Scale. *J. Phys. Chem. A* **2003**, *107*, 9830–9386.
(30) Pompe, M. Variable connectivity index as a tool for solving the 'anti-connectivity' problem. *Chem. Phys. Lett.* **2005**, *404*, 296–299.
(31) Pompe, M.; Randić, M. Variable Connectivity Model for Determination of pK_a Values for Selected Organic Acids. *Acta. Chim. Slov.* **2007**, *54*, 605–610.
(32) *MOE: Molecular Operating Environment, version 2007.0902*; Chemical Computing Group: Montreal, Quebec, Canada, 2007.
(33) Dean, J. A. In *Lange's Handbook of Chemistry*, 15th ed.; McGraw-Hill: New York, 1999; Chapter 8, pp 8.24−8.72. http:// www.knovel.com (accessed Apr 2007).
(34) *MDL CrossFire commander, version 7*; Elsevier MDL: San Leandro, CA, 2007.
(35) *ChemDraw Ultra, version 10*; CambridgeSoft: Cambridge, MA, 2007.
(36) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comp. Sci.* **1992**, *32*, 244–255.
(37) Daylight Chemical Information Systems Inc. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html#RTFToC35 (accessed July 8, 2008).
(38) Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science* **1960**, *132*, 1115–1118.
(39) Lee, A. C.; Shedden, K.; Rosania, G. R.; Crippen, G. M. Data Mining the NCI60 to Predict Generalized Cytotoxicity. *J. Chem. Inf. Comput. Sci.* **2008**, *48*, 1379–1388.
(40) SPARC Performs Automated Reasoning in Chemistry v4.2. http://ibmlc2.chem.uga.edu/sparc/ (accessed May 7, 2008).
(41) Szegezdi, J.; Csizmadia, F. New method for pK_a estimation. Proceedings of the eCheminformatics 2003 - Virtual Conference and Poster Session, Zeiningen, Switzerland, 2003; Hardy, B., Ed.; Douglas Connect: Zeiningen, Switzerland, 2003.
(42) ChemAxon. Marvin and Calculator Demo. http://www.chemaxon.com/demosite/marvin/index.html (accessed May 7, 2008).
(43) Advanced Chemistry Development ACD/Labs Online (I-Lab). http://www.acdlabs.com/ilab/ (accessed May 7, 2008).
(44) ADME/Tox WEB. http://pharma-algorithms.com/webboxes/ (accessed July 9, 2008),
(45) Dearden, J. C.; Cronin, M. T. D.; Lappin, D. C. A comparison of commercially available software for the prediction of pK_a. *J. Pharm. Pharmacol.* **2007**, A7, Suppl. 1.
(46) Meloun, M.; Bordovská, S. Benchmarking and validating algorithms that estimate pK_a values of drugs based on their molecular structures. *Anal. Bioanal. Chem.* **2007**, *389*, 1267–1281.
(47) Wan, H.; Ulander, J. High-throughput pK_a screening and prediction amenable for ADME profiling. *Expert Opin. Drug Metab. Toxicol.* **2006**, *2*, 139–155.
(48) Tetko, I. V.; Bruneau, P.; Mewes, H.-W.; Rohrer, D. C.; Poda, G. I. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today* **2006**, *15/16*, 700–707.

CI8001815