Journal of
CHEMICAL INFORMATION
AND MODELING

Rationalization of the pK_a Values of Alcohols and Thiols Using Atomic Charge Descriptors and Its Application to the Prediction of Amino Acid pK_a's

Ilke Ugur,[†,‡,§] Antoine Marion,[†,‡] Stéphane Parant,[†,‡] Jan H. Jensen,[∥] and Gerald Monard*,[†,‡]

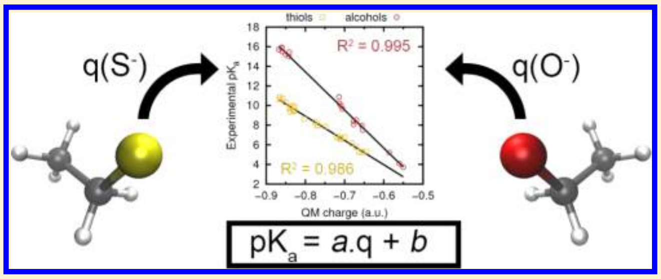
[†]Université de Lorraine, UMR 7565 SRSMC, Boulevard des Aiguillettes B.P. 70239, F-54506 Vandoeuvre-les-Nancy, France

[‡]CNRS, UMR 7565 SRSMC, Boulevard des Aiguillettes B.P. 70239, F-54506 Vandoeuvre-les-Nancy, France

[§]Department of Chemistry, Boğaziçi University, 34342 Bebek, Istanbul, Turkey

[∥]Department of Chemistry, University of Copenhagen, Copenhagen, Denmark

S *Supporting Information*



ABSTRACT: In a first step toward the development of an efficient and accurate protocol to estimate amino acids' pK_a's in proteins, we present in this work how to reproduce the pK_a's of alcohol and thiol based residues (namely tyrosine, serine, and cysteine) in aqueous solution from the knowledge of the experimental pK_a's of phenols, alcohols, and thiols. Our protocol is based on the linear relationship between computed atomic charges of the anionic form of the molecules (being either phenolates, alkoxides, or thiolates) and their respective experimental pK_a values. It is tested with different environment approaches (gas phase or continuum solvent-based approaches), with five distinct atomic charge models (Mulliken, Löwdin, NPA, Merz−Kollman, and CHelpG), and with nine different DFT functionals combined with 16 different basis sets. Moreover, the capability of semiempirical methods (AM1, RM1, PM3, and PM6) to also predict pK_a's of thiols, phenols, and alcohols is analyzed. From our benchmarks, the best combination to reproduce experimental pK_a's is to compute NPA atomic charge using the CPCM model at the B3LYP/3-21G and M062X/6-311G levels for alcohols (R^2 = 0.995) and thiols (R^2 = 0.986), respectively. The applicability of the suggested protocol is tested with tyrosine and cysteine amino acids, and precise pK_a predictions are obtained. The stability of the amino acid pK_a's with respect to geometrical changes is also tested by MM-MD and DFT-MD calculations. Considering its strong accuracy and its high computational efficiency, these pK_a prediction calculations using atomic charges indicate a promising method for predicting amino acids' pK_a in a protein environment.

INTRODUCTION

Among the 20 standard amino acids that compose proteins, some of them are ionizable. Arginine (Arg), aspartic acid (Asp), glutamic acid (Glu), cysteine (Cys), histidine (His), lysine (Lys), and to some extent tyrosine (Tyr), serine (Ser), and threonine (Thr) all have a side chain that can be protonated or unprotonated depending on acidic conditions. While the pK_a's of these individual amino acids are well-known in aqueous solutions,[1] their pK_a can be changed in a protein environment. This change can be due to different factors: desolvation, hydrogen bonding, or charge−charge interactions.[2−7]

The desolvation effect is due to the nature of proteins. Many are globular polymers, and some residues can be deeply buried inside their tertiary structure. In this case, buried amino acids are not in contact with solvent water anymore, and their pK_a is shifted to resemble more their respective pK_a in a nonpolar medium (i.e., they tend to reach a neutral form).[8] Many amino acids can be involved in hydrogen bonds either with their backbone or with their side chains. These hydrogen bonds between amino acids can also modify the amino acids' side chain pK_a's. For example, Li et al. have shown that if an aspartic acid side chain is involved in a hydrogen bond with a serine side chain, its pK_a can be lowered by one pK_a unit.[9] Finally, when two charged side chains are in close contact (a.k.a., a salt-bridge), this stabilizing interaction shifts their relative pK_a's to reflect this stable interaction: an acid becomes more acidic, and a base becomes more basic.[3]

Experimentally, amino acid pK_a's in proteins are difficult to measure. Usually, pK_a values are assigned from titration curves obtained using NMR spectroscopy.[10,11] In minor cases, other techniques like protein thermodynamics stability measurements can be used[8,12] or UV spectroscopy.[13] Overall, the reported number of measured pK_a values[14] is somewhat modest compared to the number of tertiary structures available at the PDB.[15]

Many theoretical approaches have been developed to estimate pK_a's or protonation states in proteins. These methods can be divided in four main classes: (i) QM based methods that usually estimate pK_a's from a thermodynamical cycle between

Received: February 9, 2014
Published: August 4, 2014

protonation/deprotonation of molecules in gas and aqueous phases;[16−21] (ii) MM based methods that use either alchemical modifications of ionizable residues to model (un)protonation or multiple protonation states;[22−26] (iii) continuum solvent based methods that use the dielectric difference between the interior of a protein and the outer solvent;[12,27−32] and (iv) knowledge based methods that use empirical parameters to estimate amino acids' pK_a's according to their respective local and global environments.[3−5] The accuracy measurement of a model is usually expressed as an error between any set of experimental pK_a's and their respective predicted pK_a's. One of the problems that modellers face is a possible systematic discrepancy between the experimental values and the modeled values. In most cases, pK_a experiments are performed in aqueous solution while models rely on 3D coordinates coming from the Protein Data Bank (PDB), which mostly contains structures obtained using X-ray or neutron diffractions. Therefore, the difference between a model pK_a and a set of experimentally measured pK_a's can be due not only to the model but also to various other origins: variability of the experimental conditions within the experimental set of pK_a's (i.e., temperature, ionic forces, presence of some ligands, etc.), X-ray resolutions of the proteins used in the model, packing effects in the crystallographic unit cell, sources (e.g., the organism for which a crystallographic structure exists can be different from the organism that was used for the experimental pK_a measurement), etc. For some ionizable residues, the number of reliable experimental data, today, can also be too small to provide meaningful statistics.

For knowledge based methods, having at least a trustworthy set of experimental pK_a's accompanied by their respective protein coordinates is crucial. Ideally, the experimental pK_a measurement should be made directly on the crystallographic structures to ensure a possible direct comparison between the experiments and the model.

Among the recent works on estimating the pK_a's of various organic compounds, the work by Zhang et al. is noteworthy.[33,34] They have designed a methodology to predict pK_a values in aqueous solution from the computation of deprotonation energies. For different acidic functional groups, they have performed a linear regression fit to model pK_a's as a function of the deprotonation energy ($E_{A-} - E_{AH}$). They have obtained coefficients for five different acidic functional groups in the form pK_a = $\alpha(E_{A-} - E_{AH}) + \beta$ with a very good correlation coefficient (R^2). They reported mean absolute deviations (MAD) of ~0.4 pK_a unit and a maximum error range of ±1.5 pK_a unit. Unfortunately, such an approach is difficult to transpose to proteins since this methodology requires one to compute the energies of both acidic (AH) and basic (A^-) states. In some proteins, steric interactions could lead to an incorrect, or nearly impossible, positioning of the proton in the AH form, especially when the basic form of an amino acid side chain is involved in a strong hydrogen bond or in a salt bridge.

To avoid any computation on the acidic form (AH), Roos et al. have suggested the use of atomic charge of the basic form (A^-) as a pK_a descriptor.[35] In the case of thiols, they have found a linear relationship between the NPA atomic charge on the sulfur atom of thiolates and the experimental pK_a of their related substituted thiols. This approach correlates the experimental pK_a for a compound, here thiols, with the tendency of its ionic form to bind a proton: the more negative the charge on a sulfur atom, the greater the proton attraction, hence the less acidity (= the higher pK_a). The set of experimental pK_a's was composed of (only) seven substituted thiols. Two of them had a pK_a around 3, while the other five had an experimental pK_a between 8 and 10. All thiolate forms were optimized at the B3LYP/6-31G* level using the Polarizable Continuum Model (PCM) to represent solvent effects, and charge calculations were done with B3LYP/6-31+G**. The linear regression fit was then applied to model the pK_a of cysteines in 3_{10}-helices and α-helices, resulting in a high correlation coefficient (R^2 = 0.980). Later, this work was successfully applied to the determiniation of the pK_a's of cysteines belonging to the thioredoxin superfamily.[36,37] Overall, this seminal work presented the potentiality of using atomic charges of a basic form to estimate pK_a as well as its possible application to problems of biological importance. From a set of acidic molecules (i.e, substituted thiols, alcohols, or phenols), atomic charges on the corresponding anionic forms are computed to yield, after a linear regression fit, to an estimation of pK_a as a function of the atomic charges. In peptides or proteins, this linear function can then be used to estimate the pK_a of related amino acid side chains (i.e., cysteine, serine/threonine, or tyrosine side chains, respectively) after the computation of the atomic charge of the corresponding anionic forms.

In the present work, we extend the approach of Roos et al.[35] to the case of substituted thiols and alcohols, including phenols. Here, we limit our study to the linear regression step and explore how experimental pK_a's can be related to the atomic charge of the basic forms RO^- (or RS^-) for alcohols and phenols (or thiols, respectively). Several questions are here addressed: Can an atomic charge accurately reproduce an experimental pK_a? Which atomic charge model best fits experimental data? What is the influence of the basis set or a DFT functional on the fitting? Does the solvent effect, using a continuum model approach, have to be included in the computation? What is the stability of the predicted pK_a with respect to the geometry of the molecules?

In the following sections, the methodology is presented; the linearity of the relationship between experimental pK_a and atomic charge is explored; the use of a continuum model to include averaged solvent effects is assessed; different computational protocols are compared; benchmarks using different DFT functionals and various basis sets are presented; and the stability of the present approach is analyzed. Finally, we present how accurately NDDO-based semiempirical Hamiltonians can be used instead of DFT functionals to predict pK_a's.

■ EXPERIMENTAL DATABASE

The calculations were performed on a total of 56 small organic molecules which consist of 14 phenols, 10 alcohols, and 32 thiols. Phenols and alcohols were treated in one subset (i.e., they all bear a hydroxyl group with the benzyl substituent in phenols having a strong effect on the resulting pK_a), while thiols were in another one. The subset containing phenols and alcohols will be simply referred to as alcohols from now on. Among all of the molecules studied, 19 alcohols and 25 thiols were used to construct the training set (see Tables 1 and 2). Five alcohols and seven thiols were added to the test set, which includes tyrosine and cysteine dipeptides (see Tables 3 and 4).

Our fundamental criterion to choose the molecules in the subsets is to be able to represent the widest range of experimental pK_a's. By merging phenols with alcohols, the

Table 1. Alcohols Training Set: IUPAC Nomenclature, Molecule Names, Experimental pK_a, Predicted pK_a, and Differences between Experimental and Predicted pK_a Values

alcohols	name	pK_a (exptl.)	pK_a (pred.)[a]	Δ pK_a
2,6-dinitrophenol	a01	3.71[38]	3.76	0.05
2,4-dinitrophenol	a02	4.09[38]	4.02	−0.07
2,5-dinitrophenol	a03	5.21[38]	4.96	−0.25
2,5-dichlorophenol	a04	7.51[39]	7.83	0.32
4-cyanophenol	a05	7.96[40]	7.93	−0.03
3-hydroxyquinoline	a06	8.06[41]	8.77	0.71
3-methylsulfonylphenol	a07	8.40[42]	8.68	0.28
5-hydroxyquinoline	a08	8.54[41]	8.44	−0.10
3-methoxyphenol	a09	9.65[43]	9.73	0.08
phenol	a10	9.97[38]	9.86	−0.11
4-*tert*-butylphenol	a11	10.23[44]	10.09	−0.14
2,4,6-trimethylphenol	a12	10.87[45]	10.16	−0.71
2-methoxyethanol	a13	15.00[46]	14.87	−0.13
methanol	a14	15.20[47]	15.50	0.30
phenyl-methanol	a15	15.44[48]	14.98	−0.46
ethanol	a16	15.50[49]	15.81	0.31
2-propanol	a17	15.70[47]	16.09	0.39
1-propanol	a18	15.87[48]	15.88	−0.04
1-butanol	a19	15.92[48]	15.48	−0.39

[a] pK_a values are computed from NPA atomic charges on optimized geometries of the anionic form using B3LYP/3-21G and CPCM (see text).

Table 2. Thiols Training Set: IUPAC Nomenclature, Molecule Names, Experimental pK_a, Predicted pK_a, and Differences between Experimental and Predicted pK_a Values

thiols	name	pK_a (exptl.)	pK_a (pred.)[a]	Δ pK_a
3-nitrobenzenethiol	t01	5.24[50]	5.49	0.25
5-mercaptouracil	t02	5.30[46]	5.65	0.35
4-acetylbenzenthiol	t03	5.33[50]	4.83	−0.50
3-chlorobenzenthiol	t04	5.78[50]	5.81	0.03
4-chlorobenzenethiol	t05	6.14[50]	6.10	−0.04
benzenethiol	t06	6.61[50]	6.65	0.04
2-methylbenzenethiol	t07	6.64[50]	6.60	−0.04
3-methylbenzenethiol	t08	6.66[46]	6.70	0.04
4-methoxybenzenethiol	t09	6.78[50]	6.38	−0.40
4-methylbenzenethiol	t10	6.82[50]	6.88	0.06
prop-1-ene-2-thiol	t12	7.86[46]	7.53	−0.33
ethyl-2-mercaptoacetate	t13	7.95[51]	8.15	0.20
2-mercaptopropan-2-one	t14	7.99[52]	8.13	0.14
N-2-mercaptopropanoyl-glycine	t15	8.33[53]	8.50	0.17
2,3-dimercapto-1-propanol	t16	8.62[46]	9.05	0.43
2-ethoxyethanethiol	t17	9.38[51]	9.86	0.48
phenylmethanethiol	t18	9.43[51]	9.66	0.23
3-mercaptopropane-1,2-diol	t19	9.51[51]	9.42	−0.09
2-mercaptoethanol	t20	9.72[51]	9.89	0.17
2-mercapto-2-methyl-1-propanol	t21	9.85[54]	9.48	−0.37
prop-2-ene-1-thiol	t22	9.96[51]	9.78	−0.18
methanethiol	t23	10.33[54]	10.17	−0.16
ethanethiol	t24	10.61[55]	10.55	−0.06
butane-1-thiol	t25	10.67[51]	10.40	−0.27
2-propanethiol	t26	10.86[56]	10.51	−0.35

[a] pK_a values are computed from NPA atomic charges on optimized geometries of the anionic form using M062*X*/6-311G and CPCM (See text).

Table 3. Alcohols Test Set: IUPAC Nomenclature, Molecule Names, Experimental pK_a, Predicted pK_a, and Differences between Experimental and Predicted pK_a Values

alcohols	name	pK_a (exptl.)	pK_a (pred.)[a]	Δ pK_a
2,3,4,5,6-pentachlorophenol	a20	5.62[57]	6.26	0.64
2,6-dichlorophenol	a21	6.79[46]	7.78	0.99
tyrosine	a22	9.84[1]	9.81	−0.03
prop-2-yn-1-ol	a23	13.60[49]	14.22	0.62
3-hydroxypropanenitrile	a24	14.03[48]	13.87	−0.16

[a] pK_a values are computed from NPA atomic charges on optimized geometries of the anionic form using B3LYP/3-21G and CPCM (see text).

Table 4. Thiols Test Set: IUPAC Nomenclature, Molecule Names, Experimental pK_a, Predicted pK_a, and Differences between Experimental and Predicted pK_a Values

thiols	name	pK_a (exptl.)	pK_a (pred.)[a]	Δ pK_a
2,2,2-trifluoroethanethiol	t27	7.30[58]	7.84	0.54
cysteine	t28	8.55[1]	8.93	0.38
2-mercaptoacetic acid	t29	10.40[46]	10.16	−0.24
2-aminoethanethiol	t30	10.53[46]	10.33	−0.20
2,3-dimercaptopropan-1-ol	t31	10.57[46]	10.52	−0.05
2-methylpropane-2-thiol	t32	11.05[56]	10.18	−0.87
2-methylbutane-2-thiol	t33	11.22[56]	10.19	−1.03

[a] pK_a values are computed from NPA atomic charges on optimized geometries of the anionic form using M062*X*/6-311G and CPCM (see text).

range in experimental pK_a goes from 3.71 to 15.92 units. For thiols, the experimental pK_a's vary from 5.24 to 11.22.

Most of the molecules in the subsets are not only small but also rather rigid molecules. The reason for avoiding flexible molecules is to overcome the risk of not obtaining the global minimum, which would raise systematical errors in pK_a predictions as already discussed by Zhang et al.[34] For the few molecules which are relatively more flexible than the rest, a conformational search was performed prior to the pK_a prediction calculations (the details are introduced in the Computational Details).

■ COMPUTATIONAL DETAILS

Quantum Mechanical Calculations. All of the Quantum Mechanical (QM) calculations were done using the Gaussian 09[59] program package. Nine different density functionals (BLYP,[60,61] B3LYP,[60,62] OLYP,[60,63] O3LYP,[60,63,64] PBE,[65] PBE0,[66] M06,[67,68] M06L,[68,69] and M062X[67,68]) and 16 different basis sets were used. Four different semiempirical approaches—AM1,[70] PM3,[71] PM6,[72] and RM1[73]—were also tested. To interpret the aqueous solvent environment, the universal solvent model (SMD[74]), the polarizable continuum model (PCM[75]), and the polarizable conductor solvent model (CPCM[76]) were used with the dielectric constant (ε) of 78.5. Five different types of atomic charge models were tested: Mulliken population analysis,[77] Löwdin population analysis,[78] Natural Population Analysis (NPA),[79] and two models within the Electrostatic Potential (ESP) framework (Merz−Kollman (MK) model[80] and Charges from Electrostatic Potentials using a Grid based method (CHelpG)[81]). Unless otherwise stated, all the charge calculations were performed on the optimized geometries (after including or not the solvent effect).

Molecular Dynamics Calculations. The amino acids tyrosine and cysteine anions were represented in their dipeptide forms: the N and C terminal ends of the residues (tyrosine or cysteine) are substituted with *N*-methylamide and acetamide, respectively. Parameters for phenol and phenolate were prepared using charges obtained from restrained electrostatic potential (RESP[82]). The RESP charges were calculated using the B3LYP/cc-pVTZ level of theory[83] on the geometries optimized at HF/6-31G**, in the spirit of the work by Duan et al.[84] Topology and coordinate files of these systems were prepared with the tleap module of AMBER 12 program package.[85] Dipeptides were treated with the ff03 force field.[84] The GAFF[86] force field of AMBER 12 was used to describe phenol and phenolate. The aqueous polar environment was mimicked by the implicit modified generalized Born model with α, β, and γ being 1.0, 0.8, and 4.85[59] as implemented in AMBER 12 (igb = 5). Following minimization, the systems were heated up to 300 K using the Andersen coupling algorithm[88] for 50 ps with velocities updated every 1000 steps, and a time step of 1 fs. Heated systems were equilibrated up to 100 ps. Finally, NVT production runs were performed for another 150 ps using the same thermostat algorithm. From the classically equilibrated structures, molecular dynamics simulations were also performed using DFT potentials (DFT-MD). These calculations were accomplished using the external interface between AMBER 12 and Gaussian 09 program packages.[89] B3LYP/3-21G was used for phenol and tyrosine; M062*X*/6-311G was used for cysteine. The choice of the DFTs will be detailed in the Results and Discussion section. DFT-MD simulations were performed using the protocol introduced for MM-MD calculations and productions were performed up to 50 ps.

RESULTS AND DISCUSSIONS

We investigated the linear relationship between atomic charges and experimental pK_a's for our two training sets of molecules representing various substituted thiols and alcohols. From the methodology suggested by Roos et al. (i.e., computing the atomic charge of the basic form RS^- and RO^- [35,90]) we found out that the choice of the computational framework plays an important role in the accuracy of the pK_a prediction. From the overall present study, we show that the combination of charge and solvent models giving the best results is NPA with CPCM for both thiols and alcohols sets. M062*X*/6-311G and B3LYP/3-21G are the DFT methods which are the most reliable for the present application on thiols and alcohols, respectively. In what follows, we first show the linear relationship between experimental pK_a's and atomic charges computed using the theoretical frameworks discussed above. Thus, using these results as a reference, we shall discuss the choice of charge model, DFT functional, basis-set and solvent model by changing one of these parameters while the others remain fixed to the best combination.

Linearity of the Relationship Between Experimental pK_a's and Atomic Charges. The anionic forms of all the molecules in the two training sets (RO^- and RS^-) were optimized in the aqueous phase (CPCM, ε = 78.5) using DFT (B3LYP/3-21G and M062*X*/6-311G for alcohols and thiols, respectively). With the same solvent model, natural population analyses were performed on these optimized geometries using the same DFT for the corresponding subsets (denoted $O^-//O^-$ and $S^-//S^-$ hereafter). The NPA charges of the atoms O^- and S^- were associated with the experimental pK_a's of each molecule in order to establish a standard linear fit. This method is referred to below as method 0 (M0). Figure 1 shows the relationship between experimental pK_a and computed NPA charge for the two training sets. A linear equation is obtained by a least-squares fit

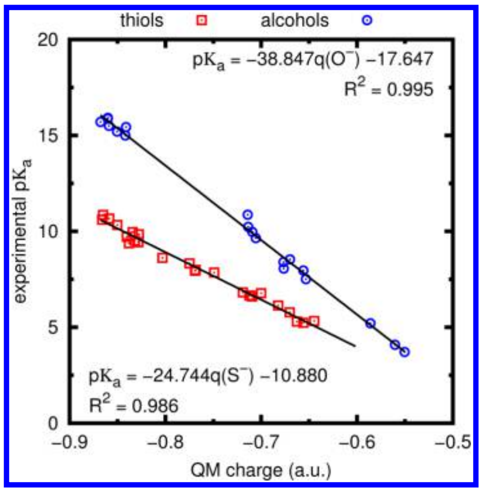


Figure 1. Linear regression between calculated NPA charges and experimental pK_a. Calculations were done using B3LYP/3-21G//CPCM and M062*X*/6-311G//CPCM for alcohols and thiols, respectively.

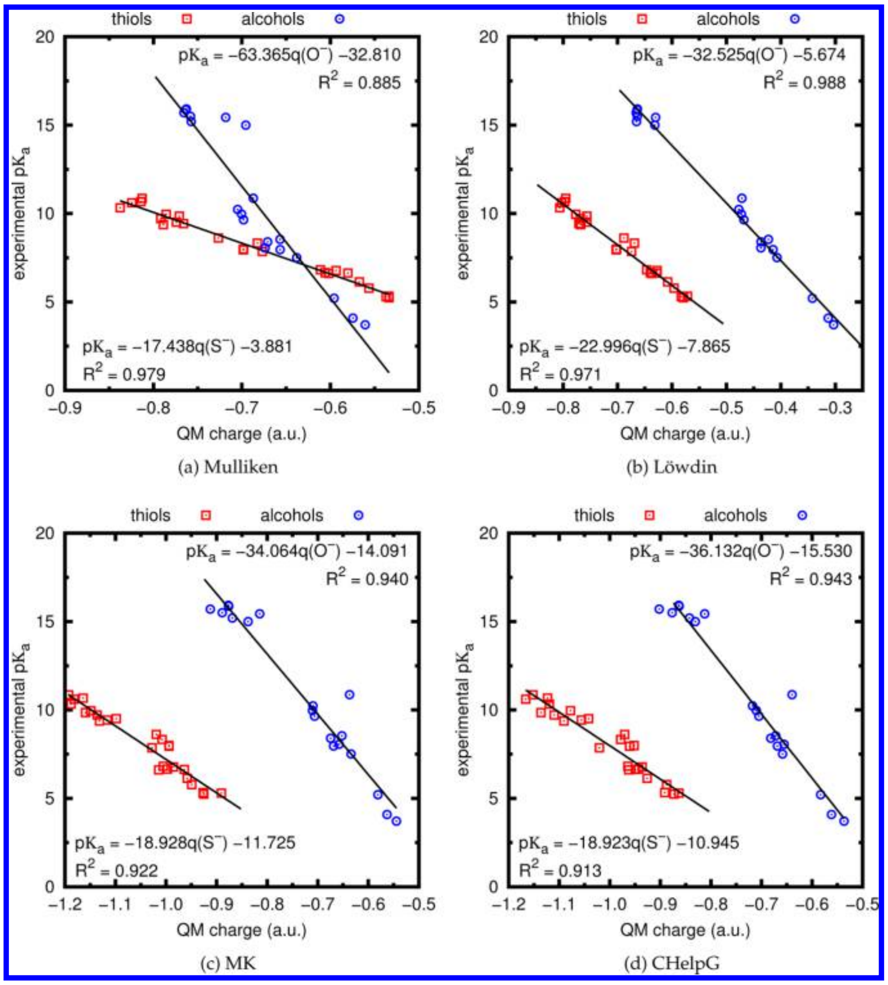
$$pK_a = a \cdot q(X^-) + b \quad (1)$$

where a and b are the fitted parameters and $q(X^-)$ is the calculated charge on $q(O^-)$ and $q(S^-)$ for alcohols and thiols, respectively. The parameters a and b and the squared Pearson correlation coefficient (R^2) are also illustrated in Figure 1. The predicted pK_a's are computed using eq 1 (i.e., by importing $q(X^-)$ of a given molecule into the parametrized equation).

For alcohols and thiols, the R^2 values were found to be 0.995 and 0.986, respectively. For both of the training sets, no strong outlier molecules were observed. The maximum difference between predicted and experimental pK_a among all the molecules was found as 0.71 and 0.50 units for alcohols and thiols, respectively (see Tables 1 and 2). These results indicate a strong correlation with the experimental pK_a's and the charges on O^- and S^-. In order to analyze the influence of the charge model on the quality of the fit, the same protocol was applied with four other charge models.

Influence of the Charge Model. Fixing the DFT methods and the solvent model for both sets of molecules to M062*X*/6-311G/CPCM for thiols and B3LYP/3-21G/CPCM for alcohols, alternative charge calculations were performed on the same optimized geometries that were used for NPA charge calculations. As was done in the case of NPA, the charges on O^- and S^- of the anions were extracted to establish a linear fit with the experimental pK_a's. The results for alternative charge models (Mulliken, Löwdin, MK, and CHelpG) are represented in Figure 2.

All of the tested charge models appear to have a strong correlation with the experimental pK_a, having $0.885 \leq R^2 \leq 0.988$. The smallest R^2 is found with Mulliken population analysis in the alcohols subset (0.885, Figure 2a). For this model, some molecules were found to be spread away from the fitting line, especially in the experimental pK_a range of 5 to 15. Using eq 1, the predicted pK_a of the strongest outliers were



(a) Mulliken (b) Löwdin (c) MK (d) CHelpG

Figure 2. Effect of the charge model on the linear regression between calculated atomic charges and experimental pK_a. Calculations were done with B3LYP/3-21G//CPCM and M062X/6-311G//CPCM for alcohols and thiols, respectively: (a) Mulliken atomic charge model; (b) Löwdin atomic charge model; (c) Merk−Kollman charge model; (d) CHelpG charge model.

found to be 3.78 and 2.73 units different than the experimental value (the molecules with the experimental values of 15.00 and 15.44, respectively). Again for the alcohols subset, better results were obtained using Löwdin charges with a considerably high value of R^2 (0.988). The accurate predictions of Löwdin charges with alcohols have already been discussed in a computational study which focuses on pK_a predictions of substituted phenols.[91] The computed atomic charges for alcohols are similar using the MK and CHelpG models, meaning that the quality of the prediction is mainly led by the ESP framework. Although both of the methods have high R^2 values (MK = 0.940 and CHelpG = 0.943), a noticeable amount of molecules deviates from the linear fit. The maximum deviation was observed for the molecule having experimental pK_a = 10.87 (a12:2,4,6-trimethylphenol), and the difference with the predicted value was found to be 3.18 and 3.38 for the methods MK and CHelpG, respectively.

In the case of thiols, Mulliken, Löwdin, and the two ESP methods (MK and CHelpG) all give quite high values of R^2: 0.979, 0.971, 0.913, and 0.922, respectively. Interestingly, considering both the R^2 values and the spreading around the linear fit function, Mulliken and Löwdin methods give more adequate results compared to ESP methods.

To summarize, all of the tested atomic charge models generate acceptable linear fit with experimental pK_a. Nevertheless, the NPA method appears to be the best when Figures 1 and 2 are compared.

Solvent Models. Zhang et al.[34] proposed CPCM to represent the solvent environment in their work focusing on pK_a prediction from deprotonation energies of a set of molecules. Roos et al. used PCM for pK_a prediction calculations using QM charges. In addition to our CPCM calculations which were introduced up to this point, here the accuracy of PCM and SMD models was tested. Gas phase calculations were also performed considering its smaller computational cost compared to either of the solvent methods. Figure 3 presents the results of the linear regression of gas phase, SMD, and PCM calculations using NPA charges and the DFT methods discussed in the previous sections.

Either gas phase, SMD, or PCM calculations are as accurate as CPCM calculations with $R^2 \geq 0.958$ (Figures 1 and 3). The results are fairly accurate in the case of alcohols (R^2 = 0.993) in the gas phase. Interestingly, in the case of alcohols, PCM calculations yield less accurate results (R^2 = 0.982) compared to gas phase calculations. The lowest R^2 was observed for the gas phase calculations of thiols (0.958). Going from gas phase to SMD and PCM solvent models, the accuracy for thiols is increased (R^2 = 0.979 and 0.986 for SMD and PCM, respectively). This trend was also observed by Roos et al.; the R^2 of seven thiols were found as 0.917 and 0.980 for gas and

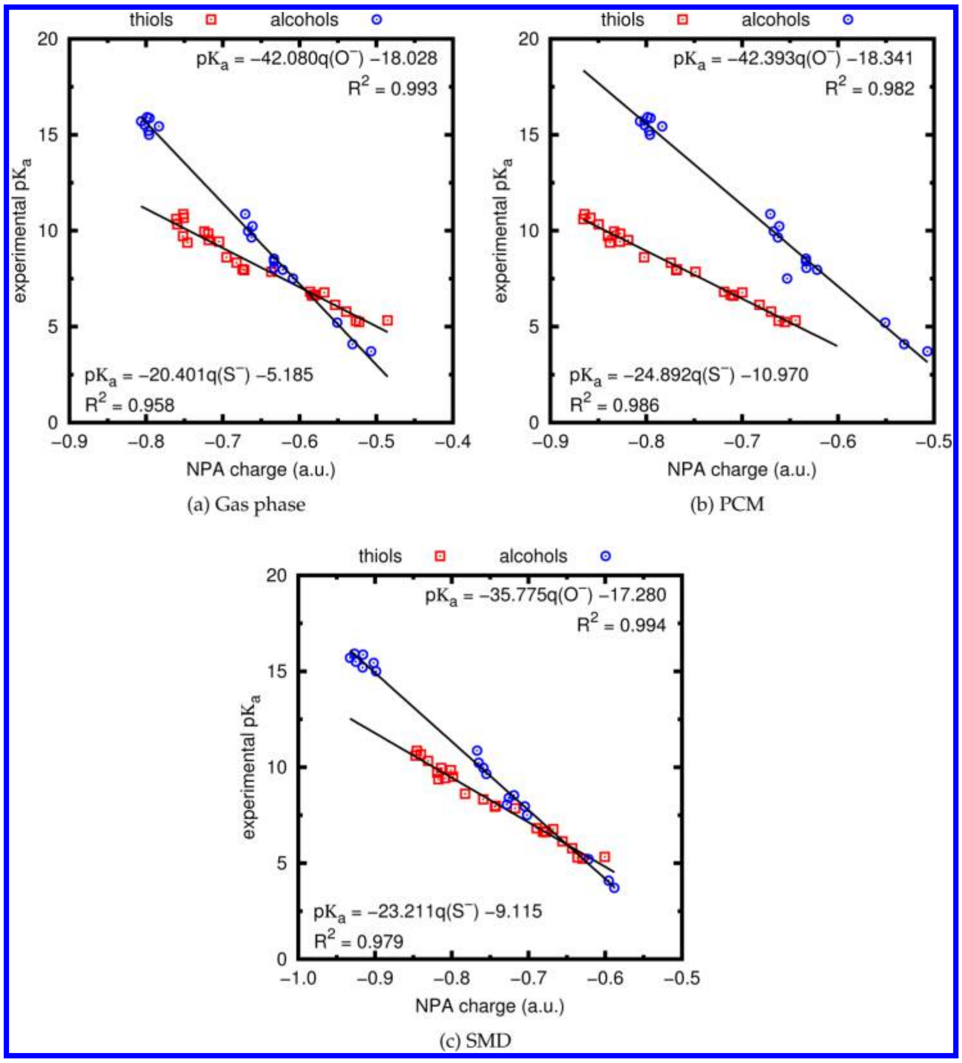


Figure 3. Effect of the implicit solvent model on the linear regression between NPA charges and experimental pK_a's. Calculations were done with B3LYP/3-21G and M062*X*/6-311G for alcohols and thiols, respectively: (a) gas phase, (b) PCM model, (c) SMD model.

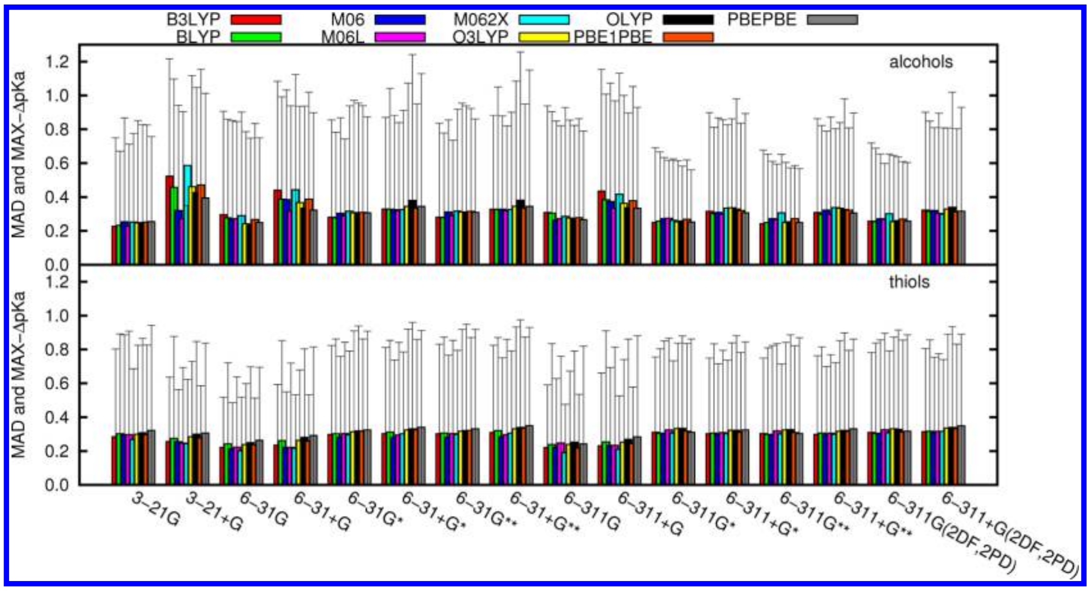


Figure 4. Mean Absolute Deviation (MAD) and maximum difference between predicted and experimental pK_a (MAX-ΔpK_a) for nine different DFT functionals and 16 different basis sets considered in this work. Geometry optimizations and NPA charge calculations were done using the CPCM model.

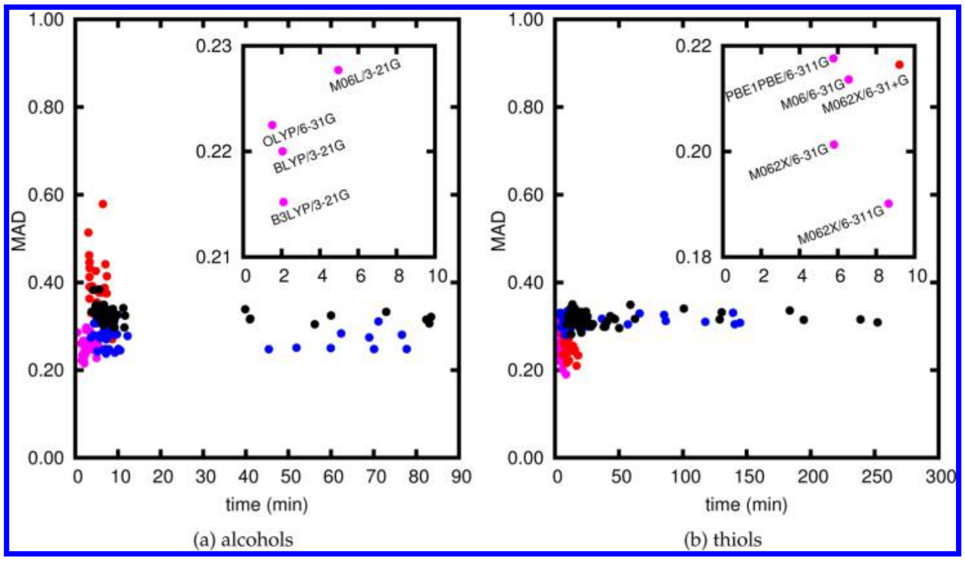


Figure 5. Mean Absolute Deviation (MAD) for all of the DFT functionals and basis sets (solvation model = CPCM, charge model = NPA) as a function of the CPU time. Red, blue, black, and magenta colored points respectively represent diffuse, polarization, both diffuse and polarization, or none of them. The region of the smallest MAD and time values are shown in expanded versions.

PCM calculations (B3LYP/6-31+G**//B3LYP/6-31G*). In the present study, relatively higher accuracy is found (R^2 = 0.986) for a larger set of molecules which have a wider range of experimental pK_a's. Comparing all of these four methods (gas phase, SMD, PCM, and CPCM), the best linearity is obtained when the CPCM solvent model is used.

DFT Functional and Basis-set Benchmarks. For a bench of DFT functionals and basis sets, the same protocol was applied to our two training sets (optimization and NPA charge calculations on RO^- and RS^-, using the CPCM solvent model). R^2, *a*, and *b* values were obtained from the linear fit with experimental pK_a's. In Figure 4, the Mean Absolute Deviation (MAD) of all the considered DFT functionals as a function of the selected basis set are presented (box representations). For each combination of DFT functional and basis set, the difference between the experimental and predicted pK_a was calculated (ΔpK_a). The maximum value of this difference (MAX-ΔpK_a) is also shown in Figure 4 (black colored lines). The highest three ΔpK_a values and also the R^2's are given in the Supporting Information for each DFT method.

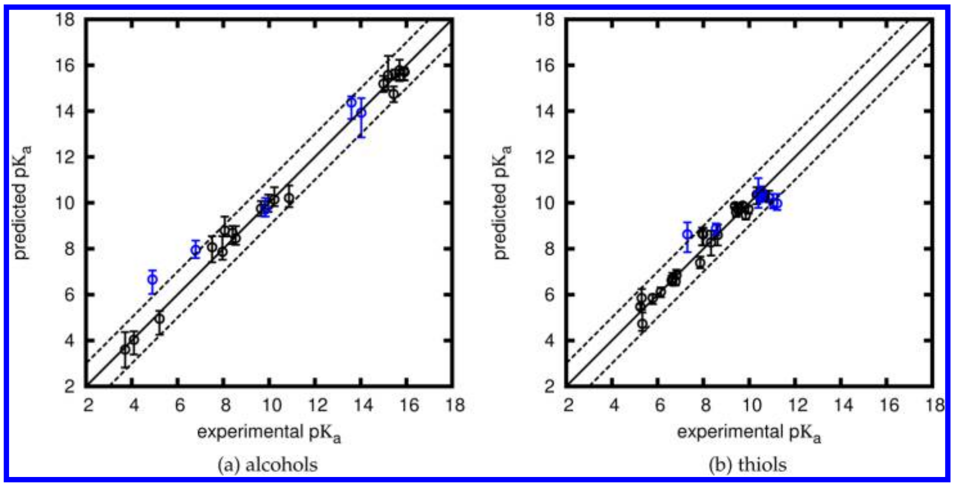
For both of the subsets, all of the DFT methods result in a strong correlation between atomic charges and experimental pK_a's with high R^2 and small MAD (or MAX-ΔpK_a) values. The R^2 was found to be $0.974 \leq R^2 \leq 0.995$ and $0.941 \leq R^2 \leq 0.986$ for alcohols and thiols, respectively (see Supporting Information). In the case of alcohols, MAD and ΔpK_a values are in a range indicating high accuracy ($0.227 \leq$ MAD ≤ 0.586, $0.560 \leq$ MAX-ΔpK_a ≤ 1.217, Figure 4). These descriptors are even smaller for thiols ($0.190 \leq$ MAD ≤ 0.350, $0.476 \leq$ MAX-ΔpK_a ≤ 0.975, Figure 4). For thiols, the quality of the results is very stable and is almost the same for all of the considered methods. Either for alcohols or thiols, increasing the basis set or adding difuse and/or polarization functions did not cause a significant improvement in accuracy. Moreover, in the case of alcohols, the accuracy is diminished when diffuse functions were added to the basis set, for any of the DFT functionals (e.g., 6-311+G* has higher MAD and MAX-ΔpK_a compared to 6-311G*).

For alcohols, the smallest MADs were found for the combination of all functionals with the 3-21G basis set, and among all the tested methods B3LYP/3-21G gave the most accurate result, having MAD = 0.227. The MAX-ΔpK_a value for this subset is well predicted by 6-311G* and 6-311G** basis sets (with the smallest MAX-ΔpK_a of 0.560 with PBE/6-311G*). The highest R^2 (0.995) is equivalent for B3LYP/3-21G, M06*L*/3-21G, and BLYP/6-311G**. To conclude, for alcohols, B3LYP/3-21G appears to be the most accurate method having the smallest MAD, highest R^2, and an acceptable value of MAX-ΔpK_a (0.749).

In the case of thiols, the smallest MAD and MAX-ΔpK_a and also the largest R^2 were obtained with M062*X*/6-311G (0.190, 0.476, and 0.986, respectively).

The methodology we propose in this paper aims at predicting accurate pK_a, not only for small organic molecules but as a final goal, for amino acids in proteins. Thus, both the accuracy of the prediction and the computational cost should be considered. Two representative molecules (1-butanol and 4-chlorobenzenethiol), which contain a number of atoms representing the average of the corresponding subset, have been chosen, and the MADs of all methods as a function of computational time have been plotted for these two molecules (Figure 5). The basis sets were divided into four classes depending on bearing a diffuse or polarization function, both or none of them. As discussed above, for both of the training sets, adding diffuse and/or polarization (red, blue, black points) functions did not improve the accuracy. For alcohols, basis sets containing only diffuse functions (red points) give systematically the worst results. The most accurate combination of DFT functionals and basis sets for alcohols and thiols (B3LYP/3-21G and M062*X*/6-311G, respectively) was also found to be one of the least time-consuming methods. This result also indicates that the two level calculation suggested by Roos et al. (B3LYP/6-31G*//B3LYP/6-31+G**) is not a necessity in order to balance between the computational cost and the accuracy.

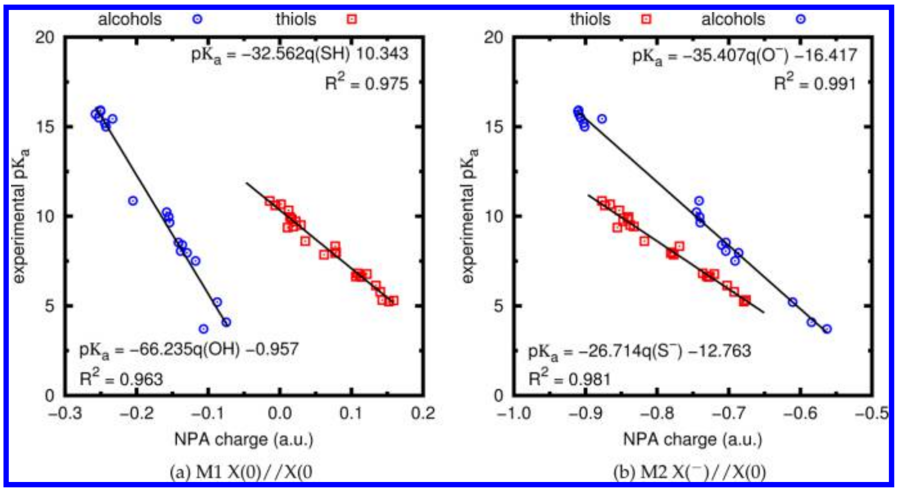
In order to have an overview on the efficiency of the DFT functionals and basis sets, the average predicted pK_a over all the methods was calculated. The minimum and maximum predicted pK_a's among all the methods have been added to the average predicted pK_a of each molecule as error bars. The predicted pK_a is plotted versus experimental values (Figure 6,



(a) alcohols

(b) thiols

Figure 6. Predicted pK_a over all the DFT functionals and basis sets versus experimental pK_a (solvation model = CPCM, charge model = NPA). Circles show the average pK_a, and the error bars denote minimum and maximum predicted pK_a. Black and blue lines correspond to the training and test sets, respectively.



(a) M1 X(0)//X(0

(b) M2 X(−)//X(0)

Figure 7. Effect of geometries on pK_a predictions. X = O or S. Method/basis set = B3LYP/3-21G and M062X/6-311G for alcohols and thiols, respectively. Solvent model = CPCM; charge model = NPA.

black colored lines). For both alcohols and thiols, average, minimum, and maximum values of the predicted pK_a were found to be within the range of ±1 unit compared to the experimental value.

The validity of the obtained protocol was tested using two small test sets which consist of five alcohols and seven thiols. These test sets also contain tyrosine and cysteine dipeptides. Average, minimum, and maximum values of the predicted pK_a's were also calculated for these test sets (Figure 6, blue colored lines).

Accurate predictions were obtained for most of the molecules in the test sets. Moreover, the average pK_a's of tyrosine and cysteine differ only by −0.13 and +0.33 units from their respective experimental pK_a's. The maximum pK_a's were found to be 0.37 and 0.45 units more than the experimental pK_a's of tyrosine and cysteine, respectively. The minimum of the calculated pK_a over all the DFT functionals and basis sets is 0.45 unit smaller for tyrosine and 0.04 units higher for cysteine compared to experimental data. For the best combination of DFT functionals and basis sets, the calculated pK_a's were found to be −0.05 and +0.38 units different than the experimental ones (B3LYP/3-21G and M062X/6-311G for tyrosine and cysteine, respectively). Thus, it can be concluded that the suggested protocol gives precise results for the predictions of amino acid pK_a's.

Few molecules in the test set have a deviation higher than 1 pK_a unit. For example, in the case of alcohols, pentachlorophenol and 2,6-dichlorophenol (with the experimental pK_a of 5.62 and 6.79) were found to be out of the ±1 range limit. However, for the corresponding molecules, the average value of the predicted pK_a was found to be very close to the minimum and maximum predicted values (error bars). This fact shows that all of the DFT methods result in similar pK_a predictions. Thus, for these two molecules, we have reasonable doubt that the experimental pK_a could be inaccurate. For thiols, a similar shift was observed for 2,2,2-trifluoroethanethiol (experimental pK_a equals to 7.30). To the opposite of the outliers in the alcohols subset, the error bars appear to be large, indicating that the pK_a predictions of this molecule are DFT method dependent. As was discussed in the Experimental Database section, this protocol would not be sufficient enough to calculate the pK_a of molecules bearing strongly electronegative atoms.

Geometries and Stability. In order to transfer the obtained pK_a prediction protocol to protein calculations, different features of proteins in terms of ionizability should

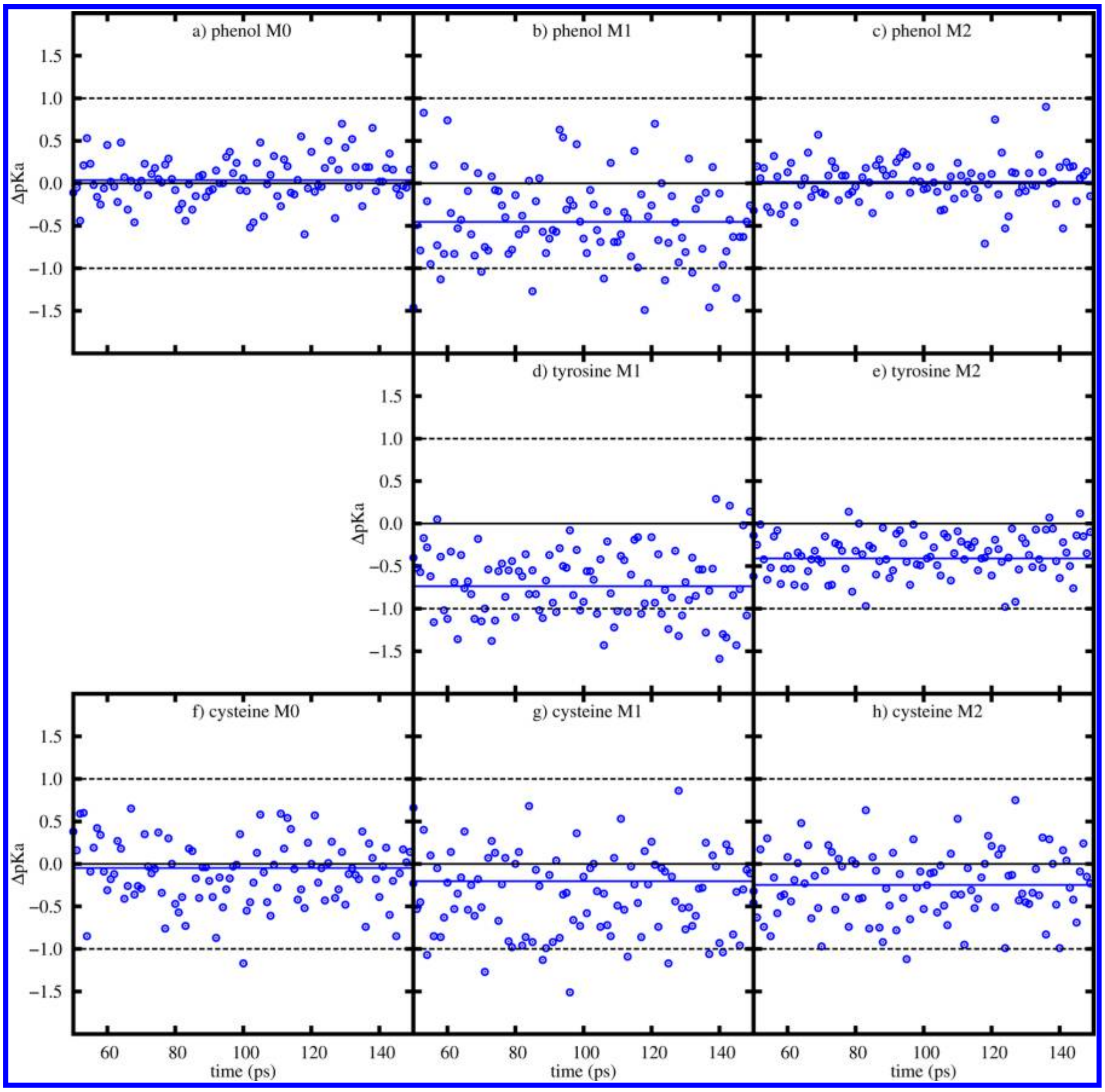


Figure 8. Deviations of predicted pK_a with respect to geometrical changes. Geometries are obtained from aqueous phase MM-MD calculations (modified generalized Born model). B3LYP/3-21G and M062X/6-311G methods for alcohols (phenol and tyrosine) and cysteine were used for single point NPA charge calculations with QM. QM calculations were performed using CPCM. The blue line shows the numerical average of the pK_a deviations.

be considered. In the crystal structure of proteins, depending on the environment, ionizable amino acids can be found in either their protonated or deprotonated forms. All of the calculations introduced up to here were performed considering only the anionic forms of the molecules. Here, a similar protocol is tested for the molecules optimized in their neutral form. All of the calculations were performed with CPCM solvent, NPA charges, and the DFT methods introduced above (B3LYP/3-21G and M062X/6-311G for tyrosine and cysteine, respectively). The charges are calculated in two ways:

- Method 1 (M1): For each optimized neutral alcohol (or thiols), the extracted atomic charge corresponds to the sum of the charge on the oxygen (or sulfur) and its bonded hydrogen. This method is referred to below as X(0)//X(0) (X = O or S)
- Method 2 (M2): The atomic charges that enter the fitting procedure are defined as the charge of the oxygen (or sulfur) atom on its anionic form. This method is referred to below as X(−)//X(0) (X = O or S), meaning that the atomic charges are extracted from an anion form after optimizing the corresponding neutral form.

In the case of alcohols, M1 gives a fit with R^2 equal to 0.963 (Figure 7a). For the same method, R^2 was obtained as 0.975 for thiols (Figure 7b). When the fit was performed on the charges obtained from single point calculations on the anions (M2), the alcohols and thiols generated linear fits with R^2 values of 0.991 and 0.981, respectively. As a comparison, for both alcohols and thiols, the best fit was obtained by the protocol discussed in previous sections (M0) with R^2 values of 0.995 and 0.986, respectively. However, both M1 and M2 also give a fairly good fit with the R^2 values higher than 0.98.

Another crucial point on pK_a predictions of proteins is the stability of the calculated pK_a's with respect to geometrical changes. For example, protein structures extracted from the PDB contain geometries of amino acids that originate from a refinement procedure. Internal distances, angles, dihedrals, etc. do not necessarily match those of the DFT minima as obtained when using the M0, M1, or M2 protocols. Moreover, applying

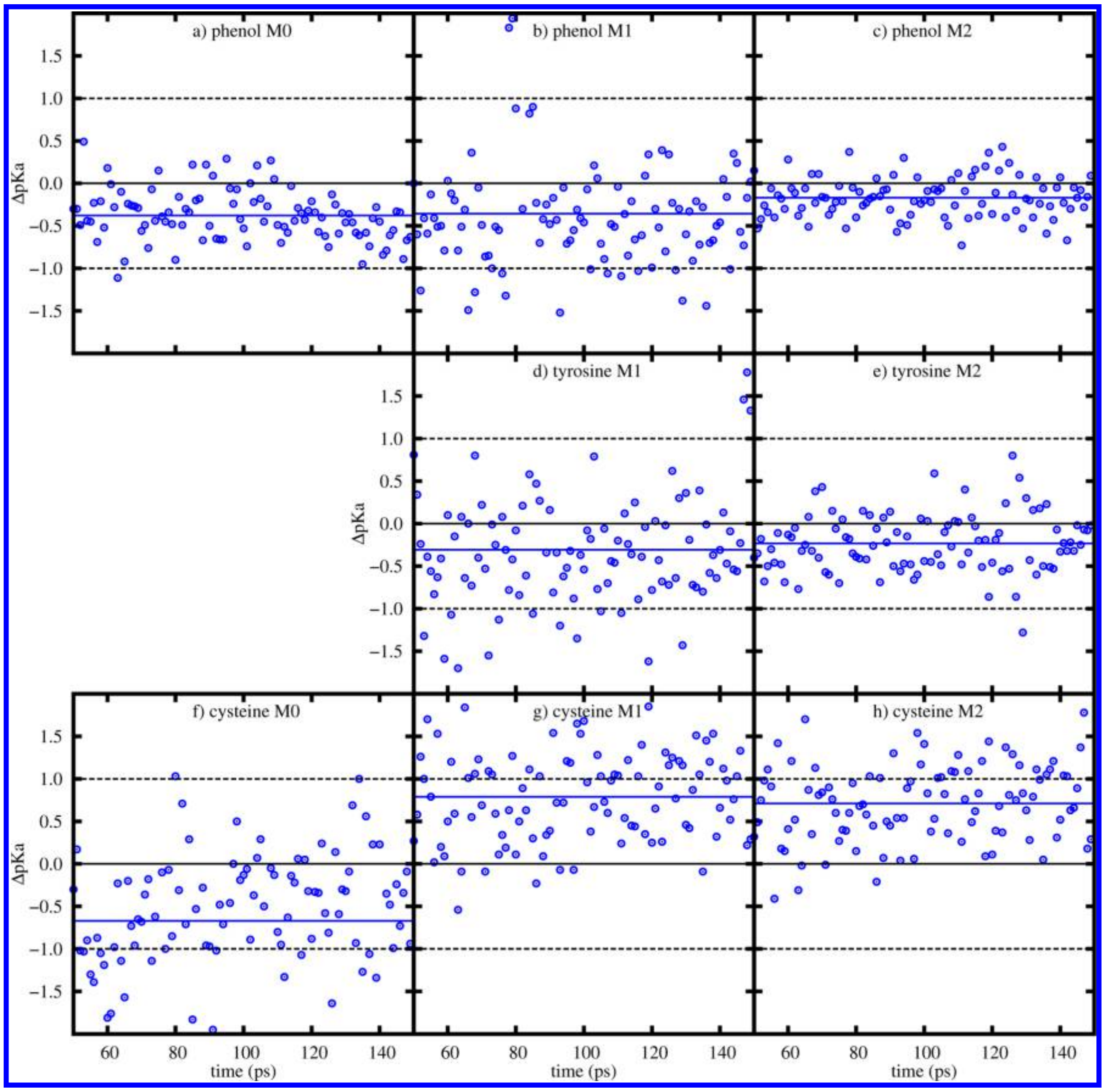


Figure 9. Deviations of predicted pK_a with respect to geometrical changes. Geometries are obtained from gas phase DFT-MD calculations. B3LYP/3-21G and M062*X*/6-311G methods for alcohols (phenol and tyrosine) and cysteine were used for both DFT-MD and single point NPA charge calculations with QM. QM calculations were performed using CPCM. The blue line shows the numerical average of the pK_a deviations.

one of these protocols to an entire protein is today computationally out of reach due to the cost of DFT calculations for large systems. Therefore, our present approach must be validated in the case where the single point calculations performed to extract NPA charges are applied on structures that are not energy minima (but still close to them). As a way to assess the variability of pK_a prediction with respect to geometrical changes, we have performed molecular dynamics (MD) simulations to provide multiple geometries around the optimum structures (i.e., the energy minima). Short MM-MD simulations (150 ps) were performed for phenol, tyrosine, and cysteine molecules using a continuum solvent model. The convergence of these simulations (i.e., do they reach all the available conformations?) is not here discussed since these short MDs are only designed to provide geometry samples. The NPA charges of the X atom (which corresponds to O and S for alcohols and cysteine, respectively) were calculated in three ways in order to represent the methods introduced above:

1. M0 ($X^{(-)}//X^{(-)}$): Deprotonated (anionic) forms of the molecules were simulated with MM-MD ($X^{(-)}$). Single point NPA charge calculations were done on these anionic forms ($X^{(-)}$).

2. M1 ($X(0)//X(0)$): MM-MD simulations on protonated (neutral) forms ($X(0)$) were followed by single point charge calculations on the same form ($X(0)$).

3. M2 ($X^{(-)}//X(0)$): Hydrogen atoms were removed from the neutral geometries obtained from MM-MD calculations ($X(0)$). The single point charge calculations were done on these anionic forms ($X^{(-)}$).

These protocols were performed on 50 frames extracted from each simulation (between 100 and 150 ps) for all methods and molecules. The predicted pK_a's were found using *a* and *b* values obtained from the fit with calculated charges and experimental pK_a's for the corresponding method (Figures 1 and 7). All of the charge calculations were performed using CPCM with B3LYP/3-21G and M062*X*/6-311G methods for alcohols (phenol, phenolate, and tyrosine) and thiols (cysteine and cysteine anion), respectively. The fluctuations of the calculated pK_a with respect to geometrical changes were monitored by taking the experimental pK_a as a reference

(Figure 8). The numerical average over all the frames was also calculated (blue line).

In the case of phenolate (M0), the predicted pK_a's deviate within the range of −1 to +1 unit with respect to experimental value (Figure 8 a). The average of these points yields a pK_a prediction which is very close to the experimental pK_a (0.04 unit). For the neutral form of this molecule (phenol), when the charges of both oxygen and hydrogen are considered (M1) the deviations were observed in a greater range (from −1.5 unit to 2 units, Figure 8 b). The predicted pK_a is 0.45 unit smaller than the experimental one. The best predictions were obtained with M2; the deviations were observed within ±0.5 unit, and the average pK_a was found to be 0.02 lower than the experimental one (Figure 8 c). Overall, in the case of phenol, M0 and M2 can be considered as stable: the pK_a prediction does not vary significantly with small geometrical deviations.

In the case of tyrosine, M0 was not studied since there are no available parameters for the tyrosine anion within the AMBER predefined library of the parameters. For the other two methods, M1 results in a wider spread of the calculated pK_a compared to M2. This is similar to the conclusions obtained in the case of phenol (Figure 8d and e). The relatively wider fluctuations in the case of tyrosine indicate the effect of the peptide backbone conformations on charge calculations. For M1 and M2, the average pK_a's are respectively 0.74 units and 0.41 units smaller than the experimental ones. This could be considered as quite precise for the latter method.

In the case of cysteine, all three of the methods give relatively wider results compared to the two other molecules tested (Figure 8f−h). Despite the large fluctuations, the average values are fairly close to the reference, which are −0.05, −0.20, and 0.07 units different from the experimental pK_a for M0, M1, and M2, respectively.

Combining the results for these three molecules, M0 and M2 give highly accurate and stable results. Hence, no matter if the simulation is performed in neutral or anionic form, the single point charge calculations should be performed on the anionic form of the molecule.

Similar calculations were also performed on structures extracted from DFT based MD. QM energies and forces were computing using M062X/6-311G for cysteine and B3LYP/3-21G for alcohols (phenol or tyrosine). In contrast to the MM-MD calculations, DFT-MD simulations were performed in the gas phase. All three methods (M0−2) were tested, and the results are shown in Figure 9.

For phenol, the pK_a deviates between −1 and 0.5 units in M0. A wider range of deviations was observed for M1 (−1.5 to 2 units). Among all three methods, M2 deviates least compared to the experimental pK_a (−0.5 to 0.5). The average calculated pK_a's were found to be 0.38, 0.35, and 0.17 units less than the experimental pK_a for M0, M1, and M2, respectively. Tyrosine behaves in a similar manner resulting in 0.31 (M1) and 0.23 (M2) units smaller predicted pK_a's than the experimental ones (Figure 9d and e). For cysteine, all three of the methods deviate from the experimental reference within a considerably wider range (Figure 9f−h). Correspondingly, the average calculated values were found to be imprecise: 0.65 unit less than the experimental pK_a for M0 and 0.79/0.71 higher than for M1/M2.

To conclude, for all three of these molecules, the narrowest range of deviations of the predicted pK_a's was obtained using M2 on the geometries extracted from classical force field calculations (MM-MD). The average pK_a of these structures also yields the closest prediction to the experimental value (−0.01, −0.41, and 0.07 unit deviation for phenol, tyrosine, and cysteine). Calculating geometries with DFT force fields did not improve the precision in most of the cases. In fact, the accuracy is lost in the case of cysteine. At this point, it should be noted that DFT-MD calculations were performed in the gas phase. The deviations of charges could be caused by the difference between gas and aqueous phase geometries. For either MM-MD or DFT-MD geometries, M1 yields the least accurate predictions for all three of the molecules.

Semiempirical Hamiltonians. The correlation between the experimental pK_a's and the atomic charges calculated by semiempirical approaches has also been analyzed. AM1, PM3, PM6, and RM1 were chosen for structure optimizations and Mulliken charge calculations. For alcohols, R^2's were found as 0.976, 0.965, 0.988, and 0.979 (for AM1, PM3, PM6, and RM1, respectively). In the case of thiols, the correlation coefficients were evaluated as 0.958, 0.962, 0.691, and 0.973 (AM1, PM3, PM6, and RM1, respectively). Disregarding the results obtained by PM6 for thiols (0.691), all of the tested semiempirical methods generate accurate results. The correlation between the experimental pK_a's and the Mulliken charges are shown in Figure 10 for the methods giving the highest R^2 value (PM6 and RM1 for alcohols and thiols, respectively). Both for alcohols and for thiols, there are not strong outliers (Figure 10). Hence, it can be concluded that these two methods provide a reasonable compromise between computational efficiency and accuracy. However, it is interesting to notice that while PM6 is the only Hamiltonian (among those considered here) which uses "d" orbitals for sulfur atoms, this method appears to be the less accurate in the case of thiols.

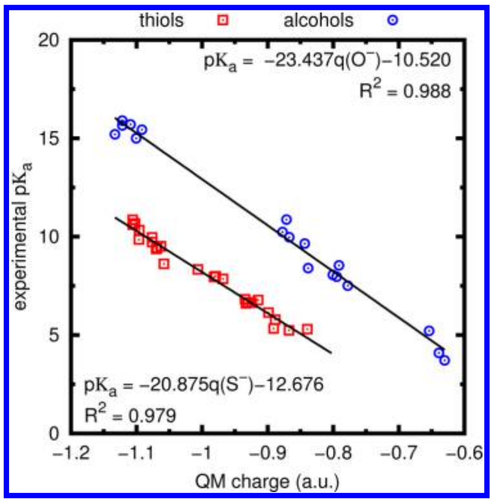


Figure 10. Experimental pK_a vs calculated Mulliken charges. Calculations were done with PM6//CPCM and RM1//CPCM for alcohols and thiols, respectively.

The stability of the calculated pK_a's with respect to geometrical changes was also tested using phenol, tyrosine, and cysteine models. The three methods (M0−2) introduced above were applied. The charge calculations were performed only on the geometries obtained from MM-MD simulations, using PM6 and RM1 for alcohols and thiols, respectively. The results are given in the Supporting Information. For all of the methods and for all of the samples, the deviations were found to be within ±2 units with respect to the experimental pK_a. However, none of the methods were found to be more accurate or more stable than what was obtained by DFT calculations (Figure 8).

■ CONCLUSIONS

In the present study, we have suggested a protocol in order to obtain accurate and fast pK_a predictions for small alcohols and thiols. In addition, we have tested their applicability to predict amino acid pK_a's. The suggested protocol is based on the linear regression of the experimental pK_a's with the atomic charges on the ionizable groups. Five charge models, three solvent models, gas phase calculations, several DFT methods (combination of nine DFT-functionals and 16 basis sets), and four semi-empirical Hamiltonians were tested. All of these methods resulted in a strong linearity having R^2 higher than 0.85 (and more than 0.95 for most of the cases). Among those, NPA charge calculations performed with the CPCM solvation model following the optimizations in CPCM give the most accurate results. The best combination of DFT functionals and basis sets are found to be B3LYP/3-21G and M062*X*/6-311G for alcohols and thiols, respectively. The energy optimization and NPA charge calculation procedures using these two most accurate DFT methods (with CPCM) came up with an acceptable computational expense.

Among all the ionizable amino acids, tyrosine and cysteine were chosen to be modeled for pK_a predictions. Using the best DFT combination, the difference between experimental and predicted pK_a was found to be −0.05 and +0.38 units for tyrosine and cysteine, respectively. The average predicted pK_a's over all the DFT methods are −0.45 (tyrosine) and +0.04 (cysteine) units different than the experimental value. MM-MD calculations showed that the predicted pK_a's deviate within ±1 unit with respect to different geometrical changes. The most accurate predictions were obtained when the charge calculations were performed on the anionic species. The average predicted pK_a over these different geometrical changes yielded precise predictions (−0.41 and 0.07 for tyrosine and cysteine, respectively). Combining these results, we have designed a protocol that accurately and efficiently can predict pK_a's of tyrosine and cysteine in solution. The next step will involve the study of the transferability of this protocol to predict pK_a's of serine, tyrosine, and cysteine within proteins and its extensions to predict pK_a's of other ionizable residues. It is hoped that the electronic mechanisms that affect the pK_a's of alcohols and thiols, when substituents are modified, will be similar to those that affect amino acids' pK_a's in proteins when their environment is modified (i.e., due to desolvation, hydrogen bondings, or charge−charge interactions).[36,37] Using NPA atomic charges extracted from B3LYP/3-21G-(CPCM) or M062*X*/6-311G(CPCM) calculations on subsets of proteins, it should be possible to evaluate the pK_a's of serine, tyrosine, and cysteine by reporting the atomic charge of the alkoxide or thiolate form, respectively, into the linear relationships reported in this work.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

R^2, MAD, and MAX ΔpK_a results for the alcohol and thiol training sets depending on the DFT functionals (B3LYP, BLYP, M06, M06L, M062X, O3LYP, OLYP, PBE0, or PBE) and the basis sets (3-21G, 3-21+G, 6-31G, 6-31+G, 6-31G*, 6-31+G*, 6-31G**, 6-31+G**, 6-311G, 6-311+G, 6-311G*, 6-311+G*, 6-311G**, 6-311+G**, 6-311G(2df,2pd), 6-311+G-(2df,2pd)). Semiempirical predictions along MM-MD simulations for phenol, tyrosine, and cysteine. Optimized geometries of the alcohol and the thiol sets. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +33 (0)383.684.381. Fax: +33 (0)383.684.371. E-mail: Gerald.Monard@univ-lorraine.fr.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

I.U. acknowledges a joint Ph.D. grant from TUBITAK Joint Ph.D. Program (2214-B). G.M. thanks Université de Lorraine and the French Embassy in Denmark for travel grants (mobility programs between Université de Lorraine and University of Copenhangen).

■ REFERENCES

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

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

(3) Li, H.; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and rationalization of protein pKa values. *Proteins* **2005**, *61*, 704−721.

(4) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins* **2008**, *73*, 765−783.

(5) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p*Ka*Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525−537.

(6) Thurlkill, 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.

(7) Laurents, D. V.; Huyghues-Despointes, B. M. P.; Bruix, M.; Thurlkill, R. L.; Schell, D.; Newsom, S.; Grimsley, G. R.; Shaw, K. L.; Treviño, S.; Rico, M.; Briggs, J. M.; Antosiewicz, J. M.; Scholtz, J. M.; Pace, C. N. Charge-Charge Interactions are Key Determinants of the pK Values of Ionizable Groups in Ribonuclease Sa (pI=3.5) and a Basic Variant (pI=10.2). *J. Mol. Biol.* **2003**, *325*, 1077−1092.

(8) Isom, D. G.; Castaneda, C. A.; Cannon, B. R.; Garcia-Moreno, B. Large shifts in pKa values of lysine residues buried inside a protein. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 5260−5265.

(9) Li, H.; Robertson, A. D.; Jensen, J. H. The determinants of carboxyl pKa values in turkey ovomucoid third domain. *Proteins* **2004**, *55*, 689−704.

(10) Quijada, J.; Lopez, G.; Versace, R.; Ramirez, L.; Tasayco, M. L. On the NMR analysis of pKa values in the unfolded state of proteins by extrapolation to zero denaturant. *Biophys. Chem.* **2007**, *129*, 242−250.

(11) Baran, K. L.; Chimenti, M. S.; Schlessman, J. L.; Fitch, C. A.; Herbst, K. J.; Garcia- Moreno, B. E. Electrostatic effects in a network of polar and ionizable groups in staphylococcal nuclease. *J. Mol. Biol.* **2008**, *379*, 1045−1062.

(12) Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; García-Moreno, E. B. Experimental pKa Values of Buried Residues: Analysis with Continuum Methods and Role ofWater Penetration. *Biophys. J.* **2002**, *82*, 3289−3304.

(13) Nelson, K. J.; Parsonage, D.; Hall, A.; Karplus, P. A.; Poole, L. B. Cysteine pK(a) values for the bacterial peroxiredoxin AhpC. *Biochemistry* **2008**, *47*, 12860−12868.

(14) Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. A summary of the measured pK values of the ionizable groups in folded proteins. *Protein Sci.* **2009**, *18*, 247−251.

(15) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(16) Abul Kashem Liton, M.; Idrish Ali, M.; Tanvir Hossain, M. Accurate pKa calculations for trimethylaminium ion with a variety of basis sets and methods combined with CPCM continuum solvation methods. *Comput. Theor. Chem.* **2012**, *999*, 1−6.

(17) Namazian, M.; Zakery, M.; Noorbala, M. R.; Coote, M. L. Accurate calculation of the pKa of trifluoroacetic acid using high-level ab initio calculations. *Chem. Phys. Lett.* **2008**, *451*, 163−168.

(18) 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.

(19) Satchell, J. F.; Smith, B. J. Calculation of aqueous dissociation constants of 1,2,4-triazole and tetrazole: A comparison of solvation modelsElectronic supplementary information (ESI) available: Calculated coordinates, atomic radii and charges of 1,2,4-triazole and tetrazole. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4314−4318.

(20) 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_aValues of Substituted Anilines. *J. Org. Chem.* **2001**, *66*, 6919−6925.

(21) Gross, K. C.; Seybold, P. G.; Hadad, C. M. Comparison of different atomic charge schemes for predicting pKa variations in substituted anilines and phenols. *Int. J. Quantum Chem.* **2002**, *90*, 445−458.

(22) Baptista, A. M.; Martel, P. J.; Petersen, S. B. Simulation of protein conformational freedom as a function of pH: constant-pH molecular dynamics using implicit titration. *Proteins: Struct., Funct., Genet.* **1997**, *27*, 523−544.

(23) Wallace, J. A.; Shen, J. K. Predicting pKa values with continuous constant pH molecular dynamics. *Methods Enzymol.* **2009**, *466*, 455−475.

(24) Itoh, S. G.; Damjanovic, A.; Brooks, B. R. pH replica-exchange method based on discrete protonation states. *Proteins* **2011**, *79*, 3420−3436.

(25) Meng, Y.; Dashti, D. S.; Roitberg, A. E. Computing Alchemical Free Energy Differences with Hamiltonian Replica Exchange Molecular Dynamics (H-REMD) Simulations. *J. Chem. Theory Comput.* **2011**, *7*, 2721−2727.

(26) Swails, J. M.; Roitberg, A. E. Enhancing Conformation and Protonation State Sampling of Hen Egg White Lysozyme Using pH Replica Exchange Molecular Dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 4393−4404.

(27) Bashford, D.; Karplus, M. pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. *Biochemistry* **1990**, *29*, 10219−10225.

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

(29) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. The determinants of pKas in proteins. *Biochemistry* **1996**, *35*, 7819−7833.

(30) Bashford, D. In *Scientific Computing in Object-Oriented Parallel Environments*; Ishikawa, Y., Oldehoeft, R., Reynders, J., Tholburn, M., Eds.; Springer: Berlin, 1997; Lecture Notes in Computer Science Vol. *1343*, pp 233−240.

(31) Dillet, V.; Van Etten, R. L.; Bashford, D. Stabilization of Charges and Protonation States in the Active Site of the Protein Tyrosine Phosphatases: A Computational Study$^+$. *J. Phys. Chem. B* **2000**, *104*, 11321−11333.

(32) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H_{++} 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.* **2012**, *40*, W537−41.

(33) Zhang, S.; Baker, J.; Pulay, P. A reliable and efficient first principles-based method for predicting pK(a) values. 2. Organic acids. *J. Phys. Chem. A* **2010**, *114*, 432−442.

(34) Zhang, S.; Baker, J.; Pulay, P. A reliable and efficient first principles-based method for predicting pK(a) values. 1. Methodology. *J. Phys. Chem. A* **2010**, *114*, 425−431.

(35) Roos, G.; Loverix, S.; Geerlings, P. Origin of the pKa perturbation of N-terminal cysteine in alpha- and 3(10)-helices: a computational DFT study. *J. Phys. Chem. B* **2006**, *110*, 557−562.

(36) Roos, G.; Foloppe, N.; Van Laer, K.; Wyns, L.; Nilsson, L.; Geerlings, P.; Messens, J. How thioredoxin dissociates its mixed disulfide. *PLoS Comput. Biol.* **2009**, *5*, e1000461.

(37) Roos, G.; Foloppe, N.; Messens, J. Understanding the pK(a) of redox cysteines: the key role of hydrogen bonding. *Antioxid. Redox Signaling* **2013**, *18*, 94−127.

(38) Hamann, S. D.; Linton, M. Influence of pressure on the ionization of substituted phenols. *J. Chem. Soc., Faraday Trans. 1* **1974**, *70*, 2239−2249.

(39) Fischer, A.; Leary, G. J.; Topsom, R. D.; Vaughan, J. Ionic dissociation of 4-substituted phenols and 2,6-dichloro- and 2,6-dimethyl-phenols in organic solvents. *J. Chem. Soc. B* **1967**, 846−851.

(40) Fickling, M. M.; Fischer, A.; Mann, B. R.; Packer, J.; Vaughan, J. Hammett Substituent Constants for Electron-withdrawing Substituents: Dissociation of Phenols, Anilinium Ions and Dimethylanilinium Ions. *J. Am. Chem. Soc.* **1959**, *81*, 4226−4230.

(41) Albert, A.; Phillips, J. N. 264. Ionization constants of heterocyclic substances. Part II. Hydroxy-derivatives of nitrogenous six-membered ring-compounds. *J. Chem. Soc.* **1956**, 1294−1304.

(42) De Maria, P.; Fini, A.; Hall, F. M. Thermodynamic acidity constants of orthosubstituted benzenethiols. *J. Chem. Soc., Perkin Trans. 2* **1974**, 1443−1445.

(43) Bolton, P. D.; Hall, F. M.; Reece, I. H. Effects of substituents on the thermodynamic functions of ionisation of meta-substituted phenols. *J. Chem. Soc. B* **1967**, 709−712.

(44) Palm, V. A. *Tables of Rate and Equilibrium Constants of Heterocyclic Organic Reactions*; Moscow, 1976.

(45) Ko, H. C.; O'Hara, W. F.; Hu, T.; Hepler, L. G. Ionization of Substituted Phenols in Aqueous Solution. *J. Am. Chem. Soc.* **1964**, *86*, 1003−1004.

(46) Serjeant, E. P.; Dempsey, B. *Ionisation Constants of Organic Acids in Aqueous Solution*; Pergamon Press: Oxford, U.K., 1979.

(47) *ACD/Structure Elucidator's Manual*; ACD/Labs: Toronto, Ontario, Canada, 2013.

(48) Takahashi, S.; Cohen, L. A.; Miller, H. K.; Peake, E. G. Calculation of the pKa values of alcohols from.sigma. constants and from the carbonyl frequencies of their esters. *J. Org. Chem.* **1971**, *36*, 1205−1209.

(49) Lide, D. *CRC Handbook of Chemistry and Physics*, 87th ed.; CRC Press: Boca Raton, FL, 2006.

(50) De Maria, P.; Fini, A.; Hall, F. M. Thermodynamic acid dissociation constants of aromatic thiols. *J. Chem. Soc., Perkin Trans. 2* **1973**, 1969−1971.

(51) Kreevoy, M. M.; Harper, E. T.; Duvall, R. E.; Wilgus, H. S., III; Ditsch, L. T. Inductive Effects on the Acid Dissociation Constants of Mercaptans. *J. Am. Chem. Soc.* **1960**, *82*, 4899−4902.

(52) Arnold, A. P.; Canty, A. J. Methylmercury(II) sulfhydryl interactions. Potentiometric determination of the formation constants for complexation of methylmercury(II) by sulfhydryl containing amino acids and related molecules, including glutathione. *Can. J. Chem.* **1983**, *61*, 1428−1434.

(53) Pettit, L. D.; Powell, K. *The IUPAC Stability Constants Database*; IUPAC and Academic Software: Otley, West Yorkshire, U.K., 2006.

(54) Kreevoy, M. M.; Eichinger, B. E.; Stary, F. E.; Katz, E. A.; Sellstedt, J. H. The Effect of Structure on Mercaptan Dissociation Constants. *J. Org. Chem.* **1964**, *29*, 1641−1642.

(55) Tsonopoulos, C.; Coulson, D. M.; Inman, L. B. Ionization constants of water pollutants. *J. Chem. Eng. Data* **1976**, *21*, 190−193.

(56) Irving, R. J.; Nelander, L.; Wadsö, I.; Halvarson, H.; Nilsson, L. Thermodynamics of the Ionization of Some Thiols in Aqueous Solution. *Acta Chem. Scand.* **1964**, *18*, 769−787.

(57) G. Kortm, W. V.; Andrussow, K. *Dissociation Constants of Organic Acids in Aqueous Solution*; Butterworths: London, 1961.

(58) Hupe, D. J.; Jencks, W. P. Nonlinear structure-reactivity correlations. Acyl transfer between sulfur and oxygen nucleophiles. *J. Am. Chem. Soc.* **1977**, *99*, 451−464.

(59) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision B.01; Gaussian, Inc.: Wallingford, CT, 2010.

(60) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785−789.

(61) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098−3100.

(62) Becke, A. D. A new mixing of Hartree-Fock and local density-functional theories. *J. Chem. Phys.* **1993**, *98*, 1372−1377.

(63) Handy, N. C.; Cohen, A. J. Left-right correlation energy. *Mol. Phys.* **2001**, *99*, 403−412.

(64) Hoe, W.-M.; Cohen, A. J.; Handy, N. C. Assessment of a new local exchange functional OPTX. *Chem. Phys. Lett.* **2001**, *341*, 319−328.

(65) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.

(66) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158.

(67) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, non-covalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215−241.

(68) Zhao, Y.; Truhlar, D. G. Density functionals with broad applicability in chemistry. *Acc. Chem. Res.* **2008**, *41*, 157−167.

(69) Zhao, Y.; Truhlar, D. G. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.* **2006**, *125*, 194101.

(70) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(71) Stewart, J. J. P. Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem.* **1989**, *10*, 209−220.

(72) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173−1213.

(73) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. RM1: a reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comput. Chem.* **2006**, *27*, 1101−1111.

(74) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009**, *113*, 6378−6396.

(75) Scalmani, G.; Frisch, M. J. Continuous surface charge polarizable continuum models of solvation. I. General formalism. *J. Chem. Phys.* **2010**, *132*, 114110.

(76) 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.

(77) Mulliken, R. S. Electronic Population Analysis on LCAO[Single Bond]MO Molecular Wave Functions. I. *J. Chem. Phys.* **1955**, *23*, 1833−1840.

(78) Löwdin, P.-O. On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals. *J. Chem. Phys.* **1950**, *18*, 365−375.

(79) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural population analysisa. *J. Chem. Phys.* **1985**, *83*, 735−746.

(80) Singh, U. C.; Kollman, P. A. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.* **1984**, *5*, 129−145.

(81) Breneman, C. M.; Wiberg, K. B. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comput. Chem.* **1990**, *11*, 361−373.

(82) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269−10280.

(83) Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007−1023.

(84) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999−2012.

(85) Case, D. et al. *AMBER 12*; University of California: San Fransisco, CA, 2012.

(86) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(87) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* **2004**, *55*, 383−394.

(88) Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **1980**, *72*, 2384−2393.

(89) Gotz, A. W.; Clark, M. A.; Walker, R. C. An extensible interface for QM/MM molecular dynamics simulations with AMBER. *J. Comput. Chem.* **2014**, *35*, 95−108.

(90) Roos, G.; Geerlings, P.; Messens, J. Enzymatic catalysis: the emerging role of conceptual density functional theory. *J. Phys. Chem. B* **2009**, *113*, 13465−13475.

(91) Svobodova Varekova, R.; Geidl, S.; Ionescu, C.-M.; Skrehota, O.; Kudera, M.; Sehnal, D.; Bouchal, T.; Abagyan, R.; Huber, H. J.; Koca, J. Predicting pK(a) values of substituted phenols from atomic charges: comparison of different quantum mechanical methods and charge distribution schemes. *J. Chem. Inf. Model.* **2011**, *51*, 1795−1806.