PERSPECTIVE

Predicting pK_a

Adam C. Lee and Gordon M. Crippen*

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

Received June 11, 2009

One of the most important physicochemical properties of small molecules and macromolecules are the dissociation constants for any weakly acidic or basic groups, generally expressed as the pK_a of each group. This is a major factor in the pharmacokinetics of drugs and in the interactions of proteins with other molecules. For both the protein and small molecule cases, we survey the sources of experimental pK_a values and then focus on current methods for predicting them. Of particular concern is an analysis of the scope, statistical validity, and predictive power of methods as well as their accuracy.

"A man would do nothing if he waited until he could do it so well that no one would find fault with what he has done." John Henry Cardinal Newman.

1. INTRODUCTION

When a weak acid dissociates according to the schematic equation $HA \rightleftharpoons A^- + H^+$, the equilibrium constant is $K_a = [A^-][H^+]/[HA]$, which is conveniently rearranged into the Henderson-Hasselbach equation

$$pH = pK_a + \log_{10}\frac{[A^-]}{[HA]} \quad (1)$$

where $pK_a = -\log_{10}K_a$, analogous to the definition of pH. Of course, one can describe the protonation of a weak base in the same terms. When the weak acid/base is titrated against a strong base/acid, the titration curve is a plot of pH as a function of equivalents of added titrant, and the curve shows a characteristic inflection when $pH = pK_a$. Experimental determination of pK_a is straightforward as long as there is only one titratable group involved. When the molecule in question has n protonatable sites, there are 2^n microspecies (particular combinations of protonations at the n sites) to be considered and $2^n - 1$ independent micro-pK_as (equilibrium constants between two microspecies), so that at any given pH there is an equilibrium mixture of some of these microspecies at non-negligible concentrations. Particularly if some of the micro-pK_as have similar values, the titration curve may show only a few inflections corresponding to macro-pK_as between the macrospecies that are the predominant mixtures of microspecies. Ullman points out that a maximum of $n^2 - n + 1$ parameters can be extracted from a titration curve of all ionizable sites, and since $2^n - 1 > n^2 - n + 1$ for $n > 3$, it is not possible to obtain micro-pK_as for polyprotic acids with more than 3 independent ionizable sites without additional information or special assumptions.[1] Expressed in these terms, there is a concern in predicting the pK_as for all microspecies.

This review focuses on pK_a predictions for molecules in an aqueous environment with the typical pK_a range of interest lying between that of the hydronium ion (−1.74) and the hydroxide anion (15.74), as the pH of blood is generally regulated to the range of 7.35−7.45 and the pH range encountered in the normal human gastrointestinal tract is 1−8.[2,3] Most of the available experimental data were obtained at 25 °C in aqueous solutions having ionic strength less than 0.1 M. Obviously 37 °C would better match human *in vivo* conditions. Many of the predictive methods discussed herein could be reparameterized for other environments, considering both temperature and solvent, as long as sufficient data were available for training and validating the model.

2. PROTEINS

Calculating the macro-pK_as for globular proteins may seem easy because there are only a few different kinds of protonation sites involved, particularly Asp, Glu, and His side chains. The problem is nontrivial because particularly the somewhat buried acidic and basic groups have long been known to have pK_as that are sometimes substantially shifted from what is observed in oligopeptides. Worse yet, there are so many protonation sites on most soluble proteins that experimentally determining the macrospecies is challenging. Prediction methods are therefore developed on the basis of several sites each of rather few well-studied proteins, which is something of a concern for validating the methods. The standard prediction task takes as input not only the amino acid sequence, i.e. the covalent structure of the molecule, but also the experimentally determined three-dimensional structure typically from X-ray crystallography, which can be a problem when conformational flexibility is important.

2.1. Experimental Data. The favored method for determining the pK_as of individual ionization sites on proteins is by NMR.[4] It can also be used for small molecules in both

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

10.1021/ci900209w CCC: $40.75
Published on Web 08/24/2009

Table 1. Proposed Null Models

group	1943[a]	1967[b]	1973−4[c]	1978[d]	1993[e]	2006[f]	2007[g]	2009[h]
α-carboxyl	3.0−3.2	3.8	3.3	—	3.5−4.3	3.67	—	—
Asp	3.0−4.7	4.0	3.91	3.9	3.9−4.0	3.67	3.47	3.49
Glu	4.4	4.4	4.145	4.2	4.3−4.5	4.25	4.16	4.39
His	5.6−7.0	6.3	(6.8)	6.9	6.0−7.0	6.54	6.30	6.6
α-amino	7.6−8.4	7.5	8.1	—	6.8−8.0	8.00	—	—
Cys	9.1−10.8	9.5	—	—	9.0−9.5	8.55	4.67	—
Tyr	9.8−10.4	9.6	(10.0)	10.2	10.0−10.3	9.84	9.90	—
Lys	9.4−10.6	10.4	10.47	11.0	10.4−11.1	10.40	10.04	9.78
Arg	11.6−12.6	12.0	—	—	12.0	—	—	—
RMSE[i]	1.44	1.54	1.48	1.56	1.46	1.46	1.43	1.42

[a] Data taken from ref 18. [b] Determined using various model compounds at 25 °C.[14,15] [c] Determined with Gly-Gly-X-Gly-Gly pentapeptides by ^{13}C NMR with unblocked termini,[16,17] while values in parentheses were taken as reported in ref 18. [d] Determined in Gly-Gly-X-Ala tetrapeptides with unblocked termini by ^{13}C NMR at 35 °C.[19] [e] Data taken from Creighton.[20] [f] Determined in Ala-Ala-X-Ala-Ala pentapeptides with blocked termini using potentiometry.[18] [g] Mean values taken from 475 different sites from 73 proteins.[21] [h] Values for each residue type were obtained by minimizing the RMSE in a benchmark set of 80 residues. [i] In all cases where a range is indicated, the midpoint was taken and used to compute the RMSE for a benchmark set of 80 residues.

aqueous and mixed solvents. It is applicable to proteins both in their folded and in their denatured states[5] and thus is useful in determining the correct order of deionization. The chemical shift of an assigned proton near the ionization site (in terms of covalent bonds or even through space) varies with pH, so the chemical shift vs pH curve is fitted to a simple model involving three adjustable parameters: the chemical shifts in the protonated and deprotonated states and the pK_a.[6] The main concern is that the chemical shift of a proton can also be influenced by other environmental factors, such as nearby solvent or other parts of the protein or by multiple conformations interconverting rapidly on the NMR time scale. It is also not possible to accurately determine the pK_a of residues that are not fully titrated at low pH (for example Glu73, Asp93, and Asp101 of barnase), as no true baseline representing the protonated state can be established. Furthermore, experimentally determining the pK_as for coupled ionizable residues is difficult, as fitting ideal titration curves to NMR chemical shift data of these residues leads to poorly resolved pK_as (for example Asp54, Asp75, and Asp86 of barnase).[7] Sometimes these questions can be resolved by difference titrations where the sequence of the protein is changed to eliminate sources of confusion.[8]

Currently, the Protein pK_a Database (PPD) exists as a free data source, providing over 1400 experimental data points taken from literature for the ionizable amino acid side chains in proteins as well as N and C termini. The vast majority of the available measurements are for Asp, Glu, His, and Lys. Very little data exist for Arg, as titrations at high pH tend to denature proteins.[9]

2.2. Predictive Methods. *2.2.1. The Null Model.* The simplest method is the trivial null model, where the experimental pK_a of the amino acid side chain in some oligopeptide is taken as the predicted value for all amino acid residues of the same type. Several variations on this theme are shown in Table 1. Many earlier works unintentionally demonstrate, or at least mention, that the performance of the null model can be difficult to beat. This is especially true when the proteins being considered have a preponderance of ionizable residues exhibiting small pK_a shifts. In these instances, the null model can be expected to have a root-mean-square error (RMSE) less than or equal to 1.0.[10−12] In protein pK_a prediction, outperforming the null model is essential, as the residues showing the most significant pK_a shifts are often the most interesting, for example buried residues, residues participating in salt bridges, or residues found in enzyme active sites.[13] The overall success of the null model is mainly due to the available data, which are dominated by surface residues and other residues that do not participate in strong intramolecular interactions.[11] Nonetheless, null model values are an important starting point for most prediction methods.

In an attempt to determine which set of null values represents a good starting point, we considered a benchmark set of 80 interesting residues from 30 different proteins used to compare some known macromolecular pK_a prediction utilities.[11] For each of four residue types (Asp, Glu, His, and Lys), the data set consists of 20 pK_a measurements, 10 having pK_a shift less than 1.0 and the other 10 having pK_a shift greater than or equal to 1.0. The data set is diverse, as it consists of a more balanced collection of residues having varied solvent exposure with over half exhibiting large pK_a shifts, and it includes residues found in active sites as well as structurally important regions.[22] There are arguments both for[12,21] and against[23,24] the correlation between solvent exposure and pK_a shift. However, in this case focus is placed on the diversity of the local environment. RMSE was calculated for each of the null models applied to the benchmark data set, as shown in Table 1. Of course our 2009 null model performs the best, since it was trained by a least-squares fit to the very benchmark data set. It is interesting to note that the most commonly used or traditional set of intrinsic values comes from the 1967 set, compiled by Nozaki and Tanford.[15] The traditional set is outperformed by almost every other proposed set of null values. The best and least controversial data sets which can be used as intrinsic values are those in the columns labeled 1943, 1993, and 2006.[14,18,20] All proposed pK_a values from these columns are based on experimental measurements and come close to matching the results from the optimized values as well as the values acquired by taking the mean value for each residue type over a large set of curated experimental values from 73 proteins.[21]

When developing any predictive model, care should be given to the separation of training and test data. That is, any protein residues used to parametrize a model should not be used for validation purposes. In this regard, some have taken a cross-validation approach to model optimization, commonly used in cases where there is a lack of available data. In the

Table 2. Software for pK_a Calculations Using FDPB Methods[25]

software	URL
Delphi	wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:DelPhi
MEAD	www.scripps.edu/mb/bashford/
PEP	casegroup.rutgers.edu/
UHBD	mccammon.ucsd.edu/uhbd.html
ZAP TK	www.eyesopen.com/products/toolkits/modeling-toolkits.html
APBS	apbs.sourceforge.net/
H++	biophysics.cs.vt.edu/H++/
HYBRID	gilsonlab.umbi.umd.edu/software1a.html
KARLSBERG	agknapp.chemie.fu-berlin.de/agknapp/
KARLSBERG+	
MCCE	www.sci.ccny.cuny.edu/~mcce/
PCE webserver	bioserv.rpbs.jussieu.fr/Help/PCE.html
WHAT IF	swift.cmbi.kun.nl/whatif/
Wade lab scripts	projects.villa-bosch.de/mcm/software/pka

cross-validation approach, the null values can be set to the mean pK_a values, or some other statistical variant thereof, for each residue type in the training set. In one example, the pK_a values for the ionizable residues in each individual protein of a set of 27 proteins were predicted using the data from the other 26 proteins. The resulting RMSE of 0.853 was a significant improvement over the traditional null model's RMSE of 1.069. The study also showed that the results of the optimized null model surpassed those of a Poisson−Boltzmann equation based approach for a data set of 122 ionizable residues of five different types that had been previously evaluated by the same group.[12] Clearly, optimizing the intrinsic null values is simple and may prove beneficial in more elaborate models.

2.2.2. Electrostatic Models. In order to account for deviations from the null model due to the spatial arrangement of the rest of the protein and the general solvent, one must estimate the free energy difference between the protonated and deprotonated states of the ionizable group in question. Most protein pK_a prediction methods are based on solving the linearized Poisson−Boltzmann equation using atomic partial charges from a molecular modeling force field. Typically, the three terms considered in calculations using finite difference Poisson−Boltzmann (FDPB) methods are the Born solvation energy, the energy due to Coulomb interactions of the ionizable group in question with fixed partial charges of the protein atoms, and pairwise Coulomb interactions of it with other titratable sites of the protein. For those interested in the underlying parametrization of an FDPB model, we direct the reader to a work by Fitch and García-Moreno[25] where the basic protocols are described for implementing the University of Houston Brownian Dynamics (UHBD) software for pK_a calculations developed by McCammon and colleagues.[26] This work serves as an excellent review of FDPB methodology, provides a list of downloadable FDPB software shown in Table 2, and discusses model parametrization including the selection of the most appropriate dielectric constant for electrostatic calculations.[25] Bashford provided a comprehensive review covering the major macroscopic electrostatic models and approximations that are used to calculate the relative energies of protonation states and the pH-titration properties of ionizable groups in proteins as well as their applications to small molecules.[27] The methods discussed are rooted in solving the Poisson−Boltzmann equation, which has been thoroughly discussed in the literature.[28−30]

When using FDPB models to calculate pK_a, two other problems need to be considered: electrostatic and thermodynamic. The electrostatic problem involves understanding the dielectric properties of the different phases in a protein-water system, thereby making it difficult to accurately calculate pK_a for all ionizable sites of a protein without considering the environment of the individual ionizable sites. For instance, buried ionizable residues can experience substantial pK_a shifts when compared to the null value for the respective residue type. It has been shown that regions with low dielectric medium provide an unfavorable environment for inducing a net charge.[31−33] Therefore, it would be expected that the pK_a of an aspartic acid residue located in a region of moderately low polarity would be shifted higher due to the absence of strong ionic and hydrogen bonding interactions found in an aqueous environment. On the other hand, buried environments exist such that a carboxylate group is favored over the unionized form, such as those in close proximity to another polar or positively charged group which lowers the pK_a relative to the null value. This is common in protein functional regions such as active sites.[34−36] Finally, the dielectric constant is inversely proportional to the electrostatic energy and can thereby significantly affect the calculated energies in continuum electrostatic calculations as well as the pK_a shift of ionizable residues. Even though the definition and parametrization of the dielectric constant is model dependent,[63] it is considered the most important adjustable parameter when performing continuum electrostatic calculations.[25] The statistical thermodynamic problem involves making FDPB calculations to solve the state of ionization and electrostatic energy for each ionizable site. The issue is purely combinatorial and limits the application of FDPB models to smaller proteins. The number of calculations increases exponentially with the number of ionizable sites in a protein. As discussed earlier, a protein with N ionizable residues can have up to 2^N microstates. The practical computational limit is thought to be 30 ionizable groups.[25] Treating larger proteins undoubtedly would require approximations.

Still other problems exist, such as protein flexibility and partial charge parametrization. One of the most common and potentially problematic simplifications used when applying Poisson−Boltzmann calculations is the assumption that protein structures are rigid and identical to the crystal structure. It is known that ions can cocrystallize with a protein affecting the fine details of the protein's surface structure.[13] Furthermore, this assumption places limitations on the calculations of changes in free energy when the state of an ionizable group changes.[37] Ionizable side chains of proteins in crystals grown at fixed pH have a fixed charge. In solution at fixed pH proteins are flexible to various degrees and constantly undergo conformational changes, as seen by NMR. During titration, the pH changes and alters the ionization state of the protein in solution, which may lead to different conformations. Static models based solely on the crystal structure coordinates are therefore not likely to be appropriate at all pHs.[38,39] On the other hand, considering conformational flexibility requires computationally expensive methods such as Monte Carlo sampling. Now, instead of considering 2^N ionizable states in the case of rigid structures,

Table 3. Selection of Protein pK_a Predictors Outperforming the Null Model[e]

ref	author	year	method[a]	no. of para	training	RMSE	no. of proteins	valid[b]	residues[c]
10	Antosiewicz	1994	FDPB		60	0.89	7		DEHIKYcn
41	Antosiewicz	1996	FDPB		52	0.7	4		CDEHKYcn
42	Demchuk	1996	FDPB		48	0.5	3		CDEHKRYcn
13	Nielsen	2001	FDPB		124	0.91	9		CDEHKRYcn
43	Georgescu	2002	FDPB		166	0.83	12		DEHKYcn
44	Czodrowski	2006	FDPB/PEOE		132	0.88	9		DEHKY
45	Barth	2007	FDPB		31	0.38	10		DE
46	Dimitrov	1997	DH		70	0.79	6		DEHKYcn
47	Warwicker	1999	DH		53	0.72	6		DEHKRcn
48	Warwicker	2004	FDPB/DH		117	0.86	15		CDEHKRYcn
49	Sham	1997	PDLD/S-LRA		9	0.73	2[d]		DE
63	Schutz	2001	PDLD/S-LRA		11	0.31	4		DEHK
50	Sandberg	1999	PDS		40	0.83	3		DEHKRYcn
51	Mehler	1999	SCP		103	<0.5	7		DEHKRYcn
				(8)		(0.504)		ext	(DE)
52	Mongan	2004	GB		18	0.82	4[d]		DEHKY
53	Kuhn	2004	GB		69	1	9		DEHKYcn
54	Pokala	2004	GB		226	0.92	15		DEH
55	Spassov	2008	GB	1	21	0.45	1		DEHKYcn
					(310)	(0.51)	(23)	ext	
56	Khandogin	2006	GB/DH		135	0.95	10		DEHc
57	Wisz	2003	Emp	24	260	0.95	41		DEHKYcn
12	Krieger	2006	Emp	4	227	0.879	(27)	cv	DEHKYcn
21	He	2007	Emp	13	405	0.593	73		DEHKcn
						(0.775)		cv	
22	Li	2005	Emp	30	314	0.89	44		CDEHKRYcn
					(77)	(0.56−0.97)	(4)	ext	

[a] FDPB - finite difference Poisson−Boltzmann; DH - Debye−Hückel; PDLD/S-LRA - protein dipoles Langevin dipoles/solvation linear response approximation; PDS - position dependent screening; SCP - screened coulomb potential; GB - generalized Born; Emp - empirical. [b] Validation method used: cross-validation (cv) or external data set (ext). [c] Modeled amino acid residue types in single character notation plus c for C-terminus and n for N-terminus. [d] Lysozyme crystal structures having different conformations were considered. [e] Information in parentheses pertains to external data used for validation.

one could consider up to $(2M)^N L^K$ possible states where each of N ionizable residues has M potential conformations and K nonionizable groups have L conformations.[40] This is further complicated by the use of molecular modeling force fields, which may lack proper parametrization or provide less than adequate partial charge assignments. Alternatively, a Quantum Mechanical (QM) or mixed Quantum Mechanical Molecular Modeling (QM/MM) approach would rely on significantly fewer empirical parameters but would be far more demanding of computational resources. In spite of the many challenges, Table 3 shows the continual improvement of FDPB and other methods for the prediction of protein pK_a.

In the 1990s, pK_a predictors tended to optimize their parameters in order to beat the null model.[41,42,58−60] It was quite common to ignore external validation and focus only on selecting the optimal dielectric constant(s) to optimize the RMSE using experimental pK_a data for the ionizable sites of a few proteins, such as hen egg white lysozyme (HEWL), ribonuclease A (Rnase A), and bovine pancreatic trypsin inhibitor (BPTI). The Antosiewicz and the Demchuk models improved their calculations by adjusting a single parameter, the dielectric constant. Antosiewicz used a fixed dielectric constant of 20 for best results.[10] The Demchuk model assumed different local dielectric constants (ranging from 15 for buried residues to 80 for highly solvated residues) in the neighborhood of each ionizable site in order to better fit the experimental data.[42]

When developing any model, being able to fit the experimental training data is essential. Even though the FDPB models are based on solid physics, the dielectric constant needs to be tweaked in order to improve performance, much as the variables in empirical linear regression models are adjusted to fit the training data. A model's predictions are not likely to be as good as its fit to the training data. One must keep in mind that an unvalidated model tells us very little about that model's ability to predict new data. While the fixed dielectric constant model was not presented with external data for validation, the creators of the multiple dielectric constant models included a small external test set which was used to compare the ability of both models to predict the pK_as of 12 buried histidine residues from triose phosphate isomerase (TIM).[42] The RMSE for the fixed dielectric was 0.40, and for the multiple dielectric it was 0.42, compared to 1.26 for the null model. In Table 3 the multiple dielectric model shows better fits than the fixed dielectric model, but the multiple local dielectric constants were freely varied throughout for best results, so no fair comparison can be made. Second, as the external test set is not diverse, one cannot draw any general conclusions about either model's ability to predict new data.

When considering the protein dielectric constant, ε_p, for use in electrostatic calculations, it is worth noting that ε_p depends on the method and system used to define it.[26,61,62] Furthermore, the best value for ε_p in modeling electrostatic effects has been found to have little to do with the protein dielectric constant but rather is a measure of the electrostatic interactions which have not been included explicitly in the model,[63] such as conformational variations due to flexibility, specific water binding,[64] or proton/hydrogen-bond network relaxation.[65]

FDPB methods have made strides by incorporating protein flexibility[40,66] and solvation models[67] and by improving efficiency through Monte Carlo sampling of the many microstates of a protein.[13] 329 data points were calculated using Kieseritzky's PACs method,[68] built on the Karlsberg+ framework, which combines continuum electrostatics with multiple pH adapted conformations selected by Monte Carlo sampling. While the overall RMSE failed to beat the null model, improvements in accuracy were found for the residues having a pK_a shift greater than or equal to 1.0. Also, the multiconformational continuum electrostatics approach of Georgescu, Alexov, and Gunner has been included in recent surveys and found to perform well on residues experiencing strong pK_a shifts.[11,43,68,69] Some other FDPB methods make use of significantly larger training sets, by including data from other proteins, and have been implemented as Web applications, such as WHAT IF.[13,70]

Alternative methods used to approximate FDPB calculations include Partial Dipole Langevin Dipole, Debye−Hückel, Generalized Born, charge screening based methods, and their combinations.

Warshel and colleagues have developed the Protein Dipole Langevin Dipole (PDLD) method, in which protein dipoles are modeled.[63,49] This method tends to relax the atom-centered partial charge assumption and treats protein relaxation in a microscopic framework using linear response approximation (LRA) allowing for structural reorganization during charge formation. The net effect of PDLD/S-LRA is reduced reliance on ε_p, such that less variance of the parameter is required in order to accurately explain pK_a. Improved accuracy is seen when using both ε_p and ε_{eff}, the effective dielectric constant of the solvent, as adjustable parameters when fitting a set of 11 protein side chains experiencing significant pK_a shift, ranging from −4.9 to 5.3 log units.[63] According to Schutz and Warshel, ε_p and ε_{eff} are model dependent scaling factors. Having less variance between parameters is a big plus, as the optimal values for other models (Generalized Born, Tanford and Kirkwood, and modified Tanford and Kirkwood)[71,72] ε_p in their survey ranged from −80 to 80. Unfortunately, there is no way to foresee what values should be used for new data, and the predictive power of these models was not evaluated.

The Debye−Hückel approximation for electrostatic pair interactions has been used individually and in combination with FDPB calculations to model pK_a. Dimitrov[46] showed how it could be used as a standalone alternative to FDPB methods with superior results for the same six proteins considered by Antosiewicz. Warwicker found that the Debye−Hückel method matched the overall pH-dependent stability, while the FDPB method provided more accurate results for active-site groups. Combining both methods provided a computational framework for distinguishing solvent-accessible groups from buried groups. In doing so, significant improvements were found for a larger set of residues than considered in previous works using only the Debye−Hückel method.[47,48]

In search for a simple, less time-consuming way for modeling electrostatic interactions in proteins, Sandberg described a distance and position dependent screening technique for the electrostatic potential. The goal was to calculate pK_as by applying this technique in conjunction with a Monte Carlo algorithm to speed up protein molecular dynamic simulations. While the results were slightly less accurate than those using FDPB calculations with the dielectric constant of 20, execution time of the algorithm was 2 orders of magnitude faster than the traditional grid based Poisson−Boltzmann calculations, with one pK_a calculation every 10 s compared to 30 min, respectively.[50]

In order to deal with significant prediction errors for the pK_a of specific residues by various FDPB methods, Mehler introduced sigmoidally screened coulomb potentials and considered microenvironment hydrophobicity, based on the hypothesis that the key factor responsible for pK_a shift is the protein microenvironment around each ionizable residue.[51] By considering the log$P$ contributions of groups of atoms very near the ionizable sites, they improved their fit to the training data, which in this case consisted of 103 ionizable residues from 7 proteins. A total of 8 Asp and Glu residues from turkey ovomucoid inhibitor domain 3 and the aspartyl dyad of HIV protease were used to validate the method, having an RMSE of ∼0.50. While the fit to the training data was an improvement, the test set was not diverse, and it is not possible to tell how this model will perform on residues without carboxyl groups. Instead of using the Rekker fragmental hydrophobic constants,[73,74] the authors suggested future results might be improved by using more recent atomistic hydrophobicity values. Perhaps Slog$P$ would help.[75] One might also consider combinations of other atomistic descriptors, such as the ratio of hydrogen bond donors or acceptors to carbon atoms within an ellipsoid of some radius surrounding the ionizable site.

The generalized Born model (GB) is an approximation of the Poisson−Boltzmann equation, and it can efficiently describe the electrostatics of molecules in an aqueous environment by implicitly representing the solvent as continuum with the dielectric properties of water and thus reducing the computational demand associated with molecular dynamics (MD) simulations. The basic idea behind the GB model is to assign each atom an effective radius, such that the solvation free energy can be calculated using the Born formula. Therefore, it is important to accurately calculate the Born radii, when using any GB model.[76] Efficiency is achieved by describing the instantaneous solvent dielectric response, which eliminates the need for equilibration of water in explicit water simulations. Furthermore, as the GB model corresponds to solvation in an infinite volume of solvent, it avoids artifacts associated with replica interactions in periodic system.

Mongan describes a method used to evaluate four different crystal structures of HEWL applying implicit solvent GB electrostatics and performing MD at constant pH with periodic Monte Carlo sampling of protonation states.[52] In contrast to most electrostatically based methods, this model was shown to be independent of the starting crystal structure.

Similar to the FDPB methods, the dielectric constant is used as an adjustable parameter. Kuhn used a molecular mechanical MM/GBSA approach, which is among the most commonly used implicit solvent model combinations and typically used for calculating biomolecular binding free energies.[53] In MM/GBSA, the GB model is augmented by a term representing the hydrophobic solvent accessible surface area (SA). Kuhn et al. question the overall accuracy of the GB method noting that in theory, complex electrostatic interactions involving several charges and electric dipoles

in close proximity should be better handled by Poisson−Boltzmann continuum solvation calculations.[53]

Prior to Pokala's publication of EGAD (Egad! A Genetic Algorithm for Protein Design) in 2004,[54] the published GB methodologies considered relatively few macromolecular pK_a data points for training and validation. EGAD is particularly attractive, as electrostatic calculations are reported to be performed 6 orders of magnitude faster than FDPB methods. This increase in speed can be attributed to an approximation of the Born radii. By applying these approximations to the GB continuum dielectric model and extending the methodology to approximate solvent accessible surface area, EGAD method was used to provide calculations for 226 ionizable groups from 15 proteins with similar accuracy to other GB models of the same time. Here, it is noteworthy that a subset of five proteins was used to parametrize the model and identified the optimal $\varepsilon_p = 8$. On the other hand, the predictive power of the model on new data is questionable. First, the number of data points used to parametrize the model was not disclosed. More importantly, statistics were provided for all 226 together, instead of independently evaluating the training and test sets. Finally, due to overall lack of data only the Asp, Glu, and His side chains were considered. Data for a small number of Lys measurements, compared to the other residues, were omitted as the inclusion can significantly increase the correlation coefficient, while minimally affecting the RMSE, as discussed in section 5.

Khandogin and Brooks presented a first principles model based on continuous constant pH molecular dynamics simulations, utilizing replica-exchange protocol for enhanced conformational and protonation state sampling.[56] The method is based on a GB implicit solvent model, which is modified by an approximate DH screening function. The DH screening function is used to account for salt effects. One of the more interesting features of this method is the scaling of the dielectric constant based on the DH length, instead of simply adjusting ε_p to empirically find the best fit for the data. RMSE for proteins with ionizable residues exhibiting low pK_a shift was approximately 0.6, whereas it was approximately 1.0 for the proteins with more highly shifted residues.

Spassov uses GB approximations with an iterative mobile clustering (IMC) approach to calculate the equilibria of protons binding to titration sites in proteins.[55] IMC is used to halt the exponential growth of GB calculations when considering conformational flexibility. Here, binding and conformational states are fully enumerated for an ionizable site within a local cluster of ionizable sites. Ionizable sites outside the cluster are treated by mean field approximation. The procedure continues to iterate through the list of all ionizable sites, repeating the calculations and using the results from previous iterations outside of the present cluster for the mean field terms of the current iteration until some convergence criteria is met.[77] This method has been incorporated into the Accelrys Discovery Studio. Not only does it appear to be highly accurate with a low RMSE, approximately equal to 0.50, but also it considers the largest external set of test data of all the GB models in this review. It is of note that this model trained its single adjustable parameter, ε_p, only on HEWL (2lzt) and validated the model on over 300 external data points from 23 proteins with RMSE of approximately 0.5. A survey[55] of 105 ionizable residues from 7 proteins showed improved accuracy for this model over the top methods in the other classes mentioned in this review.[22,43,48,51,56] A close inspection of the data set used in the survey revealed that approximately 20% of the residues were shifted more than one pK_a unit. Evaluating the data set with the null model proposed by Thurkill[8] resulted in an RMSE and mean absolute error of less than 0.80 and 0.70, respectively, and a maximum error of 2.43. While the IMC method posted the best results, four of the other five methods surveyed had lower RMSE than the null model. It is rather curious that the IMC model exhibited poorer performance on its own training set, residues from hen egg white lysozyme (2lzt), than five of six of the other proteins considered in the survey. Typically, it is expected that a model's performance decreases when used to evaluate external test data. However, this may be explained, as 2lzt has a larger proportion of highly shifted residues than most of the other proteins considered. Apparently, the IMC approach to handling conformational flexibility is responsible for the high accuracy reported by this model.

2.2.3. Empirical Models. Wisz used four model equations to determine 24 independent parameters, which could be used to simulate the electrostatic interactions in proteins. Monte Carlo methodology was used to achieve convergence for all 24 parameters based on titration curves derived at 1 pH unit intervals from 0 to 14 using the model equations.[57] The training set consisted of 260 ionizable groups of which over 20% of the residues had pK_a shift greater than or equal to 1 unit. In order to investigate the stability of the parameters, additional rounds of Monte Carlo simulations were run, but no true validation was performed on external data.

Krieger published a method in which the electrostatic potential is evaluated using Ewald summation. Ewald summation is fast and can be used to monitor pK_a shifts during MD simulations and effectively handles periodic crystal environments.[12] Naive electrostatic calculations in periodic systems may diverge. Here Ewald summation allows for simplification within the periodic environment by combining a rapidly converging short-range variable with a long-range term evaluated in reciprocal space. The particle-mesh Ewald algorithm, standard in many MD programs, was used to identify models based on three and four parameters. 227 ionizable sites from 27 proteins were considered, and a leave-one-protein-out cross-validation was performed. Both three and four parameter models outperformed the null method, which also had RMSE less than 1.0. It is worth noting that the null model was optimized using a similar cross-validation technique where the mean was taken of the respective pK_a values of the amino acids of 26 proteins and used to predict the pK_a of the remaining protein. The optimization technique improved null model predictions by over 0.2 RMSE units.

From an empirical standpoint, it is highly unlikely that a model trained on data, having relatively low variance from their respective null values, could accurately predict the pK_a for ionizable sites having significantly shifted pK_a values. This is especially true if the pK_a shift is due to an environment which was not considered by the model. This is best explained by the statistical optimization performed by He et al.[21] Here a significantly larger data set was considered by mining the PPD, such that 1122 pK_a values belonging to 667 ionizable sites could be utilized for training and validation purposes. Structures were validated against the PDB.[78] After curation 475 unique sites from 73 proteins

were accepted. According to He et al. "In the data set, the pK_a values of 46 sites are unusual because of physical or chemical factors, such as salt bridges or disulfide bridges (Table II). To obtain reasonable parameters, these data were excluded from the fitting procedure and were predicted using parameters obtained from the remaining sites.". In total 13 parameters were considered using multiple linear regression, where each parameter represented one or more amino acid types. pK_a shift was induced by residue−residue interaction determined by the amino acids surrounding the C^{α} of the ionizable residue within a sphere of some radius (minimum RMSE at 11 Å). Based on the data set used to train the model, it is obvious why the model performed reasonably well on the 405 residues with low pK_a shift (RMSE = 0.775, using 6-fold cross-validation) and quite poorly on the external data, composed of the 46 unusual ionizable sites, which also had the highest pK_a shifts (RMSE = 4.258).

Seemingly, the most accepted empirical method for predicting protein pK_a in the literature today is PROPKA.[22] The most thorough surveys, often entitled benchmarks, include PROPKA performance as the method to beat.[11,55,68,69] In each survey, PROPKA's RMSE is less than 1.0, including those that consider highly shifted residues. PROPKA's origins began with quantum mechanical/molecular mechanical studies by Jensen, where analyses of pK_a determinants led to a set of quantitative structure property relationships forming the basis of PROPKA.[79] The model was trained on 314 experimental values using 30 parameters, 20 of which are distance related. Validation was performed on four proteins not appearing in the test set where the RMSE of the predictions for each individual protein ranged from 0.56 to 0.97 with a maximum predicted error of 2.0 units. The original release of PROPKA version 1.0 was noted to ignore pK_a shifts caused by ligands, ions, and waters interacting with the protein. More recently, version 2.0 incorporates protein−ligand interactions affecting ionizable groups as well as providing predictions for the ionizable groups of ligands in the protein environment.[80] Calculations are usually complete for an entire protein in a matter of seconds. Desolvation, hydrogen-bonding, and charge−charge interactions are considered in calculating the shifts as well as groups with a fixed charged, such as Zn^{2+}. Current limitations noted by the developers include the assumption that intraligand interactions are included in the pK_a model value, while both pK_a shifts due to interligand interactions and the effects of side chain motion as well as other conformations are not considered.

Desirable qualities for empirical models are large diverse training sets fitted to as few parameters as possible and high predictive accuracy on a diverse data set that was not used for training purposes.

3. SMALL MOLECULES

Predicting the pK_as for small molecules is a substantially different problem than for proteins, based on their respective isolated environments. There are far fewer microstates to consider, and most three-dimensional effects are commonly neglected, such as the local electrostatic field and degree of solvent exposure. However, the range of chemical structures is far greater, so care must be taken when assessing the diversity of training and test sets. It has been shown that understanding the site-specific charges and concentrations of the microspecies can allow for more realistic predictions regarding a molecule's pharmacokinetic behavior.[81] Unfortunately, most models are simplifications and provide only macrospecies predictions due to the limitations of the data available for training. However it is possible to make predictions regarding the microspecies by interpolating the approximated titration curves based on the macrospecies. At least three available applications, ChemAxon's MARVIN[82] and ACD/PhysChem Suite,[83] and Simulations *Plus* ADMET Predictor,[84] provide microspecies predictions for small molecules.

3.1. Experimental Data. *3.1.1. Data Curation.* While a great deal of experimental pK_a data can be found in the literature and 'in house', their reliability is sometimes questionable. A large portion of the literature containing data on small molecules is recorded in the Beilstein database and is accessible using the MDL Crossfire Commander.[85] Lange's Handbook of Chemistry also provides a good source of pK_a data.[86] Software tools, such as ACD and SPARC, provide access to experimental data with references. SPARC will only provide the reference data based on queries for an exact structural match.[87] ACD iLabs has the added benefit of providing literature references for molecules based on a structural similarity search, returning references for exact structural matches and a limited number of similar structures based on a user defined similarity score threshold.[88] Another potential source which can identify articles that may contain pK_a data based on a structure or text based query is SciFinder Scholar.[89] Unfortunately, such data obtained from sources other than the original literature references are not necessarily clean, complete, or standardized. Sometimes even published data are unreliable, and often conflicts are found between sources. For example, when mining Beilstein for the experimental pK_a measurements of phenol, approximately 30 records exist in a bimodal distribution. The distribution consists of a cluster of six values around the pK_a of −1.0 and over 20 values clustered around the pK_a of 10. Evidently, −1.0 values arose from entering the negative logarithm of the pK_a. Other common but less frequently identified errors include the following: typographical errors, predicted values (rather than experimental), incorrect transcription of temperature and/or solvent used, and the incorrect associations between the experimental pK_a and the ionizable sites on polyprotic molecules. In a previous effort to identify and resolve some of these problems, a method for curating pK_a data from multiple sources having redundancies was discussed.[90] Finally, so much pK_a data is associated with proprietary chemical structures that public training sets are not as chemically diverse as they could be. Hence predictive models may be less accurate for some molecules in proprietary data sets. Table 4 provides an updated list of free and commercial electronic sources of pK_a data from the review article published in 2006 on '*in silico*' prediction of ionization constants by Fraczkiewicz.[91] Another significant source of pK_a data is the six-volume "Critical Stability Constants" by Martell and Smith.[92] These data are available for purchase as the NIST Standard Reference Database 46 from the National Institute of Standards and Technology.[93] Buyer beware, as the database is a self-contained application which has very limited search capability and does not support data

Table 4. pK_a Data Sources

data source	vendor	url
ACD/pK_a DB	Advanced Chemistry Development	www.acdlabs.com
ADME INDEX	Lighthouse Data Solutions	www.bio-rad.com
Beilstein Database	Elsevier B.V.	www.info.crossfiredatabases.com/
BioLoom Database	BioByte Corporation	www.biobyte.com
The Merck Index, 14th edition	Cambridgesoft Corp.	www.cambridgesoft.com
Lange's Handbook of Chemistry, 15th edition	Knovel	www.knovel.com
CRC Handbook of Chemistry and Physics, 89th edition	CRC	www.hbcpnetbase.com/
HSDB	National Institutes of Health	toxnet.nlm.nih.gov/
LOGKOW	Sangster Research Laboratories	logkow.cisti.nrc.ca/logkow/
MolSuite DB	ChemSW	www.chemsw.com
Pallas	CompuDrug	www.compudrug.com
p*K* database	University of Tartu, Estonia	mega.chem.ut.ee/tktool/teadus/pkdb/
PHYSPROP	Syracuse Research, Inc.	www.syrres.com
SPARC	University of Georgia	ibmlc2.chem.uga.edu/sparc/

export to common machine readable formats such as comma space delimited (CSV) or standard data files (SDF).[94]

In the absence of available data, ideally one should simply measure the pK_a by titration. However, this is not an option for large libraries of *in silico* small molecules that have yet to be synthesized. When dealing with the vastness of chemical space, often a good computational approximation is superior to experiment in order to overcome cost and time limitations.

3.1.2. Experimental Methods. Analytical chemistry has provided a plethora of experimental tools for making pK_a measurements, some of which are amenable to automation. For those interested in the history of titration and its development for the use of colorimetric and electrometric analysis used in the determination of pK_a, The *History of Analytic Chemistry* describes the achievements during the early 20th century.[95] Today, there are two main titration methods: volumetric and coulometric. The volumetric method entails adding titrant directly to the sample, whereas the coulometric method generates titrant electrochemically. Some of the associated indicator methods include the following: colorometric, potentiometric, conductometric, spectrophotometric, amperometric, thermometric, solubility, cryometric, and NMR (discussed above for protein pK_as). In order to show the most commonly used and preferred methods, queries of data obtained from the Beilstein database relating to the analytical methods were performed as shown in Table 5. The potentiometric, spectrophotometric, and conductometric methods have been used predominantly to determine pK_a. Interestingly enough, approximately half of the pK_a measurements obtained from the Beilstein database are not associated with a method. Furthermore, over the past year the total number of measurements increased by approximately 20%. Only 10% of the new pK_a data are associated with a method. Regardless, over 98% of the measurements with stated indicator methods used potentiometry, spectrophotometry, and conductometry. Seemingly, there is a wealth of available information; however, after curating the data for monoprotic molecules and accounting for redundancies, the authors have found reliable pK_a data for fewer than 2000 molecules.[90]

Potentiometric titrations are possible in turbid, deeply colored, highly absorptive, or dilute solutions. Simplicity, speed of measurement, robustness, and ease of automation have made it the historically preferred method to measure pK_a. On the other hand, potentiometry does have pitfalls when it comes to pK_a measurement. Measurements on compounds sparingly soluble in water require mixed solvent extrapolation.[96] Nonhigh throughput applications require large amounts of reagent (not favorable for newly synthesized compounds or natural products) and time to prepare solvent from carbonate free solutions.[97] Furthermore, impurities of both reagent and analyte can affect the observed pK_a values.[96,98] When dealing with nonaqueous media, repeated measurements are often recommended, as the signals from the glass electrodes are less reliable than those from aqueous media, due to high liquid junction potentials.

Table 5. Beilstein Database (DE.MET): Dissociation Exponent Method Category[b]

	count based on Beilstein[a] version	
indicator method	2007.04	2008.03
(blank)	56832	79602
potentiometric	45639	46906
spectrophotometric	18339	18872
conductometric	2127	2163
NMR	687	774
kinetics	359	367
calorimetric	76	143
solubility	31	32
polarographic	11	15
distribution	5	6
total	124106	148880

[a] Beilstein was accessed using the Molecular Design Limited (MDL) CrossFire Commander. [b] Numbers reflect pK_a measurements in all conditions including same molecule measurements in various solvents and temperatures.

Similar to the potentiometric method, conductometric pK_a measurements are possible in turbid, deeply colored, highly absorptive, and dilute solutions. They are relatively simple, quick to perform, and have been automated. On the other hand, it does not suffer from the same degree of reagent solubility limitations experienced in potentiometry, although conductometric methods are rather problematic in the presence of foreign electrolytes, which decrease the accuracy of the measurement. Purity is not the only issue, as conductometric methods are also sensitive to temperature. Raising the temperature one degree results in a 2−2.5% increase in the conductivity of most salts. Furthermore, conductometry is generally considered inferior in accuracy compared to potentiometric and spectrophotometric methods.[99]

Ultraviolet spectroscopy hinges on the principle that the uncharged and ionic species of a compound exhibit different

spectra. Spectroscopy is known for its excellent precision in pK_a measurements. While it is comparable in accuracy to potentiometry, spectroscopy can be used to measure pK_a for compounds having poor aqueous solubility. Amenable to high throughput automation, UV spectroscopy has the added advantage of requiring ten times lower concentrations of reagent than similar high throughput potentiometric titrations do.[97,100]

In the early 1990s, capillary electrophoresis (CE) began to show usefulness as a universal analytical technique for determining pK_a over a wide pH range.[101,102] It relies on the principle that the solute exhibits an electrophoretic mobility continuum versus pH (uncharged has no mobility; charged has maximum mobility). It exhibits both higher sensitivity and selectivity than potentiometry and spectrophotometry do, producing accurate pK_a values for small molecules. The method is capable of handling compounds of diverse solubility and samples of low concentration since it relies on migration times and does not require measurement of the reagent or titrant concentrations, as potentiometry does. The solute is purified during migration, as impurities have inherently different migration times, so purity is not an issue with CE. Finally, CE is not limited by media, as measurements can be made in aqueous, aqueous−organic, or nonaqueous media.[96,103]

pK_a can also be determined through reverse phase high performance liquid chromatography (HPLC) by measuring the capacity factor based on a compound's retention time in a column against a series of solvents having mobile phases at different pH values.[104,105] The advantages of using HPLC for pK_a screening include the minimization of solubility restrictions, eliminating the effect of impurities on measurements (thus allowing for screening combinatorial complexes directly), and the potential for high throughput automation when coupled with a mass spectrometer. However, the main potential disadvantage is the loss of accuracy when using organic solvents in a nonpolar column, due to the potential interaction between analytes and the stationary phase.

Several high throughput pK_a assays have been recently reviewed, such as capillary electrophoresis in conjunction with ultraviolet detection, capillary electrophoresis coupled with mass spectroscopy, pH gradient HPLC with mass spectroscopy, and a mixed buffer linear pH gradient system with measurements based on ultraviolet absorbance.[106] While it is not the goal of this paper to focus on the details of experimental methods, it is useful to note the development of newer technologies increasing the efficiencies of physicochemical property measurements for large chemical libraries of compounds lacking such data.

Familiarity with potential experimental problems can help in data curation. When performing a titration to determine pK_a, the following conditions are preferable to avoid unnecessary errors.[100] First, the analyte needs to be water-soluble. For those molecules with poor water solubility, Yasuda-Shedlovsky plots can be used to extrapolate the theoretical aqueous pK_a from a gradient of semiaqueous (water:methanol) pK_a measurements.[107] In a similar approach, a series of measurements at various ionic strengths allows for extrapolation to zero ionic strength. An alternative approach, sacrificing accuracy for number of measurements, requires only one measurement and uses a linear equation with slope and intercept based on limited families of compounds, such as phenols and protonated amines.[108,109] Second, in order to acquire an accurate measurement, compounds need to be stable enough to establish an equilibrium between two states of ionization. Third, compounds need to be pure in simple titration experiments, although with high throughput techniques purity is much less of an issue. Fourth, in order to eliminate experimental errors and for the purpose of validation, it is usually preferred to use a series of homologous compounds. Fifth, each titration should be performed in a thermostatic environment, as dissociation is an endothermic process. It has been found that pK_a values for some common organic acids change by less than 1 pK_a unit between 5 and 60 °C, typically decreasing with the increase in temperature. Anomalies may result when approaching 0 °C when the solvent is water.[110] Sixth, the presence of carbonate has been shown to affect data measurements leading to anomalies in the titration curve, which may require correction.[111] Carbonate acts as a base in aqueous solution and has pK_b of 6.36. Carbonic acid is diprotic and has pK_as of 3.60 and 10.25. The amount of carbonate in solution is proportional to the partial pressure of carbon dioxide in the atmosphere. Therefore, when performing titrations using a low concentration of acid, anomalies in measurements and the titration curve can occur when the pH is around 3.6, 6.36, and 10.25. Seventh, in the cases of most manual titrations, high amounts of test materials are needed. Finally, care must be taken when performing acid−base titrations with perfluorinated compounds, which exhibit an artifact of sorption. The unionized form of a perfluorinated compound tends to preferentially adhere to interfaces, both water surface and glassware. This nonuniform distribution affects the overall measurable concentrations of the unionized species in solution. Unfortunately, the extent to which sorption occurs with "ordinary" compounds is unknown, but it can be conjectured that a similar but lesser effect may be experienced for molecules having highly halogenated lipophilic tails. In cases where the concentration of the nonionic species is reduced due to sorption, one can expect the titration curve to be right-shifted, indicating higher than actual pK_a values for acids.[112,113] Furthermore, there are cases where the addition of cosolvents did not decrease the sorption of an organic acid in the aqueous phase.[114] Considering these points, even seemingly well curated data may be corrupted due to potential artifacts beyond simple human transcription errors, leading to unexpected biases in otherwise well constructed and validated predictive models.

3.2. Predictive Methods. *3.2.1. Linear Free Energy Relationships.* Seminal publications[115,116] and reviews[110,117] on the prediction of pK_a base their model on linear free energy relationships (LFER), applying the Hammett equation where pK_{a0} is the ionization constant for the parent molecule, pK_{aS} is that of the substituted molecule, ρ is the constant for a particular class of molecules, and σ_i is the effect of the i^{th} substituent on the ionization constant of the parent molecule.

$$pK_{aS} = pK_{a0} - \rho \sum \sigma_i \quad (2)$$

The flaw in this method is that the parent molecules inherently carry the majority of the chemical information, and without training on a particular parent, predictions for such compounds are impossible. A good example of this

problem is the MCASE study. Until the mid 1990s, there had been few attempts to model diverse sets of chemicals. Like eq 2, the MCASE approach used a linear regression equation with terms for quantitative properties such as log*P*, water solubility, molecular weight, absolute electronegativity, hardness, Hückel molecular orbital charge densities, and HOMO and LUMO coefficients, plus possibly 58 indicator variables for the presence of certain molecular substructures. The set of compounds was broken into 22 subsets based on the presence of molecular fragments called biophores that were particularly associated with acidity, the most important one being the carboxyl group. Different regression equations were determined for each subset. The total training and test set consisted of 2464 organic acids. When using 1848 molecules in the training set, the r^2 based on the predictions of the remaining 616 compounds was 0.91 with standard deviation 0.774. When training the model with the entire data set and attempting to predict the pK_a for 214 drug molecules, which were most likely not well represented by the parent molecules in the training set, the r^2 was 0.70 with a standard of deviation of 1.44.[118]

LFER models are still used and have been implemented in popular commercial and freely available software packages, such as EPIK and SPARC.[119−122] To help overcome the aforementioned problem, the SPARC methodology combines LFER, perturbed molecular orbital theory, and QSPR to deal with the effects of π-bonding and electron delocalization. SPARC scored an r^2 of 0.80 with RMSE of 1.05 on a set of 537 molecules from an internal Pfizer data set[122] and is able to predict both macro- and micro-pK_as calculated from molecular structure alone. It is worth noting that while SPARC is not parametrized for all atom types, one can modify the molecule being tested by substituting the closest weight atom of the same group and achieve reasonable, often good, results. For example, we have found that most predictions where silicon atoms were replaced by carbons and seleniums by sulfur were within 1 pK_a unit from the experimental value. However, since there are few heavy atom compounds to test, the accuracy of such replacements remains questionable.

3.2.2. Quantitative Structure−Property Relations. 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).[118,123−129] Other methods include artificial neural networks,[84,91,130,131] quantum mechanical continuum solvation models,[132−134] anticonnectivity models,[135,136] and tree based methods.[90,127,137,138] 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.[124,129,130,132,134−136] Others have tackled the problem of chemical diversity by devising and combining multiple models, each applied to a relatively small set of molecules when compared to the complete set of experimental data.[127,130] Here the overall combination of models is 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. The following sections address some of the most significant methods that have been used in property prediction. Table 6 shows the performance of several commercial pK_a prediction models and other methods used in the literature within the past two decades.

One of the reasons QSPRs are so popular is that linear regression always leads to a unique, easily computed model. Typically, highly correlated descriptors are undesirable, but with partial least-squares, highly correlated descriptors are handled appropriately. Like the MCASE approach, when dealing with complex properties such as pK_a, it is common to break up the training set into groups based on ionizable site type and chemical group, because it is impossible to establish a single robust model on a chemically diverse set of molecules. In doing so, multiple models are generated which may or may not use the same set of descriptors. Descriptors can be qualitative or binary, representing a <has> vs < has not> property, or they can be quantitative experimental or calculated features. Reducing the size of the training sets by separating the molecules into classes often leads to a better fit but does not necessarily reduce the number of descriptors considered. Caution needs to be exercised in order to avoid overfitting, especially when nonlinear descriptors are considered. If we have a descriptor named x, the regression equation can also have terms for $\ln x$, $\log_2 x$, $\log_{10} x$, x^n (where n is any real number), and so on. Overfitting the model leads to a high correlation between the calculated and experimental values in the training set but a poor or nonexistent correlation in cross-validation or validation on a separate test set. Forward or reverse stepwise linear regression are ways to select the smallest subset of descriptors which still fit the property being modeled. In 1972 Topliss and Costello pointed out the risk of chance correlations in quantitative structure activity relationships and gave recommendations for the number of descriptors to be used in linear regression models given the number of observations to be fit.[139] Hence, the major step in deriving a robust QSPR model is finding the smallest set of molecular descriptors that best represents the structural variations in a set of chemically diverse molecules. In QSPR the most common methods to identify a good set of descriptors are stepwise multiple linear regression, partial least-squares (PLS), and principal components analysis.

Comparative molecular field analysis (CoMFA) has been used to model pK_a values for a small series of chemical homologues where partial least-squares found a linear correlation using four parameters.[140−143] While the statistics appear very good, one must note two limiting factors: first the models were trained on very small, chemically similar training sets, and second the results depend on the chosen conformations and spatial alignments. More recently, a CoMSA (comparative molecular surface area analysis) study using similar 3D descriptors and PLS fitted the pK_a values for a series of benzoic acids.[128] The previous CoMFA methods appear to outperform the CoMSA method in predicting pK_a.

3.2.3. Quantum Mechanical and Continuum Electrostatic Methods. Continuum electrostatics models and quantum mechanical descriptors have also been a focal point in small molecule pK_a prediction. Similar to the work by Antosiewicz on protein pK_a,[10] a continuum electrostatics model for small molecule diamines using UHBD to make FDPB calculations resulted in an r^2=0.86 and RMSE=1.1 for the 12 ionizable sites in six aliphatic diamines with experimental pK_as ranging

Table 6. Survey of pK_a Prediction Methods[f]

			training set			test set			external test set		
method	ref	class	N	r^2	RMSE	n	q^2	RMSE	n	r^2	RMSE
QSPR/PLS	137	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 - MoKa	127, 141	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 (CoMSA)	128		49			49	0.86		23	0.77	
QSPR/PLS (CoMFA)	140	benzoic acids	49	0.916	0.102[b]						
	142	imidazoles	23	0.99	0.19[b]			0.27[b]	5	0.98	0.15[a,b]
		imidazolines	16	0.99	0.35[b]			0.69[b]			
	143	nucleic acid components	18	0.99	0.19[b]		0.89				
QSPR/MLR	125		15	0.97	0.12				3	0.99[a]	0.10[a]
QSPR/MLR	126	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	129	aromatic acids	74						33	0.99	0.27
QSPR/LFER	124	monoprotic oxy acids	135	0.993	0.455				14		0.471
QSPR/LFER - MCASE	118		2464			616	0.91	0.774[b]	214	0.70	1.52[b]
QSPR/LFER - EPIK	119, 120		4057		1.27[b]				123		1.37[b]
QSPR Anti-Connectivity	136		31			31	0.87	0.463			
ANN - ChemSilico	130	12 classes	>16000						665	0.83	
		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				
ANN/PCA/GA	131	nitrogen	170(282)	0.99	0.30						
ADMET Predictor	84, 91		9075	0.971	0.593				2253	0.961	0.644
Semiempirical/PLS (Novartis In-House)	123	all			0.48				350		0.81
		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				
Semiempirical MO	144	phenols				175	0.93	0.599[b]			
	145	benzoic acids				99	0.85	0.357[b]			
		amines and anilines				132	0.94	0.985[b]			
		N containing heterocycles				150	0.69	1.168[b]			
Semiempirical RM1+solv.	146	carbon containing aliphatic amines	26	0.948	0.68[b]						
Quantum (MEP)	147	phenols and carboxylic acids	228	0.896							
Quantum (MEP-V_{min})	148	anilines	36	0.945	0.301[b]						
(MEP-$V_{S,min}$)				0.932	0.336[b]						
($I_{S,min}$)				0.949	0.285[b]						
(Hammet σ)				0.940	0.310[b]						
Quantum (MEP-$V_{S,min}$)	149	phenols	19	0.938	0.300[b]						
(MEP-$V_{S,max}$)				0.932	0.314[b]						
($I_{S,min}$)				0.941	0.292[b]						
($I_{S,min}$ and $V_{S,max}$)				0.953	0.271[b]						
Quantum (MEP-$V_{S,min}$)	149	benzoic acids	17	0.942	0.120[b]						
(MEP-$V_{S,max}$)				0.970	0.085[b]						
($I_{S,min}$)				0.941	0.120[b]						
Quantum (philicity)	150		63	0.98	0.57[b]						
Quantum solvation	132	carboxylic acids	16	0.69	0.72						
Quantum solvation	151	phenols	20		0.38						
Quantum solvation	152		11	0.88	2.2						
COSMO-RS	133	bases	43	0.98	0.56[b]				58		0.66
	134	acids	64	0.98	0.49[b]						
Jaguar	153								191	0.98	0.66
MD continuum solvation	154	diprotic acids	12	0.96	2.02						
QSPR/LFER/PMO-SPARC	87, 121, 122		2500	0.99	0.36[b]				4338	0.99	0.37[b]
		Pfizer data set[c]							123	0.92	0.78[b]
		Pfizer internal data set[d]							537	0.80	1.05[b]
									185[c]	0.84	1.15

Table 6. Continued

			training set			test set			external test set		
method	ref	class	N	r^2	RMSE	n	q^2	RMSE	n	r^2	RMSE
MARVIN	82, 155					208[c]	0.98	0.38[b]	185[c]	0.88	1.03
ACD/I-Lab v8.03	88		>31000						185[c]	0.90	0.93
ADME Boxes	156								185[c]	0.93	0.69
SMARTS pK_a	90		1693	0.95	0.65	10%	0.91	0.80	185	0.94	0.68
Consensus[e]	90								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. See ref 90. [d] These molecules were unlikely to be found in the SPARC training set. [e] The consensus model used predictions from SPARC, MARVIN, ACD/I-Lab 8.03, and SMARTS pK_a. [f] 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.

from 1.09 to 10.34. Significant errors occurred in the calculations for the primary pK_a of 1,2-diaminopropane and the secondary pK_a of succinic acid, which were both calculated approximately 3 units below the experimental values.[154]

A polarizable continuum model was used to evaluate 15 small simple monoprotic molecules with experimental pK_a ranging from −6 to 33 with $r^2 = 0.96$ and RMSE = 2.02. All in all, there were 11 compounds that deprotonated within the range 0 to 16 with $r^2 = 0.88$, RMSE = 2.2, and a maximum error of 4.7.[152] Electrostatic models have also been used to predict pK_a for small multiprotic tetrahedral and triangular oxyacids, such as arsenic (H_3AsO_4) and arsenious (H_3AsO_3) acid, with close to the same accuracy as for the simpler organic acids.[157]

QM descriptors offer a promising means to accurately calculate pK_a. The *ab initio* aspect allows for greater confidence when calculating pK_a for molecules than when using strictly empirically derived descriptors. That is, QM methods are not restricted to the chemical diversity of a training set of molecules. However, the calculations are time-consuming and not feasible when considering large databases of theoretical molecules or the analysis of macromolecules. Some QM descriptors that have a strong correlation to pK_a include superdelocalizability,[123] polarizability,[123] group philicity,[150,158] molecular electrostatic potential (MEP),[147−149,161] and molecular surface local ionization energy ($I_{S,min}$).[148,149,159−161]

Group philicity refers to the electrophilic nature of the ionizable group, such as a carboxyl group, and is equivalent to the sum of the local electrophilicities of each group atom, which are determined by the electrophilicity of their respective bonded neighboring atoms and calculated using density functional theory. The philicity descriptor is a modification of a molecule's electrophilic index.[162] The reciprocal of the group philicity showed a strong correlation to the pK_a for 63 molecules including carboxylic acids, substituted phenols, anilines, phosphoric acids, and alcohols.[150]

Three classes of MEP calculations have shown a strong correlation to pK_a: spatial minima (V_{min}), surface minima ($V_{S,min}$), and surface maxima ($V_{S,max}$).[148,149] It was recently shown that the MEP minus a given reference value for each category of compounds (as in the FDPB calculations for protein pK_a prediction with the electrostatic methods) has a single unique linear relationship to the experimental pK_a data for thiols, sulfonic acids, alcohols, carbonyl acids, amines, and analines.[147]

The investigations of Brinck et al. established the correlation between a molecule's pK_a and an ionizable atom's minimum surface local ionization energy ($I_{S,min}$), as defined by self-consistent-field molecular orbital theory.[159−161] The location of the $I_{S,min}$ is related to charge/transfer polarization and indicates the areas where the least amount of energy is needed to abstract an electron from the surface of the molecule. Early investigations found a single linear relationship between the $I_{S,min}$ and pK_a for four sets of carbon and oxygen acids as well as three nitrogen acids. It is easy to see from the scatter plots correlating $I_{S,min}$ to pK_a that some calculations missed by approximately 10 pH units and that the high correlation coefficient ($r = 0.97$) was due largely to the broad range (−5 to 40) of experimental pK_as considered.[160] Later investigations confirmed that not all ionizable groups could be represented by a single linear equation due to key structural differences between the ionizable groups.[149]

While MEP descriptors and $I_{S,min}$ both show good pK_a correlations for different series of compounds taken separately, it is interesting to note that the derivation of $I_{S,min}$ and $V_{S,min}$ correspond to different atoms and generally do not correlate.[149,161] When performing simple linear regression on one descriptor, the $I_{S,min}$ has been shown to be slightly superior to the V_{min}, $V_{S,min}$, and $V_{S,max}$ as well as the natural charge and relative proton-transfer enthalpy.[148,149] Furthermore, it was found that no significant improvement could be obtained by linear regression on combinations of $I_{S,min}$, $V_{S,min}$, and $V_{S,max}$ QM descriptors.[149] Although these QM descriptors appear quite promising, no strong correlations have been found between them and pK_a for aliphatic amines. More recently, good correlations between neutral amines (excluding ammonia and hydroxylamine) and their cations were found by using the SM5.4A solvent model and performing calculations at both the semiempirical RM1 and density functional theory (DFT) B3LYP/6-31G* levels.[146]

Ab initio quantum mechanical methods are always the slowest but have often been found to be the most accurate, such as Jaguar, which performs geometry optimization at the DFT B3LYP/6-31G* level.[153] Shields et al. used QM calculations to accurately predict pK_a for 20 phenols[163] and 6 carboxylic acids[164] using a CPCM[165] continuum solvation method in Gaussian 98.[166] CPCM utilizes COSMO,[167] a conductor-like screening model to calculate the polarization charges of a molecule, in a polarizable continuum model (PCM) framework.

Another popular QM package that has been shown to perform as well as Jaguar in aqueous environments, capable of accurate pK_a calculations, is COSMO-RS,[168,169] where the RS stands for real solvents. It is a statistical thermodynamics postprocessing of COSMO calculations that extends the applicability of quantum chemistry to the entire range of fluid thermodynamics including mixtures and variable temperatures. An assessment of several *ab initio* programs for the prediction of pK_a, not including Jaguar, tested neutral, cationic, and carbon acids with experimental pK_as ranging from 14 to 36. Ignoring three outliers, the overall r^2 was 0.89, while considering all data points lowers the r^2 to 0.72.[170] Still there are foreseeable complications. The COSMO-RS model was able to fit a set of 43 bases very well, but when aliphatic amines were considered, correction factors needed to be introduced for secondary and tertiary amines which uniformly deviated from the regression line. Furthermore, two compounds (hexamethylenetetramine and 1,2-diazabicyclo[2,2,2]octane) did not share this deviation. In these cases, ionizable nitrogens act as bridging atoms for a bicyclic ring system.[133] Strictly empirically based methods could not even hope to achieve this level of accuracy based on the relatively small sampling of chemical space considered for both acids and bases.

For pK_a predictions outside the physiological range, the *ab initio* QM methods tend to be more robust and often more accurate than the empirical and less complicated continuum electrostatic methods. This class of pK_a predictors also allows for a broader range of analysis than is afforded by empirical models trained on pK_a data obtained from titrations in H_2O alone. On the other hand, validation has been performed on rather small data sets, and it is not clear that even the QM methods will be able to maintain their statistics, based on a study of carbonaceous ionizable sites, involving COSMO-RS and other contemporary theoretical methods.[170]

3.2.4. Artificial Neural Networks. The theory behind artificial neural networks (ANN) has been described in the literature.[171−173] ANNs are a powerful tool for making nonlinear approximations and are designed to emulate how the brain processes information through a network of neurons. As such, the neurons of an ANN act as interconnected units, processing information based on mathematical functions. Like the Internet, each neuron acts like a minicomputer receiving requests and sending responses to other neurons. Multilayered ANNs with enough neurons have been said to have the capability to approximate almost any nonlinear mapping of input to output to any required accuracy.[174] They are well suited to handle large data sets and identify complex nonlinear patterns that could easily be missed by a simple equation or set of equations modeling a system, such as those derived through linear regression.

A principal component, genetic algorithm, artificial neural network was used to calculate the pK_a of 282 various nitrogen containing molecules in water.[131] The training set consisted of 170 molecules, the cross-validation set 56 molecules, and the prediction set 56 molecules. The model uses 406 descriptors to identify 179 principal components that explained 99.9% of the total pK_a variance. Of these, 15 principal components were included in the final model. While it appears that the model generation made use of the training and validation sets, we were not able to determine if only the molecules in the training and validation sets were used to perform the principal components analysis. The RMSE of the 56 molecule prediction set for the artificial neural network was 0.0750, compared to 1.4863 when multiple linear regression was used to build the model. The extreme accuracy of the artificial neural network model, the excessive number of descriptors used to perform the principal components analysis, and the small number of molecules suggests that PCA was performed on the entire data set to identify the relevant principal components, and that overfitting is an issue. As with many other empirical models, this one is unable to identify the site of ionization.

A well validated ANN model has been implemented by Simulations Plus Inc.[84,91] This model has diverse training and test sets consisting of 9075 and 2253 pK_a data points, respectively. It is also the only neural network model and one of the few prediction utilities that predicts micro-pK_as. Based on the large training and test sets and the respectable score from the Chem Silico data set, this utility appears to be not only the most accurate,but also one of the most robust methods for pK_a prediction. On the other hand, this model, like most others, is only parametrized for C, N, O, S, P, F, Cl, Br, and I atoms. However, it will process molecules with other atom types as well as salts (which are washed away), providing a warning to the user. Substances that are mixtures of compounds are not processed.

3.2.5. Database Methods. Tree methods and database lookup methods are becoming more popular tools for pK_a prediction. The simplest form of database lookup method is to assign to the test molecule the pK_a of the database molecule that is most similar, according to some similarity metric, such as the Tanimoto similarity coefficient, based on some fingerprint. The fingerprint is typically made up of qualitative descriptors. The accuracy of the model depends on how comprehensive (large and chemically diverse) the molecular database of experimental pK_a data is and the design of the fingerprint. To represent the database well, ideally every molecule in it would have a unique fingerprint. In one study it was found that pK_a assignment based on a simple atom type method using SMARTS strings was able to assign pK_a to simple molecules with an r^2 of 0.80 and a standard deviation of 0.95.[175]

Other methods are far more rigorous, such as that of Kogej and Muresan.[138] Here a pK_a database was mined using fingerprints based on 64 atom types represented by SMARTS strings and bond distance from ionizable site. A fingerprint was taken at each level of removal from the ionizable atom (level 1). Atoms one covalent bond removed from the ionizable site were at level 2, atoms at two bonds removed were at level 3, etc. After the entire training database is fingerprinted, a fingerprint exists for every concentric level of removal from the ionizable site for each ionizable site in the molecule. This style of submolecular fingerprints has also been coined 'circular fingerprints', as it dissects local structural information in expanding concentric levels of bond removal from the ionizable site.[176] This method allows for identification and predictions for each ionizable site in a molecule based on the theory that the ionization of a particular group is dependent on these topological subenvironments. Furthermore, it allows for predictions of the microspecies, but it is likely that microspecies would be poorly represented, as the vast majority of published pK_as are for the macrospecies. pK_a assignment is made by

identifying the highest level where exact matches to the fingerprints are found. In the event that there are multiple matches, the average pK_a value for the highest level fingerprint matches is taken. Typically accuracy was good for level 4 or greater matches, but level 6 or more is needed for substituted aromatic ring systems such as substituted phenol. Note that 48% of the compounds cannot be predicted with level greater than 4, which is a problem for the substituted aromatic compounds and amines, indicating the need to expand the database of 4700 compounds. One advantage of this approach is that when attempting to predict the pK_a for a compound found in the database, the exact value is returned as in a lookup. While the authors mention a 20-fold cross-validation (training the model using 95% of the data and testing on the remaining 5%) no statistics (r^2 or RMSE) for overall performance were provided. At level greater or equal to 5, 16% (4%) of the tests had a mean absolute deviation greater than 0.5 (1.0), indicating high accuracy for the vast majority of compounds and a need to increase the chemical space covered by the database.

Xing also used molecular tree structured fingerprints but included the number of hits at each level of removal.[137] Like the previous approach, this method allows the individual treatment of specific chemical classes, as it generates tree-structured molecular descriptors for each class. A problem was experienced with an external data set where four ionizable sites could not be classified because of the specificity trained into the chemical classes. Applying more general rules in their previous work[177] allows the missed ionizable sites to be combined with a similar class of molecules. While this appears to reduce the overall fit of the model on the training data, it shows that generalizations can lead to an improvement in the overall robustness, while refinements lead to improvements in accuracy. Here we would suggest enlarging the training set while retaining the more general classifications in a parallel scheme for background operations. This way one could gain all the improvements of the more specific descriptors without suffering loss of information. A training set consisting of 625 acids and 214 bases had a standard error of 0.41 for acids and 0.30 for bases. Similar tree based methods were also investigated in industry. In 2007, Jelfs et al. described a method extending the molecular tree structured fingerprints by including 2D substructural fingerprints, which were used to flag the presence of other important structural features that affect pK_a.[123] As before, they found that the molecular trees (circular fingerprints) needed to consider at least five bonds of removal from the central atom for adequate results. This makes sense, for example in the case of the acidic OH group of phenol, where a para-substituent would be five bonds removed or at level 6 in the database lookup method of Kogej. The software package MoKa implements this concept, where the descriptors are based on molecular interaction fields precomputed on a set of molecular fragments.[141]

3.2.6. Decision Trees. Decision tree methods have also been considered recently in pharmaceutical research, both for the prediction of biological activity and for physicochemical property predictions.[90,178,179] The benefits of decision tree methods include the following: (1) explaining nonlinear response, (2) ease of interpretation, as they provide a clear decision path for better understanding of the test compound, (3) the ability to ignore irrelevant descriptors, (4) the ability to handle large sets of both quantitative and qualitative descriptors, (5) the ability to handle large sets of structurally diverse compounds, and (6) speed. The main drawback with decision tree methods is how to deal with multiple data points for the same molecule, as in the case of polyprotic acids. Other drawbacks include instability and lack of accuracy when compared to other algorithms. While decision trees have commonly been used for classifications, it is possible to derive a regression tree to provide a quantitative rather than qualitative result.[180] Predictive decision trees can be derived through recursive partitioning, which often leads to an unbalanced tree, when smaller groups of compounds having similar property values are filtered out early on, rather than making clever choices that favor a more balanced tree. By defining a pool of both backbone and substituent molecular fragments in terms of highly generalized and specific SMARTS strings, Lee and Crippen were able to iteratively construct a more balanced decision tree where each decision gave weight to the evenness of each split and the reduction in pK_a variance at each child node for a large set of monoprotic molecules.[90] Performance of SMARTS pK_a was competitive with and even exceeded that of several well-known applications as described in Table 6. It appears that accuracy and stability can be maintained by not heavily relying on highly specific descriptors or by making decisions leading to terminal nodes (where predictions are assigned) close to the root node. As with any other empirical model, regression trees can only be as good as the data used in training, particularly the accuracy and spread of the experimental pK_as and the chemically diversity of the training set compounds.[90]

4. PROTEIN−LIGAND COMPLEXES

In rational drug design, one of the ultimate goals is to understand protein−ligand interaction; therefore, it is significant to note a recent change in direction for pK_a prediction, which attempts to account for binding effects. Experimental studies have demonstrated that ligand binding induces protonation state changes.[181−184] Dullweber et al. examined a series of congeneric ligands and identified significant changes in protonation states when binding to thrombin and trypsin.[184] Recently, Czodrowski et al. have developed a method for predicting protein pK_a that also accounts for pK_a shift due to a bound ligand.[44,185,186] Here, the pK_a shifts for the ionizable sites in proteins were calculated with MEAD,[187] parametrized with partial charges from a modified version of Gasteiger's PEOE[188] method. As part of the method validation, pK_as were calculated for 132 ionizable groups of 9 proteins (RMSE = 0.88) and showed significant improvements over the null model and FDPB calculations using partial charges from PARSE[189] and CHARMM22.[190]

PROPKA 2.0 also attempts to address the issue of pK_a shifts in relation to protein−ligand binding for the ionizable groups of both protein and ligand.[80] The underlying empirical rules of PROPKA 1.0[22] have been modified in PROPKA 2.0 to include the effects of the functional groups of the ligand. The model (or null) pK_a values for the ligand are taken from literature or from MARVIN, when no experimental data are available. In all 26 protein−ligand complexes were studied. Of these, PROPKA 2.0 was shown to identify

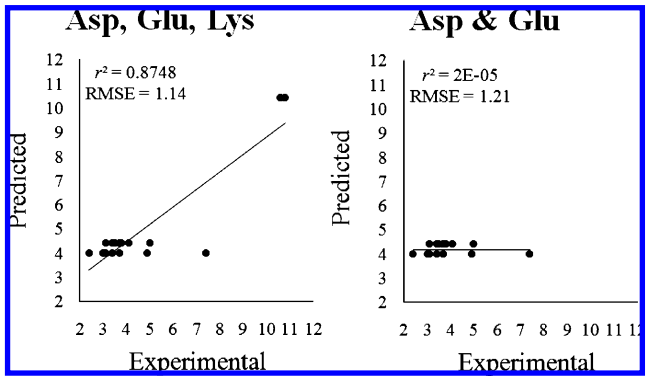

Figure 1. Simulation of the null model used to predict the pK_a of 18 residues (9 Asp, 7 Glu, and 2 Lys). One predicted value is assigned to each residue type, although the respective experimental values vary due to environmental effects. Adding in the two Lys data points raises the correlation coefficient (r^2) from nearly zero to 0.87 while improving the RMSE only slightly due to the close fit for the two extra points.

changes in protonation states that agree with the majority of the experimental data; however, no statistics on the pK_a predictions were provided. Clearly, more NMR pK_a data for protein−ligand complexes are required for retraining and validating both models, but it is encouraging to see that the combination of methodologies can provide some promising results.

5. STATISTICS AND BENCHMARKING

Often experimental data for a particular ionizable site are provided as a range. The range is due to estimated experimental error or data curation, where multiple measurements were obtained from different sources possibly using different techniques. One can only hope that care was taken in the curation, as outliers can significantly extend the range. Using an experimental range leads to a lower RMSE than when single values are used. Typically any predicted value falling within the range has an error of 0, and otherwise the error is the absolute difference between the value and the nearer limit of the range. Ranges are more commonly used when computing statistics for protein pK_a predictors, whereas the small molecule pK_a predictors generally compare to single experimental values in terms of Pearson's correlation coefficient squared, r^2. On the other hand, r^2 is often not considered when comparing experimental to calculated data for proteins, as pK_a data for proteins are dominated by carboxyl groups in aspartic and glutamic acid residues, and by the imidazole group of histidine residues, most of which tend to have pK_as in the range 2 to 6. If only a few lysine or tyrosine residues (pK_a 9 to 11) are added to the comparison, the r^2 increases substantially, while the RMSE will remain close to the same, providing that there is no significant difference in the range of errors for the different residue types. It was for this very reason that Pokala omitted the few available Lys data points from the evaluation, as it exaggerated the null model's apparent accuracy. Including the Lys data points increased the r^2 for the null model from 0.36 to 0.90.[54] When considering r^2, it is better to apply the statistic to the correlation between experimental and predicted pK_a shifts for each residue type separately. Figure 1 illustrates how using the correlation coefficient to evaluate data can be misleading. Similarly, it can be worthwhile to consider r^2 for different classes of small molecules to best evaluate the strengths and weaknesses of different pK_a predictors. When comparing methods, it is important that a discriminative benchmark is used.[63]

Table 7. Predictive Abilities of Ten pK_a Prediction Utilities[191]

software	url	no. of molecules predicted	r^2	MAE[a]
ADME Boxes	ap-algorithms.com	627	0.959	0.32
VCCLAB	vcclab.org	610	0.931	0.40
ADMET Predictor	simulationsplus.com	653	0.899	0.67
Pipeline Pilot	accelrys.com	626	0.852	0.43
SPARC	ibmlc2.chem.uga.edu/sparc	644	0.846	0.78
MARVIN	chemaxon.com	653	0.778	0.90
QikProp	schrodinger.com	645	0.768	0.93
ACD/Labs	Acdlabs.com	644	0.678	1.07
PALLAS	compudrug.com	646	0.656	1.17
CSpK_a	chemsilico.com	642	0.565	1.48
not surveyed				
ASTER	www.epa.gov			
COSMOtherm	www.commologic.de			
Epik	www.schrodinger.com			
Jaguar	www.schrodinger.com			
MoKa	www.moldiscovery.com			

[a] MAE - mean absolute error.

Empirical pK_a predictors are capturing ever increasing attention, given the vast amount of available data. We once again refer to Table 4 for a comprehensive list of large data sources. While the data available to the public may be vast, they are by no means comprehensive, as is shown by a survey of several commercial and public small molecule pK_a predictors made by Dearden et al.[191] Chem Silico provided a 653 molecule test set for the survey and verified that it was not used in training their model, but it could not be verified what portion of the data was external to the other models. Table 7 is a reproduction of this survey in hopes that it, along with the broader survey in Table 6, can help the reader select the most suitable software. No statistics for training and or validation data were found for Pallas, Pipeline Pilot, and QikProp. Furthermore, the Chem Silico pK_a prediction utility is no longer described or offered on their Web site. It should also be noted that the only software packages which provided predictions for the entire data set were ADMET Predictor and MARVIN. One other hidden anomaly in the results is that VCCLAB uses pK_a prediction data from ADME Boxes. VCCLAB obtains log$P$, log$S$, and pK_a predictions from ADME Boxes as part of their suite of properties returned by their Web utility ALOGPS.[192] For all of the molecules that were predicted in common between ADME Boxes and VCCLAB, the performance was equivalent. Apparently the performance differences are due to either differing SMILES interpretations or a transmission problem between the two Web sites.

Benchmarking models is a major issue in literature. To date, no true benchmarks for pK_a exist. Ideally, one would have a universal training set to train all models, and a universal disjoint and similarly diverse test set would be used to test their predictions. Even *ab initio* methods may be based on some small subset of the universal training set. A more practical way to compare two methods would be to (1) examine the fit of both on the intersection of their training sets and then (2) compare their predictions on a test set outside the union of their training sets.

With no true benchmarks for pK_a prediction utilities, the only way to identify a superior model is by trusting the statistics. All empirically based models should have r^2 close to 1.0 and RMSE as close to 0.0 as possible over a wide range of compounds. The statistics for both training and test data should be separate; unfortunately this is not the case with current surveys for both macromolecules[11,68,69] and small molecules,[193] including the one in this review. Consensus models further confuse the issue, as the training sets of all models are considered and the test set of molecules under consideration is more likely to be represented by one or more of those data sets. Statistics on the training data indicate the upper threshold for accuracy, while statistics on the test set (data outside the training set) suggest how the model will perform on data having similar chemical diversity. Therefore, it would be useful to have a diversity statistic based on some chemical property space defined by calculable orthogonal descriptors based on a large chemical database such as PubChem. The other obvious but nonetheless relevant aspect of the training and test sets are their respective sizes. Empirical pK_a models with small training sets and good statistics are either specific to an ionizable group, such as carboxylic acids, or their accuracy cannot be trusted without more rigorous validation. Robust evaluation of predictive models comes down to five factors: the statistics, the range, and variance of the property values and the size and chemical diversity of the test set. As shown in Figure 1, clustering of data in chemical property space can drastically affect the statistics, especially leave-some-out cross-validation. When choosing the method(s) keep the following questions in mind. *How chemically diverse is the data set?* Testing on substituted phenols alone only indicates how well the pK_a of phenols will be predicted. *What is the range and distribution of experimental values of the external test set?* It is impossible to trust predictions unless validation across the entire pK_a spectrum of the user's desired application is performed. For example, it is useless to only validate pK_a predictions in the range of 10 to 16, if the intended application is pharmaceutical research relating to the oral bioavailability of potential lead candidates. *What is the size of the test set?* Large and diverse test sets not only validate the accuracy of a predictive model but also its robustness. *What can be learned from the outliers (poor predictions)?* If model A has superior performance on acids and model B has superior performance on bases, it is common sense to use both in their respective areas of strength. If it is not possible to discern a superior group of models, it has been shown that consensus models tend to improve accuracy,[69,90] apparently due to the increased diversity of the combined training sets of the models used in the consensus.

6. CONCLUSIONS

The main advantage of *in silico* pK_a prediction is that physical samples are not needed. Still, some new compounds surely need to be synthesized for experimental evaluation of physicochemical properties to better understand chemical space and expand the diversity of the molecules available to update existing models and develop new prediction methods. Even when one considers newer methods of high throughput pK_a, there are two limiting factors: the costs and time associated with obtaining or synthesizing the molecules of interest. Hence, there is a need for a quick, accurate, and robust model for pK_a prediction for large as well as small molecular libraries. There is also a need for better benchmarking and comparison of methods. For example, the common belief is that consensus models tend to improve the accuracy of predictions. Here, common sense should throw up red flags. At least with the methods and software discussed above, there is no comprehensive database indicating what molecules were used for training and testing. Therefore, the consensus could be based on a selection of methods where some or all of the training sets included the molecule being evaluated. Statistics obtained from a consensus model may not reflect its performance on new data.

So, which method is best? While *ab initio* quantum mechanical methods are broadly applicable, they are computationally expensive. With regard to small molecules, QM descriptors are easier to calculate, but they suffer some of the same limitations as Hammett based methods, as one needs to first group compounds and establish linear correlations between the descriptors and pK_a. Unlike the Hammett and Taft equation, a σ-like variable needs to be established for each descriptor type for each class of compounds. In order to accurately predict pK_a using QM methods, it is clear that solvation needs to be considered for some if not all compound classes. Any improved accuracy invariably is associated with large computational cost; hence these methods may be impractical when calculating pK_a for large *in silico* molecular databases. On the other hand, QM analysis can identify the sites and order of ionization, which many empirical methods cannot. Furthermore, QM may be used to provide added insight in the analysis of microspecies. It has been shown that QM continuum-solvation methods are still a viable tool for providing predictive regression equations for various chemical classes. While it may not represent the fastest solution to high throughput *in silico* pK_a analysis, it would behoove us to identify a single comprehensive level of DFT and solvation theory such that a single equation could deal with all compound classes.

Again, which method is best? It is impossible to know until true assessments have been made. Being able to fairly assess these models is paramount, but how? Data are limited, fallible, hidden, unorganized, and often found to be conflicting, yet it is the basis of each and every model discussed in this review. These factors are compounded when considering pK_a models for macromolecules. NMR titrations are by far the most difficult, and it is not obvious which shift (^{1}H, ^{13}C, ^{15}N) will provide the most interpretable titration curve. In some cases the titration curves were obtained by measuring the shifts of atoms in the vicinity of the ionizable site but far removed considering covalent bond connectivity. The main limiting factor in order to improve pK_a prediction is the need for a large quantity of new well curated data for both small and macromolecules. Especially, more data are needed for buried ionizable residues and ionizable residues in active sites of proteins. Steps toward a comprehensive pK_a database have already begun with the PPD, but the data cannot be efficiently downloaded into a text delimited or other common computer readable format such as SD files. There is need for a similar database for experimental pK_a data for small molecules. The initial challenge is to collect and curate all of the freely available information and then

offer data to the public in an organized and computer accessible format.

Proprietary data are still an issue, but it has already been seen that Big Pharma has been willing to participate in the assessment of predictive utilities on their own proprietary data sets.[194] These tests, while relevant to the pharmaceutical industry, may not be a fair assessment of the overall predictive quality of the models being tested. The proprietary data sets are likely to contain highly skewed sets of molecules based on the investigation of structure activity relationships. Therefore, it is likely that the proprietary data sets do not represent a well distributed sampling of chemical space, resulting in a less than adequate predictive performance of empirical and semiempirical methodologies. A robust predictive model is expected to exhibit uniform performance across a defined segment of chemical space. It is possible and in fact desirable that performance will be maintained outside such definitions, but it cannot be expected.

Accepting a new definition of chemical space, capable of differentiating all known classes of chemical compounds, could serve as a basis for identifying the strengths and weakness of existing physicochemical property prediction utilities. This is important, as different methodologies are likely to demonstrate higher accuracy when accessing molecules in various localized regions of chemical space. Such an analysis would allow researchers to select the optimal combination of predictive models for their specific purposes.

A final note: regardless of the model used, validation on an external test set is necessary. In this regard, we would draw the reader's attention to an article entitled *Beware* q^2*!*,[195] where leave one out (LOO) cross-validation was explored utilizing k nearest neighbors QSARs for three data sets. All-in-all 160 LOO models for each of the data sets were explored, very few of which were found to have desirable statistics for predicting new data, and it was impossible to identify the best LOO model without assessing it on new data. In order to validate that their models did not suffer from overfitting, chance correlations were explored by performing 160 randomizations on each data set and respectively retraining the dependent variables for each randomization. It was verified that the q^2 for chance correlations were significantly lower than those of the trained models derived from the nonrandomized dependent variables. Furthermore, there was little to no correlation between the q^2 of the cross-validated training sets and the respective r^2 on the external test data. The authors concluded that LOO cross-validation could not be used to identify a robust model, nor could it be used to identify the best model for making predictions without validation on an external test set consisting of new data. We again emphasize, the statistics for a model based on fitting training data represents the maximum predictive power for that model and in no way determines how the model will predict new data.

REFERENCES AND NOTES

(1) Ullmann, G. M. Relations between protonation constants and titration curves in polyprotic acids: A critical view. *J. Phys. Chem. B* **2003**, *107*, 1263–1271.

(2) Prasad, R.; Mahajan, V.; Verma, S.; Gupta, N. Arterial blood gas: Basics and interpretation. *Pulmon* **2007**, *9*, 82–87.

(3) Hoener, B. A.; Benet, L. Z. In *Modern Pharmaceutics*, 3rd ed.; Banker, G. S., Rhodes, C. T., Eds.; Marcel Dekker Inc.: New York, 1996; pp 121−153.

(4) Nielsen, J. E. Analyzing protein NMR pH-titration curves. In *Annual Reports in Computational Chemistry*, 1st ed.; Wheeler, R. A., Spellmeyer, D. C., Eds.; Elsevier: Amsterdam, The Netherlands, 2008; Vol. 4, pp 89−106.

(5) Quijada, J.; López, G.; Versace, R.; Ramírez, L.; Tasayco, M. L. On the NMR analysis of pK_a values in the unfolded state of proteins by extrapolation to zero denaturant. *Biophys. Chem.* **2007**, *129*, 242–250.

(6) Bartik, K.; Redfield, C.; Dobson, C. M. Measurement of individual pK_a values of acidic residues of hen and turkey lysozymes by two-dimensional ^{1}H NMR. *Biophys. J.* **1994**, *66*, 1180–1184.

(7) Oliveberg, M.; Arcus, V. L.; Fersht, A. R. pK_a Values of carboxyl groups in the native and denatured states of barnase: The pK_a values of the denatured state are on average 0.4 units lower than those of model compounds. *Biochemistry* **1995**, *34*, 9424–9433.

(8) Thurkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. Hydrogen bonding markedly reduces the pK of buried carboxyl groups in proteins. *J. Mol. Biol.* **2006**, *362*, 594–604.

(9) Protein pK_a Database. http://www.jenner.ac.uk/PPD/ (accessed on June 8, 2009).

(10) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. Prediction of pH-dependent properties of proteins. *J. Mol. Biol.* **1994**, *238*, 415–436.

(11) Stanton, C. L.; Houk, K. N. Benchmarking pK prediction methods for residues in proteins. *J. Chem. Theory Comput.* **2008**, *4*, 951–966.

(12) Krieger, E.; Nielsen, J. E.; Spronk, C. A. E. M.; Vriend, G. Fast empirical pK_a prediction by Ewald summation. *J. Mol. Graphics Modell.* **2006**, *25*, 481–486.

(13) Nielsen, J. E.; Vriend, G. Optimizing the hydrogen-bond network in Poisson-Boltzmann equation-based pK_a calculations. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 403–411.

(14) Edsall, J. T. In *Proteins, amino acids and peptides as ions and dipolar ions*, 1st ed.; Cohn, E. J., Ed.; Reinhold Publishing Corp.: New York, 1943; Chapter *20*, pp 444−505.

(15) Nozaki, Y.; Tanford, C. Examination of titration behavior. In *Methods in Enzymology*, 1st ed.; Hirs, C. H. W., Ed.; Academic Press: New York, 1967; Vol. 11, pp 715−734.

(16) Keim, P.; Vigna, R. A.; Morrow, J. S.; Marshall, R. C.; Gurd, F. R. N. Carbon 13 nuclear magnetic resonance of pentapeptides of glycine containing central residues of serine, threonine, aspartic and glutamic acids, asparagine, and glutamine. *J. Biol. Chem.* **1973**, *248*, 7811–7818.

(17) Keim, P.; Vigna, R. A.; Nigen, A. M.; Morrow, J. S.; Gurd, F. R. N. Carbon 13 nuclear magnetic resonance of pentapeptides of glycine containing central residues of methionine, proline, arginine, and lysine. *J. Biol. Chem.* **1974**, *249*, 4149–4156.

(18) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. pK values of the ionizable groups in proteins. *Protein Sci.* **2006**, *15*, 1214–1218.

(19) Richarz, R.; Wüthrich, K. Carbon-13 NMR chemical shifts of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH. *Biopolymers* **1978**, *17*, 2133–2141.

(20) Creighton, T. E. In *Proteins: Structures and molecular properties*, 2nd ed.; W. H. Freeman and Company: New York, 1993; p 6.

(21) He, Y.; Xu, J.; Pan, X.-M. A statistical approach to the prediction of pK_a values in proteins. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 75–82.

(22) Li, H.; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and rationalization of protein pK_a values. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 704–721.

(23) Forsyth, W. R.; Antosiewicz, J. M.; Robertson, A. D. Empirical relationships between protein structure and carboxyl pK_a values in proteins. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 388–403.

(24) Edgcomb, S. P.; Murphy, P. M. Variability in the pK_a of histidine side-chains correlates with burial within proteins. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 1–6.

(25) Fitch, C. A.; García-Moreno, E. B. Structure-based pK_a calculations using continuum electrostatics methods. *Curr. Prot. Bioinf.* **2006**, 8.11.18.11.22.

(26) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; McCammon, J. A. Electrostatics and diffusion of molecules in solution: Simulations with the University of Houston Brownian Dynamics Program. *Comput. Phys. Commun.* **1995**, *91*, 57–95.

(27) Bashford, D. Macroscopic electrostatic models for protonation states in proteins. *Front. Biosci.* **2004**, *9*, 1082–1099.

(28) Fogolari, F.; Brigo, A.; Molinari, H. The Poisson-Boltzmann equation for biomolecular electrostatics: A tool for structural biology. *J. Mol. Recognit.* **2002**, *15*, 377–392.

(29) Bashford, D.; Karplus, M. pK_as of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. *Biochemistry* **1990**, *29*, 10219–10225.

(30) Yang, A.-S.; Gunner, M. R.; Samponga, R.; Sharp, K.; Honig, B. On the calculation of pK_as in proteins. *Proteins: Struct., Funct., Genet.* **1993**, *15*, 252–265.

(31) Hill, T. On intermolecular and intramolecular interactions between independent pairs of binding sites in proteins and other molecules. *J. Am. Chem. Soc.* **1956**, *78*, 3330–3336.

(32) Tanford, C.; Kirkwood, J. Theory of protein titration curves. I. General equations for impenetrable spheres. *J. Am. Chem. Soc.* **1957**, *79*, 5333–5339.

(33) Warshel, A. Energetics of enzyme catalysis. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 5250–5244.

(34) Warshel, A. What about protein polarity. *Nature* **1987**, *330*, 15–16.

(35) Elcock, A. H. Prediction of functionally important residues based solely on computed energetics of protein structure. *J. Mol. Biol.* **2001**, *312*, 885–896.

(36) Ondrechen, M. J.; Clifton, J. G.; Ringe, D. THEMATICS: A simple computational predictor of enzyme function from structure. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 12473–12478.

(37) Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995**, *268*, 1144–1149.

(38) Barker, P. D.; Mauk, A. G. pH-Linked conformational regulation of a metalloprotein oxidation-reduction equilibrium: Electrochemical analysis of the alkaline form of cytochrome *c*. *J. Am. Chem. Soc.* **1992**, *114*, 3619–3624.

(39) Turano, P.; Ferrer, J. C.; Cheesman, M. R.; Thomson, A. J.; Banci, L.; Bertini, I.; Mauk, A. G. pH, electrolyte, and substrate-linked variation in active site structure of the Trp5l Ala variant of cytochrome c peroxidase. *Biochemistry* **1995**, *34*, 13895–13905.

(40) Alexov, E. G.; Gunner, M. R. Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophys. J.* **1997**, *74*, 2075–2093.

(41) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. The determinants of p*K*s in proteins. *Biochemistry* **1996**, *35*, 7819–7833.

(42) Demchuk, E.; Wade, R. C. Improving the continuum dielectric approach to calculating p*K*s for ionizable groups in proteins. *J. Phys. Chem.* **1996**, *100*, 17373–17387.

(43) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. Combining conformational flexibility and continuum electrostatics for calculating pK_as in proteins. *Biophys. J.* **2002**, *83*, 1731–1748.

(44) Czodrowski, P.; Dramburg, I.; Sotriffer, A.; Klebe, G. Deelopment, validation, and application of adapted PEOE charges to estimate pK_a values of functional groups in protein-ligand complexes. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 424–437.

(45) Barth, P.; Alber, T.; Harbury, P. B. Accurate, conformation-dependent predictions of solvent effects on protein ionization constants. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4898–4903.

(46) Dimitrov, R. A.; Crichton, R. R. Self-consistent field approach to protein structure and stability. I: pH dependence of electrostatic contribution. *Proteins: Struct., Funct., Genet.* **1997**, *27*, 576–596.

(47) 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.

(48) Warwicker, J. Improved pK_a calculations through flexibility based sampling of water-dominated interaction scheme. *Protein Sci.* **2004**, *13*, 2793–2805.

(49) Sham, Y. Y.; Chu, Z. T.; Warshel, A. Consistent calculation of p*K*s of ionizable residues in proteins: Semi-microscopic and microscopic approaches. *J. Phys. Chem. B* **1997**, *101*, 4458–4472.

(50) Sandberg, L.; Edholm, O. A fast simple method to calculate protonation states in proteins. *Proteins: Struct., Funct., Genet.* **1999**, *36*, 474–483.

(51) Mehler, E. L.; Guarnieri, F. A self-consistent, microenvironment modulated screened coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins. *Biophys. J.* **1999**, *75*, 3–22.

(52) Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem.* **2004**, *25*, 2038–2048.

(53) Kuhn, B.; Kollman, P. A.; Stahl, M. Prediction of pK_a shift in proteins using a combination of molecular mechanical and continuum solvent calculations. *J. Comput. Chem.* **2004**, *25*, 1865–1872.

(54) Pokala, N.; Handel, T. M. Energy functions for protein design I: Efficient and accurate continuum electrostatics and solvation. *Protein Sci.* **2004**, *13*, 925–936.

(55) Spassov, V. Z.; Yan, L. A fast accurate computational approach to protein ionization. *Protein Sci.* **2008**, *17*, 1955–1970.

(56) Khandogin, J.; Brooks, C. L., III. Toward the accurate first-principles prediction of ionization equilibria in proteins. *Biochemistry* **2006**, *45*, 9363–9373.

(57) Wisz, M. S.; Hellinga, H. W. An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 360–377.

(58) Karshikoff, A. A simple algorithm for the calculation of multiple site titration curves. *Protein Eng.* **1995**, *8*, 243–248.

(59) Antosiewicz, J.; Briggs, J. M.; Elcock, A. H.; Gilson, M. K.; McCammon, J. A. Computing ionization states of proteins with a detailed charge model. *J. Comput. Chem.* **1996**, *17*, 1633–1644.

(60) Gibas, C. J.; Subramaniam, S. Explicit solvent models in protein pK_a calculations. *Biophys. J.* **1996**, *71*, 138–147.

(61) Warshel, A.; Sharma, P. K.; Kato, M.; Parson, W. W. Modeling electrostatic effects in proteins. *Biochim. Biophys. Acta* **2006**, *1764*, 1647–1676.

(62) Simonson, T.; Perahia, D. Microscopic dielectric properties of cytochrome c from molecular dynamics in aqueous solution. *J. Am. Chem. Soc.* **1995**, *117*, 7987–8000.

(63) Schutz, C. N.; Warshel, A. What are the dielectric "constants" of proteins and how to validate electrostatic models. *Proteins: Struct., Funct., Genet.* **2001**, *44*, 400–417.

(64) Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; Garcia-Moreno, E. B. Experimental pK_a values of buried residues: Analysis with continuum methods and role of water penetration. *Biophys. J.* **2002**, *82*, 3289–3304.

(65) Nielsen, J. E.; Andersen, K. V.; Honig, B.; Hooft, R. W. W.; Klebe, G.; Vriend, G.; Wade, R. C. Improving macromolecular electrostatics calculations. *Protein Eng.* **1999**, *12*, 657–662.

(66) Alexov, E. G.; Gunner, M. R. Calculated protein and proton motions coupled to electron transfer: Electron transfer from Q_A- to Q_B in bacterial photosynthetic reaction centers. *Biochemistry* **1999**, *38*, 8253–8270.

(67) Simonson, T.; Carlsson, J.; Case, D. A. Proton binding to proteins: pK_a calculations with explicit and implicit solvent models. *J. Am. Chem. Soc.* **2004**, *126*, 4167–4180.

(68) Kieseritzky, G.; Knapp, E.-W. Optimizing pK_a computation in proteins with pH adapted conformations. *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 1335–1348.

(69) Davies, M. N.; Toseland, C. P.; Moss, D. S.; Flower, D. R. Benchmarking pK_a prediction. *BMC Biochem.* [Online] **2006**, *7*, Article 18. http://www.biomedcentral.com/1471-2105/7/18 (accessed Jun 09, 2009).

(70) Vriend, G. WHAT IF: A molecular modeling and drug design program. *J. Mol. Graph.* **1990**, *8*, 52–56. http://swift.cmbi.kun.nl/whatif/ (accessed May 27 2009).

(71) Edinger, S. R.; Cortis, C.; Shenkin, P. S.; Friesner, R. A. Solvation free energies of peptides: Comparison of approximate continuum solvation models with accurate solution of Poisson-Boltzman equation. *J. Phys. Chem.* **1997**, *101*, 1190–1197.

(72) Warshel, A.; Russell, S. T.; Churg, A. K. Macroscopic models for studies of electrostatic interactions in proteins: Limitations and applicability. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 4785–4789.

(73) Rekker, R. F. In *The hydrophobic fragmental constant, its derivation and application: A means of characterizing membrane systems*; Elsevier Scientific Pub. Co.: New York, 1977.

(74) Rekker, R. F.; Kort, H. M. The hydrophobic fragmental constant: An extension to a 1000 data point set. *Eur. J. Med. Chem.* **1979**, *14*, 479–488.

(75) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.

(76) Onufriev, A.; Case, D. A.; Bashford, D. Effective Born radii in the generalized Born approximation: The importance of being perfect. *J. Comput. Chem.* **2002**, *23*, 1297–1304.

(77) Spassov, V. Z.; Bashford, D. Multiple-site ligand binding to flexible macromolecules: Separation of global and local conformational change and iterative mobile clustering approach. *J. Comput. Chem.* **1999**, *20*, 1091–1111.

(78) RCSB Protein Data Bank. http://www.rcsb.org/pdb/home/home.do (accessed Jul 15, 2009).

(79) Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. Prediction and rationalization of protein pK_a values using QM and QM/MM methods. *J. Phys. Chem. A* **2005**, *109*, 6634–6643.

(80) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very fast prediction and rationalization of pK_a values for protein-ligand complexes. *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 765–783.

(81) Marosi, A.; Kovács, Z.; Béni, S.; Kökösi, J.; Noszál, B. Triprotic acid-base microequilibria and pharmacokinetic sequelae of cetirizine. *Eur. J. Pharm. Sci.* **2009**, *37*, 321–328.

(82) *Calculator Plugins for structure property prediction, Marvin version 5.1.4*; ChemAxon: Budapest, Hungary, 2009. http://www.chemaxon.com/demosite/marvin/index.html (accessed May 7, 2008).

(83) *ACD/PhysChem Suite, version 12.0*; Advanced Chemistry Development Inc.: Toronto, Ontario, Canada, 2009.

(84) *ADMET Predictor, version 3.0*; Simulations Plus, Inc.: Lancaster, CA, 2009.

(85) *MDL CrossFire commander, Version 7.1*; Elsevier MDL: San Leandro, CA, 2009.

(86) 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).

(87) SPARC Performs Automated Reasoning in Chemistry v4.2. http://ibmlc2.chem.uga.edu/sparc/ (accessed Dec 16, 2008).

(88) Advanced Chemistry Development ACD/Labs Online (I-Lab). http://www.acdlabs.com/ilab/ (accessed May 7, 2008).

(89) *SciFinder Scholar, version 2007*; Chemical Abstract Services: Columbus, OH, 2007.

(90) Lee, A. C.; Yu, J.-Y.; Crippen, G. M. pK_a prediction of monoprotic small molecules the SMARTS way. *J. Chem. Inf. Model.* **2008**, *48*, 2042–2053.

(91) Fraczkiewicz, R. In silico prediction of ionization. In *Comprehensive Medicinal Chemistry II*; Testa, B., van de Waterbeemd, H., Eds.; Elsevier: Oxford, U.K., 2006; Vol. 5, Chapter 25, pp 603−626.

(92) Martell, A. E.; Smith, R. M. In *Critical Stability Constants*; Plenum Press: New York, NY, 1974, Vols. *1*−*6*.

(93) *NIST Standard Reference Database 46, version 8.0*; National Institute of Standards and Technology: Gaithersburg, MD, 2009. http://www.nist.gov/srd/nist46.htm (accessed 07/15/2009).

(94) 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. Comput. Sci.* **1992**, *32*, 244–255.

(95) Szabadváry, F. Electrometric Analysis. In *History of Analytic Chemistry*, 1st English ed.; Belcher, R., Gordon, L., Eds.; Pergamon Press: Long Island City, NY, 1966; Vol. 26, pp 375−387.

(96) Barbosa, J.; Barrón, D.; Jiménez-Lozano, E.; Sanz-Nebot, V. Comparison between capillary electrophoresis, liquid chromatography, potentiometric and spectrophotometric techniques for evaluation of pK_a values of zwitterionic drugs in acetonitrile-water mixtures. *Anal. Chim. Acta* **2001**, *437*, 309–321.

(97) Kim, H.-s.; Chung, T. D.; Kim, H. Voltametric determination of the pK_a of various acids in polar aprotic solvents using 1,4-benzoquinone. *J. Electroanal. Chem.* **2001**, *498*, 209–215.

(98) Ishihama, Y.; Oda, Y.; Asakawa, N. Microscale determination of dissociation constants of multivalent pharmaceuticals by capillary electrophoresis. *J. Pharm. Sci.* **1994**, *83*, 1500–1507.

(99) Kolthoff, I. M. Conductometric titrations. *Ind. Eng. Chem.* **1930**, *2*, 225–230.

(100) Niazi, S. Dissociation, Partitioning, and Solubility. In *Handbook of Preformulation: Chemical, Biological, and Botanical Drugs*, 1st ed.; Informa Healthcare: New York, 2007; pp 112−115.

(101) Beckers, J. L.; Everaerts, F. M.; Ackermans, M. T. Determination of absolute mobilities, p*K* and separation numbers by capillary zone electrophoresis. Effective mobility as a parameter for screening. *J. Chromatogr. A* **1991**, *537*, 407–428.

(102) Cai, J.; Smith, T.; Rassi, Z. E. Determination of the ionization constants of weak electrolytes by capillary zone electrophoresis. *J. High Resolut. Chromatogr.* **1992**, *15*, 30–32.

(103) Cleveland, J. A., Jr.; Benko, M. H.; Gluck, S. J.; Walbroehl, Y. M. Automated pK_a determination at low solute concentrations by capillary electrophoresis. *J. Chromatogr., A* **1993**, *652*, 301–308.

(104) Kaliszan, R.; Wiczling, P.; Markuszewski, M. J. pH gradient high-performance liquid chromatography: Theory and applications. *J. Chromatogr., A* **2004**, *1060*, 165–175.

(105) Wiczling, P.; Markuzewski, M. J.; Kaliszan, R. Determination of pK_a by pH gradient reversed-phase HPLC. *Anal. Chem.* **2004**, *76*, 3069–3077.

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

(107) Avdeef, A.; Comer, J. E. A.; Thomson, S. J. pH-metric log *P*. 3. Glass electrode calibration in methanol-water applied to pK_a determination of water-insoluble substances. *Anal. Chem.* **1993**, *65*, 42–49.

(108) Rosés, M.; Rived, F.; Bosch, E. Dissociation constants of phenols in methanol-water mixtures. *J. Chromatogr., A* **2000**, *867*, 45–56.

(109) Ruiz, R.; Ràfols, C.; Rosés, M.; Bosch, E. A potentially simpler approach to measure pK_a of insoluble basic drugs containing amino groups. *J. Pharm. Sci.* **2003**, *92*, 1473–1481.

(110) Harris, J. C.; Hayes, M. J. Acid dissociation constant. In *Handbook of Chemical Property Estimation Methods*; Lyman, W. J., Reehl, W. F., Rosenblatt, D. H., Eds.; McGraw-Hill, Inc.: New York, 1982; pp 6.1−6.28.

(111) Chen, J.-F.; Xia, Y.-X.; Choppin, G. R. Derivative analysis of potentiometric titration data to obtain protonation constants. *Anal. Chem.* **1996**, *68*, 3973–3978.

(112) Goss, K.-U.; Bronner, G.; Harner, T.; Hertel, M.; Schmidt, T. C. Partition behavior of fluorotelomer alcohols and olefins. *Environ. Sci. Technol.* **2006**, *40*, 3572–3577.

(113) Goss, K.-U. The pK_a values of PFOA and other highly fluorinated carboxylic acids. *Environ. Sci. Technol.* **2008**, *42*, 456–458.

(114) Lee, L. S.; Bellin, C. A.; Pinal, R.; Rao, P. S. C. Cosolvent effects on sorption of organic acids by soils from mixed-solvents. *Environ. Sci. Technol.* **1993**, *27*, 165–171.

(115) Clark, J.; Perrin, D. D. Prediction of the strengths of organic bases. *Q. Rev. Chem. Soc.* **1964**, *18*, 295–320.

(116) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. In *pK_a Prediction for Organic Acids and Bases*; Chapman & Hall: New York, 1981.

(117) Livingstone, D. J. Theoretical property predictions. *Curr. Top. Med. Chem.* **2003**, *3*, 1171–1192.

(118) Klopman, G.; Fercu, D. Application of the multiple computer automated structure evaluation methodology to a quantitative structure-activity relationship study of acidity. *J. Comput. Chem.* **1994**, *15*, 1041–1050.

(119) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A software program for pK_a prediction and protonation state generation for drug-like molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681–691.

(120) EPIK, version 2.0109; Schrödinger LLC: New York, NY, 2009. http://www.schrodinger.com (accessed 05/18/2009).

(121) 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.

(122) 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.

(123) 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.

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

(125) Soriano, E.; Cerdán, S.; Ballesteros, 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.

(126) 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.

(127) 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.

(128) 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.

(129) 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. (THEOCHEM)* **2007**, *805*, 27–32.

(130) CS_prpKa; ChemSilico: Tewksbury, MA, 2008. http://www.chemsilico.com/CS_prpKa/PKAexp.html (accessed Mar 11, 2008 - no longer available).

(131) Habibi-Yangjeh, A.; Pourbasheer, E.; Danandeh-Jenagharad, M. Application of principal component-genetic algorithm-artificial neural network for prediction acidity constant of various nitrogen-containing compounds in water. *Monatsh. Chem.* **2009**, *140*, 15–27.

(132) 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. A* **1998**, *102*, 6706–6712.

(133) Eckert, F.; Klamt, A. Accurate prediction of basicity in aqueous solution with COSMO-RS. *J. Comput. Chem.* **2006**, *27*, 11–19.

(134) 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.

(135) Pompe, M. Variable connectivity index as a tool for solving the 'anti-connectivity' problem. *Chem. Phys. Lett.* **2005**, *404*, 296–299.

(136) Pompe, M.; Randić, M. Variable connectivity model for determination of pK_a values for selected organic acids. *Acta. Chim. Slov.* **2007**, *54*, 605–610.

(137) 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.

(138) Kogej, T.; Muresan, S. Database mining for pK_a prediction. *Curr. Drug Discovery Technol.* **2005**, *2*, 221–229.

(139) Topliss, J. G.; Costello, R. J. Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.* **1972**, *15*, 1166–1068.

(140) Kim, K. H.; Martin, Y. C. Substituent effects from 3D structures using comparative molecular field analysis. 1. Electronic effects of substituted benzoic acids. *J. Org. Chem.* **1991**, *56*, 2723–2729.

(141) *MoKa, version 1.1*; Molecular Discovery Ltd.: Middlesex, U.K., 2009. http://www.moldiscovery.com (accessed June 09, 2009).

(142) Kim, K. H.; Martin, Y. C. Direct prediction of dissociation constants (pK_as) of clonidine-like imidazolines, 2-substituted imidazoles, and 1-methyl-2-substituted imidazoles from 3D structures using a comparative molecular field analysis (CoMFA) approach. *J. Med. Chem.* **1991**, *34*, 2056–2060.

(143) Gargallo, R.; Sotriffer, C. A.; Liedl, K. R.; Rode, B. M. Application of multivariate data analysis methods to comparative molecular field analysis (CoMFA) data: Proton affinities and pK_a prediction for nucleic acids components. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 611–623.

(144) Tehan, B. G.; Lloyd, E. L.; Wong, M. G.; Pitt, W. R.; Montana, J. G.; Manallack, D. T.; Gancia, E. Estimation of pK_a using semiempirical molecular orbital methods. Part 1: Application to phenols and carboxylic acids. *Quant. Struct. Act. Relat.* **2002**, *21*, 457–472.

(145) Tehan, B. G.; Lloyd, E. L.; Wong, M. G.; Pitt, W. R.; Gancia, E.; Manallack, D. T. Estimation of pK_a using semiempirical molecular orbital methods. Part 2: Application to amines, anilines, and various nitrogen containing heterocyclic compounds. *Quant. Struct. Act. Relat.* **2002**, *21*, 473–485.

(146) Seybold, P. G. Analysis of the pK_as of aliphatic amines using quantum chemical descriptors. *Int. J. Quantum Chem.* **2008**, *108*, 2849–2855.

(147) Liu, S.; Pedersen, L. G. Estimation of molecular acidity via electrostatic potential at the nucleus and valence natural atomic orbitals. *J. Phys. Chem. A* **2009**, *113*, 3648–3655.

(148) Gross, K. C.; Seybold, P. G.; Peralta-Inga, Z.; Murray, J. S.; Politzer, P. Comparison of quantum chemical parameters and Hammett constants in correlating pK_a values of substituted anilines. *J. Org. Chem.* **2001**, *66*, 6919–6925.

(149) Ma, Y.; Gross, K. C.; Hollingsworth, C. A.; Seybold, P.; Murray, J. S. Relationships between aqueous acidities and computed surface-electrostatic potentials and local ionization energies of substituted phenols and benzoic acids. *J. Mol. Model.* **2004**, *10*, 235–239.

(150) Parthasarathi, R.; Padmanabhan, J.; Elango, M.; Chitra, K.; Subramanian, V.; Chattaraj, P. K. pK_a prediction using group philicity. *J. Phys. Chem. A* **2006**, *110*, 6540–6544.

(151) Liptak, M. D.; Gross, K. C.; Seybold, P. G.; Feldgus, S.; Shields, G. C. Absolute pK_a determinations for substituted phenols. *J. Am. Chem. Soc.* **2002**, *124*, 6421–6427.

(152) Pliego, J. R.; Riveros, J. M. Theoretical calculation of pK_a using cluster-continuum model. *J. Phys. Chem. A* **2002**, *106*, 7434–7439.

(153) *Jaguar; version 4.2; User Guide*; Schrödinger LLC: New York, NY, 1991−2000. http://yfaat.ch.huji.ac.il/jaguar-help/manTOC.html (accessed May 18, 2009).

(154) Potter, M. J.; Gilson, M. K.; McCammon, J. A. Small molecule pK_a prediction with continuum electrostatics calculations. *J. Am. Chem. Soc.* **1994**, *116*, 10298–10299.

(155) 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.

(156) *ADME/Tox WEB, version 3.5*; Pharma Algorithms: Toronto, ON, Canada, 2008 http://pharma-algorithms.com/webboxes/ (accessed July 9, 2008).

(157) Bickmore, B. R.; Rosso, K. M.; Tadanier, C. J.; Bylaska, E. J.; Doud, D. Bond-valence methods for pK_a prediction. II. Bond-valence, electrostatic, molecular geometry, and solvation effects. *Geochim. Cosmochim. Acta* **2006**, *70*, 4057–4071.

(158) Parthasarathi, R.; Padmanabhan, J.; Elango, M.; Subramanian, V.; Chattaraj, P. K. Intermolecular reactivity through the generalized philicity concept. *Chem. Phys. Lett.* **2004**, *394*, 225–230.

(159) Brinck, T.; Murray, J. S.; Politzer, P.; Carter, R. E. A relationship between experimentally determined pK_as and molecular surface ionization energies for some azines and azoles. *J. Org. Chem.* **1991**, *56*, 2934–2936.

(160) Brinck, T.; Murray, J. S.; Politzer, P. Relationships between the aqueous acidities of some carbon, oxygen, and nitrogen acids and the calculated surface local ionization energies of their conjugate bases. *J. Org. Chem.* **1991**, *56*, 5012–5015.

(161) Brinck, T.; Murray, J. S.; Politzer, P. Molecular surface electrostatic potentials and local ionization energies of group V - VII hydrides and their anions: Relationships for aqueous and gas-phase acidities. *Int. J. Quantum Chem.* **1993**, *48*, 73–88.

(162) Parr, R. G.; Szentpaly, L. V.; Liu, S. J. Electrophilicity index. *J. Am. Chem. Soc.* **1999**, *121*, 1922–1924.

(163) Liptak, M. D.; Gross, K. C.; Seybold, P. G.; Feldgus, S.; Shields, G. Absolute pK_a determinations for substituted phenols. *J. Am. Chem. Soc.* **2002**, *124*, 6421–6427.

(164) Toth, A. M.; Liptak, M. D.; Phillips, D. L.; Shields, G. C. Accurate relative pK_a calculations for carboxylic acids using complete basis set and Gaussian-*n* models combined with continuum solvation methods. *J. Chem. Phys.* **2001**, *114*, 4595–4606.

(165) Barone, V.; Cossi, M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A* **1998**, *102*, 1995–2001.

(166) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G. J. A.; Montgomery, J.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98, revision A.6*; Gaussian, Inc.: Pittsburgh, PA, 1998.

(167) Klamt, A.; Schüürmann, G. COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.

(168) Klamt, A. Conductor-like screening model for real solvents: A new approach to quantitative calculation of solvation phenomena. *J. Phys. Chem.* **1995**, *99*, 2224–2235.

(169) Klamt, A.; Jonas, V.; Brürger, T.; Lohrenz, C. W. Refinement and parametrization of COSMO-RS. *J. Phys. Chem. A* **1998**, *102*, 5074–5085.

(170) Ho, J.; Coote, M. pK_a calculation of some biologically important carbon acids - An assessment of contemporary theoretical procedures. *J. Chem. Theory Comput.* **2009**, *5*, 295–306.

(171) Zupan, J.; Gasteiger, J. In *Neural Networks for Chemists; An Introduction*; VCH Publishers: New York, NY, 1993.

(172) Zupan, J.; Gasteiger, J. In *Neural Networks in Chemistry and Drug Design*; Wiley-VCH Verlag GmbH: Weinheim, Germany FRG, 1999.

(173) Hagan, M. T.; Demuth, H. B.; Beale, M. In *Neural Network Design*; PWS: Boston, MA, 1996; Chapters 1−4. http://hagan.ecen.ceat.okstate.edu/nnd.html (accessed June 7, 2009.)

(174) Khayamian, T.; Kardanpour, Z.; Ghasemi, J. A new application of PC-ANN in spectrophotometric determination of acidity constants of PAR. *J. Braz. Chem. Soc.* **2005**, *16*, 1118–1123.

(175) Sayle, R. Physiological ionization and pK_a prediction. Metaphorics LLC: 2000. http://www.daylight.com/meetings/emug00/Sayle/pkapredict.html (accessed May 18, 2009).

(176) Glen, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *iDrugs* **2006**, *9*, 199–204.

(177) Xing, L.; Glen, R. C. Novel methods for the prediction of log*P*, pK_a, and log*D*. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796–805.

(178) Blower, P. E.; Cross, K. P. Decision tree methods in pharmaceutical research. *Curr. Top. Med. Chem.* **2006**, *6*, 31–39.

(179) Lee, A. C.; Shedden, K.; Rosania, G. R.; Crippen, G. M. Data mining the NCI60 to predict generalized cytotoxicity. *J. Chem. Inf. Model.* **2008**, *48*, 1379–1388.

(180) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. In *Classification and Regression Trees*; Wadsworth Inc.: Belmont, CA, 1984; Chapter *8*, pp 216−265.

(181) Yamazaki, T. NMR and X-ray evidence that the HIV protease catalytic aspartyl groups are protonated in the complex formed by the protease and a non-peptide cyclic urea-based inhibitor. *J. Am. Chem. Soc.* **1994**, *116*, 10791–10792.

(182) Wang, Y. X.; Freedberg, D. I.; Yamazaki, T.; Wingfield, P. T.; Stahl, S. J.; Kaufman, J. D.; Kiso, Y.; Torchia, D. A. Solution NMR evidence that the HIV-1 protease catalytic aspartyl groups have different ionization states in the complex formed with the asymmetric drug KNI-272. *Biochemistry* **1996**, *35*, 9945–9950.

(183) Singer, A. U.; Forman-Kay, J. D. pH titration studies of an SH2 domain-phosphopeptide complex: Unusual histidine and phosphate pK_a values. *Protein Sci.* **1997**, *6*, 1910–1919.

(184) Dullweber, F.; Stubbs, M. T.; Musil, Đ.; Stürzebecher, J.; Klebe, G. Factorising ligand affinity: A combined thermodynamic and crystallographic study of trypsin and thrombin inhibition. *J. Mol. Biol.* **2001**, *313*, 593–614.

(185) Czodrowski, P.; Sotriffer, C. A.; Klebe, G. Protonation upon ligand binding to trypsin and thrombin: Structural interpretation based on pK_a calculations and ITC experiments. *J. Mol. Biol.* **2007**, *367*, 1347–1356.

(186) Czodrowski, P.; Sotriffer, C. A.; Klebe, G. Atypical protonation states in the active site of HIV-1 protease: A computational study. *J. Chem. Inf. Model.* **2007**, *47*, 1590–1598.

(187) Bashford, D. Scientific computing in object-oriented parallel environments. In *Lecture Notes in Computer Science*, 1st ed.; Ishikawa, Y.,

Oldehoeft, R. R., Reynders, J. V. W., Tholburn, M., Eds.; Springer: New York, NY, 1997; Vol. 1343, pp 233−240.
(188) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228.
(189) Sitkoff, D.; Ben-Tal, N.; Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **1994**, *98*, 1978–1988.
(190) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher III, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wirkiewicz-Kuczera, D.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
(191) 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**, *59*, A7.
(192) Virtual Computational Chemistry Lab. http://www.vcclab.org/ (accessed July 9, 2008).
(193) 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.
(194) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log*P* methods on more than 96000 compounds. *J. Pharm. Sci.* **2009**, *3*, 861–893.
(195) Golbraikh, A.; Tropsha, A. Beware q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.

CI900209W