# A Reliable and Efficient First Principles-Based Method for Predicting p$K_a$ Values. 1. Methodology

**Shuming Zhang,[†] Jon Baker,[†,‡] and Peter Pulay*,[†]**

*Department of Chemistry and Biochemistry, University of Arkansas, Fayetteville, Arkansas 72701, and Parallel Quantum Solutions, 2013 Green Acres Road, Suite A, Fayetteville, Arkansas 72703*

*Received: July 15, 2009; Revised Manuscript Received: November 10, 2009*

We have developed an efficient and reliable protocol for the calculation of p$K_a$ values in aqueous solution from density functional calculations. We establish a standard linear regression fit using only calculated energies of deprotonation and experimental p$K_a$ values; all other factors, including most entropic effects, are absorbed into the fitting constants. In this article we fit a small training set of 34 experimentally well-characterized molecules to determine the best level of theory among those tested (i.e., the optimum compromise between efficiency and accuracy for the basis set, the exchange−correlation functional, the (continuum) solvation model and the level of geometry optimization). Our main findings are that a relatively modest basis set (6-311+G**) suffices for the calculation of the energy differences, with an even small basis set (3-21G*) sufficient for the preceding geometry optimization. Using a solvation model (COSMO in our case) throughout is essential to achieve reliable results. The exchange−correlation functional plays only a modest role; in particular, pure DFT functionals that allow the efficient calculation of the Coulomb term are perfectly adequate. The final protocol will be applied subsequently to data sets much larger than commonly used in such studies.

## 1. Introduction

The p$K_a$ value plays a very significant role in many aspects of drug **a**bsorption, **d**istribution, **m**etabolism, and **e**xcretion (ADME). The p$K_a$ is defined as $-\log K_a$, where $K_a$ is the equilibrium constant for the deprotonation reaction. The p$K_a$ and pH values determine the relative concentration of protonated and deprotonated forms through the Henderson−Hasselbalch equation:[1] log([unprotonated])/([protonated]) = pH − p$K_a$. Most drugs contain at least one site that is able to protonate or deprotonate reversibly. If the pH is higher than the p$K_a$, the site is mostly deprotonated; otherwise it is mostly protonated. The ratio of the protonated and deprotonated forms of a molecule largely determines its binding and transport properties. A drug commonly has to pass through at least one biomembrane via passive diffusion or by carrier-mediated uptake before it can produce any biological effect. Only neutral species can easily penetrate the cell membrane because the lipid bilayers in the cell wall have very low permeability for ions and most polar molecules. This is the main reason for the importance of a drug's p$K_a$ value in pharmacokinetics. About 63% of drug molecules listed in the World Drug Index[2] can be ionized between pH 2 and 12.[3]

Under equilibrium conditions, the p$K_a$ is related to the Gibbs free energy of protonation $\Delta G_{aq}$ through

$$\mathrm{p}K_a = \Delta G_{aq}/(RT \ln 10) = \Delta G_{aq}/(2.303RT) \qquad (1)$$

where $R$ is the gas constant and $T$ is the absolute temperature. The computational determination of accurate p$K_a$ values is very demanding, as an error of only 1.36 kcal/mol in $\Delta G_{aq}$ is equivalent to a unit difference in p$K_a$ value at room temperature.

Of course, the p$K_a$ is not the only factor that influences the biological activity of a drug candidate. Other important physicochemical properties include solubility in aqueous and lipid environments and the partition coefficient between the two, structural effects such as resonance, induction, redox potentials, bond types and isomerism, as well as specific characteristics such as molecular size and shape, and stereochemistry. However, the p$K_a$ is probably the most important factor and is often used as a preliminary measure to select suitable drug candidates.

A fully first-principles calculation of p$K_a$ values in the gas phase requires a computationally demanding level of theory together with a complete thermodynamic analysis, including computation of zero-point energies and entropy effects (see for example Topol et al.[4]). Such calculations require large basis sets and a high level of electron correlation.[5] Most early and virtually all high-accuracy solution studies utilize a thermodynamic cycle that corrects the gas-phase deprotonation energy with the solvation energies of the participating species. One motivation for this is that high-accuracy calculations are easier to carry out in the gas phase. However, we have found, in agreement with others,[6,7] that the detour through the gas phase is counterproductive. At the computational levels that are routinely applicable today, the dominant source of error is not the gas-phase deprotonation energy. Any gain from improved accuracy in gas-phase calculations is outweighed by errors caused by geometry and conformational differences between the gas and the solution phase. Therefore, as will be seen, we calculate our energy differences directly in the solvated model.

A fully first principles calculation is not appropriate if the aim is to achieve high throughput, for example, the initial screening of a large drug candidate database. Because of the very high cost of rigorous thermodynamic simulations, solvent effects must be approximated by a continuous solvation model. The calculation of zero-point energies and entropy terms is also too expensive computationally. For this reason, we, like many researchers before us, take only the dominant term, the energy difference between the protonated and deprotonated forms, from quantum chemical calculations. Replacing $\Delta G$ in eq 1 by the

[†] University of Arkansas.

[‡] Parallel Quantum Solutions.

10.1021/jp9067069
Published on Web 12/04/2009

computed enthalpy difference $\Delta H$ assumes that entropic effects cancel, and that solvation is adequately described by the selected solvation model. None of these conditions hold sufficiently to produce reliable results. In addition, the direct use of eq 1 requires the solvation energy of the proton which is still not well established.[8] One generally utilized solution to these problems (see, for example, refs 6, 7, and 9) is to approximate the p$K_a$ values by a linear function of $\Delta H$:

$$\mathrm{p}K_a(\mathrm{f}) = \alpha_f \Delta H + \beta_f \tag{2}$$

where *f* denotes a class of ionizable compounds, for instance carboxylic acids. $\Delta H$ for the acid HA is calculated as $E(\mathrm{A}_{aq}^-) - E(\mathrm{HA}_{aq})$ at an appropriate theoretical level. The empirical parameters $\alpha_f$ and $\beta_f$ are determined through a least-squares fit to accurately known experimental p$K_a$ values for representative molecules in the class *f*. These empirical parameters can to a certain extent absorb systematic errors of the quantum chemical and the solvation model, as well as entropic effects. We hope to develop a number of parameter sets (i.e., values of $\alpha_f$ and $\beta_f$) for various functional groups (e.g., carboxylic acids, alcohols, amines, etc.) so that p$K_a$ values can be rapidly and accurately estimated from a couple of relatively straightforward energy calculations.

On the basis of a number of preliminary calculations and literature evaluation, density functional theory (DFT) was selected as the main calculational method. DFT provides a good compromise between accuracy and computational speed. In this work we have addressed several questions:

1. Is pure DFT capable of yielding the same degree of accuracy as hybrid DFT?
2. What is the smallest basis set size required to yield converged relative energies?
3. Is it important to optimize geometries at the DFT level or can we use molecular mechanics geometries?
4. If we need to optimize geometries, what is the smallest basis set that we can use and still get acceptable results?
5. Is it essential to optimize the geometries using a solvation model (specifically the COSMO continuum solvation model[10])?

## 2. Computational Methods

All calculations were done with the PQS program package,[11] initially on a fairly old home-built PC cluster (800 MHz PIV Xeon processors) and subsequently on the University of Arkansas Red Diamond supercomputer.[12]

We selected a set of 34 small molecules for which the p$K_a$ values were experimentally well established. These systems, together with their p$K_a$ values and references, are given in Table 1. Despite the huge number of papers published annually on p$K_a$'s, most experimental values quoted are derived from two large IUPAC compilations published in the 1960s,[13,14] now over forty years ago. There are also supplements[15] and later compilations, e.g., from 1979[16] and 1999,[17] although in the latter no references are provided, so it is not clear how old the data reported actually are. In many cases, these books list more than one experimental reference for a particular p$K_a$ value, and in every one of the 34 molecules we selected there are at least two, and often more, values given that lie within 0.1 p$K_a$ units of each other. In some cases we have supplemented the values reported with more recent experimental data.

The molecules were built at first using the PCModel[18] and later the PQSMol model builder[11] and preoptimized using one of the built-in molecular mechanics force fields. (In the case of PQSMol this was principally Sybyl_5.2.[19]) On this training set we carried out the following series of calculations:

(1) DFT using either (a) the B3LYP[20] or (b) OLYP[21] functional; (2) gas-phase or COSMO (solvation) calculations; (3) geometries either (a) taken directly as constructed, i.e., optimized using molecular mechanics, or optimized with (b) the 3-21G basis set,[22] (c) the 3-21G(*) basis set,[23] (d) the 6-31G* basis set,[24] or (e) the 6-311+G** basis set;[25] (4) energy differences computed using (a) the 6-31+G* basis, (b) the 6-311+G** basis set, or (c) the much larger 6-311++G(3df,3pd) basis set.[26] All possible combinations of (1) DFT functional, (2) gas-phase or COSMO, (3) optimized geometry, and (4) final single-point energy were utilized, for a total of 80 sets of calculations.

There are now a huge number of different DFT functionals; we selected B3LYP as perhaps the most popular and widely used, and OLYP as a general high quality nonhybrid functional known to give good results for organic molecules[27] (where it typically performs far better than the more popular BLYP functional[28]).

The compounds in our training set were divided into four separate data sets, (1) carboxylic acids, (2) alcohols and phenols, (3) pyridines, and (4) anilines and amines, and each set was fitted separately. Final energy differences were plotted against experimental p$K_a$ values and the best least-squares linear fit was obtained. Full tables showing our computed p$K_a$ values and the standard and mean absolute deviations are provided as Supporting Information. Only the mean absolute deviations for selected fits are shown here graphically in Figures 1−4.

The smallest of our four data sets contains only seven compounds. This, of course, is nowhere near enough to derive a pair of reliable $\alpha_f$ and $\beta_f$ values applicable to all compounds in that class. But that is not the aim here. In this initial paper we are simply selecting the best theoretical approach among those that we have chosen to test. Once the theoretical approach has been determined, it will be used to fit much larger data sets and it is these fits that will form the basis of our fast throughput analysis.

In conformationally flexible molecules, several low-energy conformers exist. The correct way of calculating the electronic part of the free energy is to include all low-lying conformers in the thermodynamic averaging. We chose to avoid this route and considered only the energy of the lowest conformer. In a typical worst case, when the ion has only one dominant conformer but the neutral molecule is chiral and has two conformers of equal energy, this approximation leads to an error of $\log 2 = 0.301$ in the p$K_a$ value, assuming that the slope (the $\alpha$ fitting parameter) attains the theoretical value of $(RT \ln 10)^{-1} = 0.734$ mol/kcal. In reality, the fitted slopes are significantly smaller, typically about 0.3 mol/kcal, and thus the error in the predicted p$K_a$ is about 0.12 units. As this is less than the expected accuracy of our method, we chose to ignore it. However, it is not expected to change our final recommendation for the method. (We note that one of the most difficult problems we have encountered is the determination of the minimum energy conformer for larger molecules. In our opinion, previous workers, dealing mainly with small, rigid molecules, have not fully appreciated this problem.)

## 3. Results and Discussion

The first things to note are that p$K_a$ values obtained with raw, unoptimized geometries (i.e., direct from the force field optimizations; Figure 1) are noticeably worse than values obtained using optimized geometries, and in the latter case p$K_a$'s obtained using the COSMO solvation model, as noted in the Introduction, are in turn noticeably better than p$K_a$'s obtained

**TABLE 1: Training Set and Reference p$K_a$ Values**

| molecule | p$K_a$ | reference values |
|---|---|---|
| | | Anilines |
| aniline | 4.61 | 4.63, 4.605, 4.606, 4.620;[14] 4.577, 4.596, 4.56, 4.58;[15] 4.61[45] |
| 3-chloroaniline | 3.52 | 3.46, 3.52;[14] 3.521, 3.52;[15] 3.52[17] |
| 3-(methylsulfonyl)aniline | 2.58 | 2.58, 2.561[14] |
| 4-cyanoaniline | 1.74 | 1.75, 1.739;[14] 1.71[15] |
| 4-methoxyaniline | 5.31 | 5.31, 5.34, 5.3, 5.29;[14] 5.36[17] |
| 4-nitroaniline | 1.00 | 1.00, 1.02, 0.99;[14] 1.019, 0.97,1.00;[15] 1.01[17] |
| | | Amines |
| methylamine | 10.67 | 10.657, 10.68, 10.67, 10.62[14] |
| dimethylamine | 10.77 | 10.73, 10.81, 10.77[14] |
| trimethylamine | 9.81 | 9.81, 9.80, 9.752;[14] 9.801, 9.987;[15] 9.8[17] |
| guanidine | 13.60 | 13.59, 13.6[14] |
| piperidine | 11.11 | 11.22, 11.123, 11.11, 11.06[14] |
| | | Pyridines |
| pyridine | 5.25 | 5.25, 5.21, 5.18;[14] 5.229, 5.21, 5.198, 5.22;[15] 5.31[45] |
| 2-methylpyridine | 5.97 | 5.94, 5.97;[14] 5.957, 6.06[15] |
| 3,4-dimethylpyridine | 6.48 | 6.46, 6.52;[14] 6.48;[15] 6.47[17] |
| 2,4,6-trimethylpyridine | 7.43 | 7.43, 7.48;[14] 7.25;[15] 7.43[17] |
| 3-cyanopyridine | 1.39 | 1.39, 1.36;[14] 1.35[15] 1.45[17] |
| 3-fluoropyridine | 2.97 | 2.97, 3.0;[14] 2.97[17] |
| 4-methoxypyridine | 6.55 | 6.47, 6.58, 6.55;[14] 6.58[15] |
| | | Phenols |
| phenol | 9.98 | 9.98, 9.95, 9.998, 9.991;[13] 9.994, 9.97;[16] 9.81[46] |
| 2-methylphenol | 10.32 | 10.28, 10.15;[13] 10.22, 10.29, 10.32[16] |
| 2-nitrophenol | 7.22 | 7.216, 7.23, 7.234;[13] 7.230, 7.22;[16] 7.21[46] |
| 3-nitrophenol | 8.36 | 8.39, 8.38;[13] 8.355, 8.36;[16] 8.29[46] |
| 2,4-dinitrophenol | 4.12 | 4.09, 4.11, 4.06, 4.02;[13] 4.10, 4.12;[16] 4.14[45] |
| 2,5-dinitrophenol | 5.20 | 5.216, 5.315[13] 5.21, 5.19, 5.20;[16] 5.22[45] |
| 2,6-dinitrophenol | 3.73 | 3.799, 3.710, 3.712, 3.713;[13] 3.695, 3.73[16] |
| | | Alcohols |
| methanol | 15.54 | 15.54;[13] 15.5[16] |
| trifluoroethanol | 12.37 | 12.43, 12.37[13] |
| | | Carboxylic Acids |
| formic acid | 3.76 | 3.74, 3.75;[13] 3.76;[47] 3.81[48] |
| acetic acid | 4.76 | 4.756, 4.754, 4.755;[13] 4.76;[46] 4.74[48] |
| cyanoacetic acid | 2.47 | 2.45, 2.46, 2.58, 2.47;[13] 2.471, 2.50[16] |
| oxalic acid | 1.25 | 1.23, 1.27, 1.25, 1.34, 1.30;[13] 1.252[16] |
| butanoic acid | 4.82 | 4.82;[13] 4.83, 4.80;[16] 4.79[49] |
| benzoic acid | 4.19 | 4.20, 4.18, 4.21;[13] 4.11;[46] 4.18;[50] 4.19[51] |
| 4-nitrobenzoic acid | 3.44 | 3.44, 3.43;[13] 3.426, 3.422, 3.48;[16] 3.74[46] |

using gas-phase (free) energy differences. We can thus concentrate on methods involving optimized geometries with COSMO and select the best overall method that includes these two options.

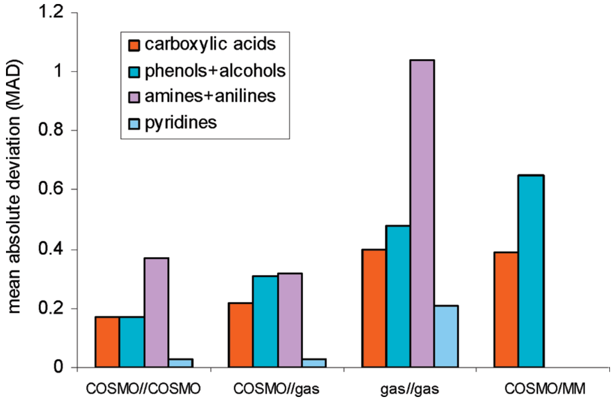

**Figure 1.** Mean absolute deviations from experiment for p$K_a$ values computed at the OLYP/6-311+G**//OLYP/3-21G(d) level: (a) entirely in solution (via COSMO); (b) final single-point energy in solution using geometries optimized in the gas phase; (c) entirely in the gas phase; (d) for carboxylic acids, phenols, and alcohols in solution using unoptimized (i.e., molecular mechanics) geometries.

Of the three basis sets used to obtain the final single-point energies, 6-31+G*, 6-311+G**, and 6-311++G(3df,3pd), it

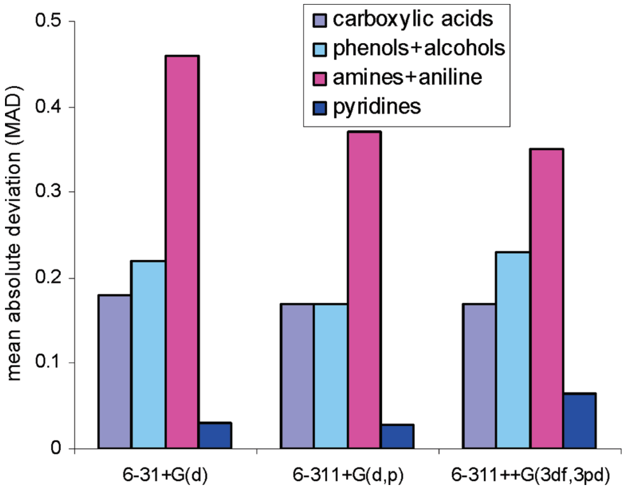

**Figure 2.** Mean absolute deviations from experiment for p$K_a$ values computed at the single-point OLYP level using the OLYP/3-21G(d) geometry optimized in solution (via COSMO) with the (a) 6-31+G*, (b) 6-311+G**, and (c) 6-311++G(3df,3pd) basis sets.

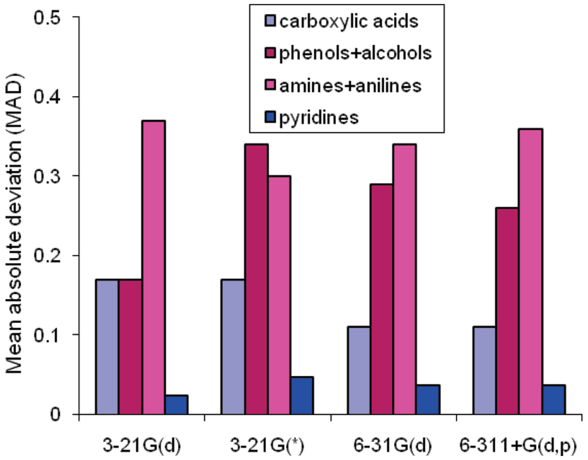

**Figure 3.** Mean absolute deviations from experiment for p$K_a$ values computed at the single-point OLYP/6-311+G** level using geometries optimized in solution (via COSMO) with the (a) 3-21G(d), (b) 3-21G(*), (c) 6-31G*, and (d) 6-311+G** basis sets.

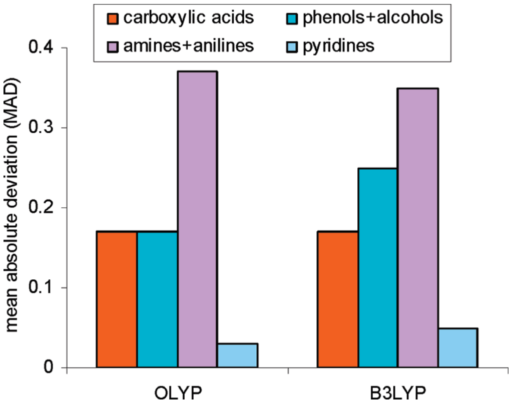

**Figure 4.** Mean absolute deviations from experiment for p$K_a$ values computed at the single-point DFT level using 3-21G(d) geometries optimized in solution (via COSMO) followed by single-point DFT energies using the 6-311+G** basis set with the (a) OLYP and (b) B3LYP functionals.

is clear that, despite comments to the contrary in the literature for methods that attempt to compute absolute p$K_a$ values,[5,29] there is nothing to be gained from using the largest basis set, at least at the DFT level. Neither the mean absolute deviation nor the standard deviation in the predicted p$K_a$ values are any better with the 6-311++G(3df,3pd) basis than with the much smaller 6-311+G** basis, and in several cases they are noticeably worse.

The situation is less clear-cut when it comes to selecting the basis set used to optimize the molecular geometry, but within each set of single-point energies (6-31+G* and 6-311+G**) there is no *overall* accuracy advantage in choosing a larger basis to optimize the geometry than a smaller one; sometimes the larger basis is best, sometimes the smaller (see Figure 3). Given our goal of fast throughput, it is clearly advantageous to optimize geometries using a smaller basis because of the significant computational savings that result; consequently, we propose to use the smallest of our chosen basis sets, 3-21G(d), for this purpose. The geometries of compounds containing the more electronegative first-row elements (N, O, and F) improve significantly in general if d functions are added to the basis set, giving the 3-21G(*) basis,[23] because of a better description of the lone pairs. However, as our data show, this does not result in a better prediction of the p$K_a$.

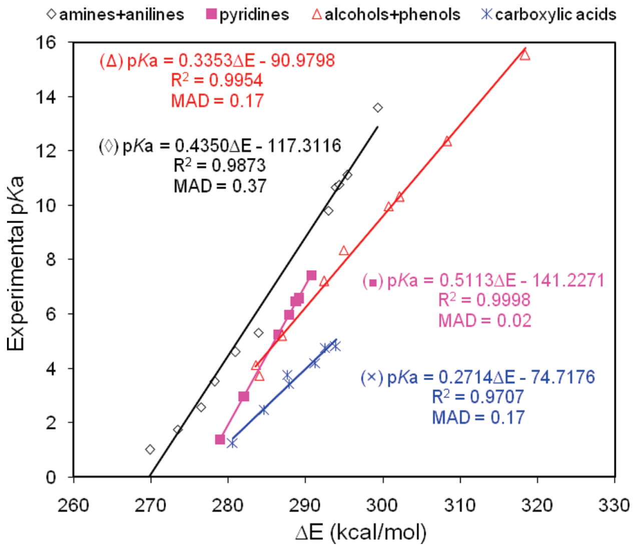

**Figure 5.** Linear regression plots for the final method (OLYP/6-311+G**//3-21G(d); COSMO solvation model with water as solvent).

Overall, we favor the larger 6-311+G** basis over 6-31+G* for the final single-point energy calculations. The mean absolute and standard deviations are at least as good and often better than with the smaller basis in every case (Figure 2).

There is little difference between the two density functionals, OLYP and B3LYP as far as the final p$K_a$ values are concerned (Figure 4), so (as was the case with the basis sets) the least expensive method computationally is favored. OLYP, being a "pure" density functional, can be significantly less expensive than B3LYP, as there is no need to compute the Hartree–Fock exchange. For larger systems and basis sets, the advantage is even more significant because a pure density functional like OLYP allows the use of very efficient algorithms for the Coulomb term, such as the Fourier Transform Coulomb (FTC) method.[30]

Our final method is thus OLYP/6-311+G**//3-21G(d) with the COSMO solvation model (using water as the solvent). We are going to use this method to derive linear regression p$K_a$ equations for over 1000 different molecules. To the best of our knowledge, this is one of the biggest p$K_a$ data sets ever fitted. For the reader's information, we show the linear regression fits we obtained for the final selected method in Figure 5; the results are summarized in detail in Table 2.

## 4. Comparison with Prior work

In this section we compare our approach and results with those of other groups who also used ab initio methods.

In a recent paper Sadlej–Sosnowska compared a number of different methods for calculating absolute p$K_a$ values for nine small, rigid molecules, looking at factors such as the choice of thermodynamic cycle, level of theory, basis set, (continuum) solvation model and, in the latter case, the size of the solvent cavity.[31] One perhaps surprising finding was that the best results were obtained at the Hartree–Fock level. This confirms the results of Liptak et al.[32] for substituted phenols and Murlowska and Sadlej–Sosnowska[33] for substituted tetrazoles. It was also noted by Barone and Cossi, who considered it as being due to underestimation of the solvent reaction field at correlated levels.[34] The good performance of the Hartree–Fock method is less surprising

**TABLE 2: Final Calculated p$K_a$ Values from OLYP/6-311+G(d,p)//3-21G(d) Calculations Using the COSMO Solvation Model**

| molecules | p$K_a^{exp}$ | p$K_a^{cal}$ | Δp$K_a$[a] | molecules | p$K_a^{exp}$ | p$K_a^{cal}$ | Δp$K_a$ |
|---|---|---|---|---|---|---|---|
| Alcohols + Phenols | | | | | | | |
| methanol | 15.54 | 15.79 | 0.25 | trifluoroethanol | 12.37 | 12.38 | 0.01 |
| phenol | 9.98 | 9.85 | −0.13 | 2.4-dinitrophenol | 4.12 | 4.09 | −0.03 |
| 2-methylphenol | 10.32 | 10.32 | 0.00 | 2.5-dinitrophenol | 5.20 | 5.23 | 0.03 |
| 2-nitrophenol | 7.22 | 7.06 | −0.16 | 2.6-dinitrophenol | 3.73 | 4.25 | 0.52 |
| 3-nitrophenol | 8.36 | 7.92 | −0.43 | MAD[b] | | | 0.17 |
| Carboxylic Acids | | | | | | | |
| formic acid | 3.76 | 3.35 | −0.41 | butanoic acid | 4.82 | 5.06 | 0.24 |
| acetic acid | 4.76 | 4.66 | −0.10 | benzoic acid | 4.19 | 4.33 | 0.14 |
| cyanoacetic acid | 2.47 | 2.51 | 0.04 | 4-nitrobenzoic acid | 3.44 | 3.40 | −0.04 |
| oxalic acid | 1.25 | 1.44 | 0.19 | MAD | | | 0.17 |
| Anilines + Amines | | | | | | | |
| aniline | 4.61 | 4.89 | 0.28 | methylamine | 10.67 | 10.52 | −0.15 |
| 3-chloroaniline | 3.52 | 3.71 | 0.19 | dimethylamine | 10.77 | 10.71 | −0.06 |
| 3-(methylsulfonyl)aniline | 2.58 | 2.94 | 0.36 | trimethylamine | 9.81 | 10.12 | 0.31 |
| 4-cyanoaniline | 1.74 | 1.63 | −0.11 | guanidine | 13.60 | 12.89 | −0.71 |
| 4-methoxyaniline | 5.31 | 6.19 | 0.88 | piperidine | 11.11 | 11.17 | 0.06 |
| 4-nitroaniline | 1.00 | 0.07 | −0.93 | MAD | | | 0.37 |
| Pyridines | | | | | | | |
| pyridine | 5.25 | 5.23 | −0.02 | 2-methylpyridine | 5.97 | 5.96 | −0.01 |
| 3-cyanopyridine | 1.39 | 1.39 | −0.00 | 3,4-dimethylpyridine | 6.48 | 6.46 | −0.02 |
| 3-fluoropyridine | 2.97 | 3.01 | 0.04 | 2,4,6-trimethylpyridine | 7.43 | 7.44 | 0.01 |
| 4-methoxypyridine | 6.55 | 6.61 | 0.06 | MAD | | | 0.02 |

[a] Δp$K_a$ = p$K_a^{cal}$ − p$K_a^{exp}$. [b] Mean absolute deviation.

than it appears, given that the number of electron pairs does not change in a deprotonation reaction.

We did not include the Hartree−Fock method in our initial comparison, as it suffers from the same disadvantage as hybrid DFT methods, i.e., the need to calculate the exact exchange, and also because molecular geometries computed at this level of theory are typically noticeably inferior to DFT geometries. However, in view of the above results, we decided to carry out a complete set of Hartree−Fock calculations on all 34 molecules in our training set at HF/6-311+G**//3-21G(d), using the COSMO solvation model. This is identical to our recommended model except that OLYP has been replaced by HF as the theoretical method.

The Hartree−Fock results were indeed quite good, with mean absolute deviations in the fitted p$K_a$ values of 0.15 (carboxylic acids), 0.18 (alcohols/phenols), 0.36 (amines/anilines), and 0.10 (pyridines). These values are similar to the corresponding OLYP values (see Figure 1) except for the pyridines where the Hartree−Fock results are markedly worse. Given the similar accuracy and the potential cost savings for larger systems, there is no reason for us to recommend Hartree−Fock over OLYP. (Details of our calculations are provided in the Supporting Information.)

Further attempts to calculate *absolute* p$K_a$ values will not be reviewed in detail. Recent work, in particular that of Shields and co-workers[32,35] and Yates and co-workers[5] has established that it *is* possible to calculate accurate p$K_a$'s purely from first principles. However, these high-level calculations are not suitable for mass screening of a large number of molecules, for instance putative drug candidates. At lower levels of theory, large errors are obtained. For instance, the calculated absolute p$K_a$ values in ref 31 exhibit mean absolute errors ranging up to 7 p$K_a$ units. Errors of this magnitude are not atypical when attempting to calculate absolute p$K_a$ values[36,37] and support the conventional wisdom, noted in the Introduction, that large basis sets and high levels of theory are required to get even close to experiment (see, e.g., ref 35).

The studies most comparable with ours are those of Klamt et al.[6] and Adam.[7] The approach of Klamt and co-workers differs from ours in two main ways. First they use *electronic* $\Delta G$ in their fitting equation and not $\Delta E_{min}$ as we do. (Vibrational and rotational contributions to the free energy are neglected in ref 6.) Second, they use the proprietary COSMO-RS method (COSMO for real solvents),[38] which involves optimization using basic COSMO in the conductor limit (i.e., infinite dielectric constant), followed by a modeling of the deviations of a real solvent (e.g., water) compared to an ideal conductor using pairwise interacting molecular surfaces, whereas we use the basic COSMO with water as solvent. Their database includes 64 acids (with acidic hydrogens at oxygen and nitrogen only), fitting all compounds with a single linear-regression equation. The fitting has a maximum error of 1.26 p$K_a$ units and a mean unsigned error of 0.37 showing the effectiveness of this kind of approach.

Adam[7] approximates, as we do, free energy differences by energy differences, $E_{A^-} - E_{HA}$, but makes a further approximation by equating this to $-E_H$ using Bader's theory of atoms in molecules (AIM),[39] where $E_H$ is the AIM energy of the ionizable proton in AH, assuming that the AIM energies of the other atoms cancel. He uses the PW91 density functional,[40] COSMO, and the 6-311+G** basis set. Unlike us, however, he carries out a full geometry optimization with this basis set. His results are impressive; for example, for a series of 19 related aliphatic carboxylic acids, his fit gives a mean unsigned error of 0.105 p$K_a$ units, and a maximum unsigned error of 0.342.

We decided to try the PW91 functional and so we repeated our calculations using all 34 molecules in the test set, replacing OLYP with PW91. The mean absolute deviations in the computed p$K_a$ values were 0.16 (carboxylic acids), 0.34 (alcohols/phenols), 0.34 (amines/anilines), and 0.06 (pyridines). These results offer no reason to switch functionals. (A full set of calculated p$K_a$ values, with mean absolute and standard deviations are provided in the Supporting Information.)

Both Klamt et al.[6] and Adam[7] comment in their papers that the slope of their regression fits are significantly lower than the value expected theoretically ($1/2.303RT$). This is apparently a long-standing problem.[41] Our own results behave similarly: at 298 K, the theoretical slope should be 0.733 $\text{kcal}^{-1}$, but all of our fits have slopes much less than this. We will return to this problem in a follow-up paper.

Prior to our work, the paper by Friesner and co-workers[9] was one of the most ambitious efforts to fit a large database of experimental p$K_a$ values through theory. These authors first use a thermodynamic cycle to obtain absolute p$K_a$ values, optimizing geometries in the gas phase at the B3LYP/6-31G* level and computing a single-point energy with the cc-pvtz-f basis.[42] Diffuse functions are added at the reactive center for the negative ion. Solvent effects are included via a self-consistent reaction field (SCRF) method.[43] As this approach was unable to achieve the desired accuracy (an average deviation of 0.5 p$K_a$ units or less), an empirical fit was then imposed on these "raw" p$K_a$ values: p$K_a = A$(p$K_a$ "raw") $+ B$. As in our case, a different fit was used for different functional groups. There are in fact three parameters for each functional group: $A$, $B$, and the ionic radius in the SCRF model. In our opinion, ref 9 overfits the data. In some functional groups, there are only four compounds fitted using three empirical constants. In addition, as our results show, the use of gas-phase geometries worsens the agreement with experiment significantly. Nonetheless the mean absolute deviation in their ~200 compound training set was a respectable 0.41 p$K_a$ units. They subsequently applied their computational protocol to a diverse set of 16 medicinal molecules from the CMC database,[44] obtaining a mean absolute deviation of 0.6 p$K_a$ units with a maximum error of 2.1 p$K_a$ units.

## 5. Conclusions

We have developed a protocol for the first principles calculation of acid dissociation constants that strikes a good compromise between computational expense and accuracy. We have explored several theoretical methods and basis sets in conjunction with the COSMO continuum solvation model. As in previous work, the calculated energy differences of deprotonation are used in an empirical linear regression fit to experimental p$K_a$ values. Our main findings: (1) Pure density functional models, specifically OLYP, that are potentially much more efficient than Hartree–Fock or hybrid functionals, perform as well as the latter. (2) Moderate basis sets, such as 6-311+G**, suffice for the final energy differences. (3) A small basis set, 3-21G*, is appropriate for geometry optimization. (4) It is important to optimize geometries in the presence of a continuum solvent model (COSMO). (5) Methods using high-level gas-phase calculations in conjunction with a thermodynamic cycle do not improve the agreement with experiment. (6) It is essential to use different fitting parameters for different classes of compounds.

**Acknowledgment.** This work has been supported by the National Science Foundation (Grant No. CHE-0515922) and the Mildred B. Cooper Chair at the University of Arkansas. Funds for the acquisition of the Red Diamond supercomputer were provided in part by the National Science Foundation under grant No. MRI-0421099.

**Supporting Information Available:** Predicted p$K_a$ values computed at different theoretical levels are tabulated. Tables S1–S4: OLYP and B3LYP gas-phase and COSMO results. Tables S5 and S6: predicted p$K_a$ values from OLYP and B3LYP COSMO single point calculations with gas-phase optimized geometries. Table S7: predicted p$K_a$ values from geometries optimized using the SYBYL force field of PQSMol. Table S8: p$K_a$ values calculated at the Hartree–Fock and PW91 levels. This material is available free of charge via the Internet at http://pubs.acs.org.

## References and Notes

(1) Hasselbalch, K. A. *Biochem. Z.* **1916**, *78*, 112.

(2) World Drug Index from Thomson Scientific (1999) (see http://scientific.thomson.com/products/wdi)

(3) Comer, J.; Tam, K. Lipophilicity profiles: Theory and measurement. In *Pharmacokinetic Optimization in Drug Reseach: Biological, Physicochemical and Computational Strategies*; Testa, B., van de Waterbeemd, H., Folkers, G., Guy, R., Eds.; Wiley-VCH: New York, 2000; pp 275–276.

(4) Topol, I. A.; Tawa, G. J.; Caldwell, R. A.; Eissenstat, M. A.; Burt, S. K. *J. Phys. Chem. A* **2000**, *104*, 9619.

(5) Magill, A. M.; Cavell, K. J.; Yates, B. F. *J. Am. Chem. Soc.* **2004**, *126*, 8717.

(6) Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M. E. *J. Phys. Chem. A* **2003**, *107*, 9380.

(7) Adam, K. R. *J. Phys. Chem. A* **2002**, *106*, 11963.

(8) Schmidt am Busch, M.; Knapp, E.-W. *ChemPhysChem* **2004**, *5*, 1513, and references therein.

(9) Klicić, J. J.; Friesner, R. A.; Liu, S.-Y.; Guida, W. C. *J. Phys. Chem. A* **2002**, *106*, 1327.

(10) Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799.

(11) *PQS*, version 3.3; Parallel Quantum Solutions: 2013 Green Acres Road, Suite A, Fayetteville, AR 72703, U.S.A.; see http://www.pqs-chem.com.

(12) Red Diamond was supported by the National Science Foundation under grant number 0421099.

(13) Kortüm, G.; Vogel, W.; Andrussow, K. *IUPAC: Dissociation constants of organic acids in aqueous solution*; Butterworths: London, 1961.

(14) Perrin, D. D. *IUPAC: Dissociation constants of organic bases in aqueous solution*; Butterworths: London, 1965.

(15) Perrin, D. D. *IUPAC: Dissociation constants of organic bases in aqueous solution: supplement*; Butterworths: London, 1972.

(16) Serjeant, E. P.; Dempsey, B. *IUPAC: Ionisation constants of organic acids in aqueous solution*; Pergamon: Oxford, U.K., 1979.

(17) Dean, J. A., Ed. *Lange's Handbook of Chemistry*, 15th ed.; McGraw-Hill: New York, 1999.

(18) *PCModel* v8.5; Serena Software: Bloomington, IN, 2003.

(19) Clark, M.; Cramer, R. D.; van Opdenbosch, N. *J. Comput. Chem.* **1989**, *10*, 982.

(20) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(21) Hoe, W.-M.; Cohen, A. J.; Handy, N. C. *Chem. Phys. Lett.* **2001**, *341*, 319.

(22) Binkley, J. S.; Pople, J. A.; Hehre, W. J. *J. Am. Chem. Soc.* **1980**, *102*, 939.

(23) The non-standard 3-21G(*) basis employs d polarization functions for first-row heteroatoms with lone pairs, principally N, O, and F. (Note that in the standard 3-21G(d) basis, there are *no* polarization functions on first-row elements.)

(24) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.

(25) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.

(26) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265.

(27) Baker, J.; Pulay, P. *J. Chem. Phys.* **2002**, *117*, 1441.

(28) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.

(29) Tran, N. L.; Colvin, M. E. *J. Mol. Struct. (THEOCHEM)* **2000**, *532*, 127.

(30) Baker, J.; Füsti-Molnár, L.; Pulay, P. *J. Phys. Chem. A.* **2004**, *108*, 3040, and references therein.

(31) Sadlej-Sosnowska, N. *Theor. Chem. Acc.* **2007**, *118*, 281.

(32) Liptak, M. D.; Gross, K. C.; Seybold, P. G.; Feldgus, S.; Shields, G. C. *J. Am. Chem. Soc.* **2002**, *124*, 6421.

(33) Murlowska, K.; Sadlej-Sosnowska, N. *J. Phys. Chem. A* **2005**, *109*, 5590.

(34) Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998**, *102*, 1995.

(35) Liptak, M. D.; Shields, G. C. *J. Am. Chem. Soc.* **2001**, *123*, 7314.

(36) da Silva, C. O.; da Silva, E. C.; Nascimento, M. A. C. *J. Phys. Chem. A* **1999**, *103*, 11194.

(37) Pliego, J. R.; Riveros, J. M. *J. Phys. Chem. A* **2002**, *106*, 7434.

(38) Klamt, A. *COSMO-RS: From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*; Elsevier: Amsterdam, 2005.

(39) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Clarendon: Oxford, U.K., 1990.

(40) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244.

(41) Chipman, D. M. *J. Phys. Chem.* **2002**, *106*, 7413.

(42) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(43) Rashin, A. A.; Young, L.; Topol, I. A. *Biophys. Chem.* **1994**, *51*, 359.

(44) *CMC-3D*, version 98.1; MDL Information Systems, Inc.: 14600 Catalina St., San Leandro, CA 94577.

(45) Jia, Z.; Ramstad, T.; Zhong, M. *Electrophoresis* **2001**, *22*, 1112.

(46) Connors, K. A.; Lipari, J. M. *J. Pharm. Sci.* **1976**, *65*, 379.

(47) Bell, J. L. S.; Wesolowski, D. J.; Palmer, D. A. *J. Solution Chem.* **1993**, *22*, 125.

(48) Flash, P. *J. Chem. Educ.* **1994**, *71*, A6.

(49) Sue, K.; Ouchi, F.; Minami, K.; Arai, K. *J. Chem. Eng. Data* **2004**, *49*, 1359.

(50) Cleveland, J. A.; Benko, M. H.; Gluck, S. J.; Walbroehl, Y. M. *J. Chromatogr. A* **1993**, *652*, 301.

(51) Bosch, E.; Bou, P.; Allemann, H.; Rosés, M. *Anal. Chem.* **1996**, *68*, 3651.

JP9067069