# Using atomic charges to describe the p$K_a$ of carboxylic acids

Zeynep Pinar Haslak,[†,‡] Sabrina Zareb,[†] Ilknur Dogan,[‡] Viktorya Aviyente,[‡] and Gerald Monard*[,†]

†*Université de Lorraine, CNRS, LPCT, F-54000 Nancy, France*

‡*Department of Chemistry, Boğaziçi University, 34342 Bebek, Istanbul, Turkey*

E-mail: Gerald.Monard@univ-lorraine.fr

## Abstract

In this study, we present an accurate protocol for the fast prediction of p$K_\mathrm{a}$'s of carboxylic acids based on the linear relationship between computed atomic charges of the anionic form of the carboxylate fragment and their experimental p$K_\mathrm{a}$ values. Five charge descriptors, three charge models, three solvent models, gas phase calculations and several DFT methods (combination of eight DFT functionals and fifteen basis sets) were tested. Among those, the best combination to reproduce experimental p$K_\mathrm{a}$'s is to compute NPA atomic charge using the SMD model at the M06L/6-311G(d,p) level of theory and selecting the maximum atomic charge on the carboxylic oxygen atoms ($R^2$ = 0.955). The applicability of the suggested protocol and its stability along geometrical changes are verified by molecular dynamics simulations performed for a set of aspartate, glutamate and alanine peptides. By reporting the calculated atomic charge of the carboxylate form into the linear relationship derived in this work, it should be possible to estimate accurately the amino acid's p$K_\mathrm{a}$'s in protein environment.

# Introduction

A large number of chemical and biological systems contain acidic and basic groups. These groups can strongly interact with their surroundings, usually via electrostatics and hydrogen bond interactions. Their impact on the functions of biological systems can be very large. At a particular pH, the extent to which an ionizable species can be protonated or deprotonated by the hydrogen transfer from/to the environment is determined by the p$K_\mathrm{a}$ of the species. Most of the drug molecules are weak acids or weak bases and when they are in solution they are in their both ionized and nonionized states. Solubility, lipophilicity and permeability of a drug ligand in a cell membrane is governed by the p$K_\mathrm{a}$'s of the acidic and basic sites within the molecule, since only the uncharged ligands can penetrate into the cell membrane.[1] Besides, the interactions between the ionizable functional groups of a ligand with the residues of its target protein, which affects the affinity, activity and efficacy of that ligand, is highly dependent on the p$K_\mathrm{a}$'s of the side chains in the active site and of the drug molecule. Moreover, the changes in the protonation states of amino acid residues can have a direct impact on establishing protein conformation and stability,[2] solubility and folding,[3] catalytic activity of enzymes[4] and their binding ability.

Carboxylic acids are the main acidic functional groups in biological systems. Glutamate and aspartate have carboxylic acid groups in their side chains and these groups help in holding the peptide together by hydrogen bonds. More than 30% of the ionizable residues (32% of the Arg residues, 19% of the Asp residues, 13% of the Glu residues, and 6% of the Lys)[5] are buried inside the hydrophobic cavities which limits the contact with solvent.[6] Since the protein matrix is heterogeneous, the fluctuations in the electrostatic environment alter the interactions between buried charges which in turn leads to modifications in the affinities of the protonation sites for ionization; and thus their p$K_\mathrm{a}$ values are re-adjusted.[7] Eventually, in polar parts of the protein the p$K_\mathrm{a}$ of the acidic groups in the residues shifts to higher values and the p$K_\mathrm{a}$ of the basic groups shifts to lower values from those of the isolated amino acids.[8] Hydrogen bondings between the amino acid's functional groups and the side chain or the backbone atoms also tend to result in p$K_\mathrm{a}$ deviations; especially when the number of H-bonds increases and if they are rigid the effect is larger such that the p$K_\mathrm{a}$ for acidic side chains are perturbed above their intrinsic p$K_a$ values and for the basic groups the reverse is observed.[3,9] Salt-bridge formation between two residues, which contributes to protein stability, is also reported to result in lower or higher p$K_a$ values with the same trends in polarization and hydrogen bonding effects.[10]

Measuring p$K_a$'s of molecules or part of molecules in large medias by experimental means is complex and difficult.[11,12] Thus, the need for accurate p$K_a$ estimations by the applications of theoretical approaches is necessary.[13] The features that determine the acidities of different classes of chemical compounds can be explained by the molecular structure. The traditional method for the calculation of p$K_a$'s is based on the free energy changes in the thermodynamic cycle. Typically electrostatic interactions are obtained by numerically solving the linearized Poisson-Boltzmann equation (LPBE). Despite the enormous number of successful p$K_a$ predictions by using the deprotonation energies and solvation free energies,[14–17] these calculations usually fail in their purpose due to the instability of the ion in gas phase and the conformational differences between the solvent and gas phase calculations.[18] Besides, empirical methods such as PROPKA and the methods based on Poisson-Boltzmann equation, Generalized Born equation, QM/MM or Molecular Dynamics or a combination of one or more; quantitative structure property relationship (QSPR) is a widely used technique in which several molecular descriptors are successfully linked to p$K_a$'s of organic molecules such as topological state,[19,20] atom type,[21,22] group philicity,[23] bond length and frequency,[24,25] maximum surface potential,[26] HOMO and LUMO energies,[27,28] atomic charge.[29,30] Among them, the concept of partial atomic charges is closely related to the relative acidity and basicity of a molecule.[31,32]

A Multiple Linear Regression model was developed by Dixon and Jurs with an accuracy of 0.5 units for the calculation of p$K_a$'s of oxyacids by using the empirical atomic charges of atoms in a molecule.[30] The model is based on the changes in the $\sigma$ and $\pi$ charges upon going from the neutral to ionic state, concerning the resonance and inductive effects of nearby atoms. Citra constructed four linear regression models by using the partial atomic charges on oxygen and hydrogen atoms which are involved in deprotonation and O-H bond order for the set of phenols, alcohols and aromatic and non-aromatic carboxylic acids.[33]Various combinations of different level of theories, basis sets and charge models were tested by Vareková *et al.* in order to create a model for phenols.[34] Recently, Ugur *et al.* made use of a similar approach with an extended study for the prediction of amino acid p$K_a$'s in proteins and developed an accurate protocol by computing the atomic charge on the anionic form of alcohols and thiols.[35] Among the tested DFT functionals, basis sets, semiempirical methods, solvation and charge models, they observed the best combination is NPA charge calculation in CPCM model at the B3LYP/3-21G ($R^2$=0.995) level of theory for alcohols and M06-2X/6-311G ($R^2$=0.986) level of theory for thiols in order to reproduce the experimental p$K_a$'s. Moreover, they tested the stability of the calculated p$K_a$'s in amino acids by MM-MD and DFT-MD calculations. Regarding the successful applications of QM charges as descriptors, in this study we aim to suggest an accurate protocol for the fast prediction of p$K_a$'s of carboxylic acids.

# Computational Details

## Experimental Database

From literature,[36,37] we have selected a total of 59 carboxylic acid compounds with p$K_a$'s ranging from 0.65 to 5.12. We have selected molecules which have the widest range of experimental p$K_a$'s as possible. Most of these molecules are also small and rather rigid molecules. We have avoided flexible molecules in order to overcome the risk of failing to obtain their global minima during geometry optimization, which would raise systematical errors in p$K_a$ predictions.[37] A training set of 30 small molecules (see Table 1 and Figure S1) and a test set of 29 small molecules (see Table 2 and Figure S2) have been extracted from the ensemble.

## Quantum Mechanical Calculations

All of the Quantum Mechanical (QM) calculations were carried out using the Gaussian 09[38] program package. Eight different density functionals (BLYP,[39,40] B3LYP,[39,41] OLYP,[39,42] PBE,[43] PBE0,[44] M06,[45,46] M06L,[46,47] M062X[45,46]) and fifteen different basis sets were used. To interpret the aqueous solvent environment, the universal solvent model (SMD[48]), the polarizable continuum model (PCM[49]), and the polarizable conductor solvent model (CPCM[50]) were employed with a dielectric constant ($\varepsilon$) of 78.5. Three different types of atomic charge models were tested: Mulliken population analysis,[51] Löwdin population analysis,[52] Natural

Population Analysis (NPA).[53] Compared to the study from Ugur et al.,[35] Electrostatic Potential (ESP) derived atomic charges, like the Merz-Kollman (MK) model[54] and the CHelpG model,[55] are not reported here since preliminary studies have shown us that, as in the cases of thiols and alcohols, they do not perform better than NPA atomic charges (data not shown). Unless otherwise stated, all the charge calculations were performed on the optimized geometries (after including or not the solvent effect) that do not contain any imaginary frequency.

## Molecular Dynamics Simulations

Molecular dynamics simulations have been performed using the AMBER biomolecular package.[56] All simulated molecules have been modeled with the AMBER ff14SB protein force field.[57] The aqueous polar environment was mimicked by the implicit modified generalized Born model with $\alpha$, $\beta$, $\gamma$ are 1.0, 0.8, and 4.85[58] as implemented in AMBER 18 (igb = 5). Following minimization, the systems were heated up to 300 K using the Langevin thermostat during 50 ps with a collision frequency $\gamma = 10$ ps$^{-1}$, and a timestep of 1 fs. Then, NVT production runs were performed for another 150 ps using the same thermostat algorithm. From each of these molecular dynamics, 1500 frames were extracted, one every 0.1 ps.

# Results and Discussions

The linear relationship between atomic charges and experimental p$K_a$'s depends on many factors: the choice of the DFT method, the choice of the basis set, the use (or not) of an implicit solvent model, the type of the atomic charge model, and which atomic charges are considered. From the overall present study (see Supplementary Information for the full detailed results), we have found that the best combination of all these factors is to consider the highest oxygen atomic charge of each carboxylate fragment computed with NPA at the M06L/6-311G(d,p) level using the SMD implicit solvent model. In what follows, we present a linear relationship between experimental p$K_a$'s and atomic charges computed using the theoretical framework discussed above. Then, using these results as a reference, we discuss the choice of charge descriptor, charge model, solvent model, DFT functional and basis set by changing one of these parameters while the others remain fixed to their best combination.

## Linearity of the Relationship Between Experimental p$K_\mathrm{a}$'s and Atomic Charges

For each moleule of the training set, a geometry optimization was performed at the M06L/6-311G(d,p) level using the SMD implicit solvent model. We ensure that no imaginary frequency remains for any molecule. Atomic charges were computed using the natural population analysis. For each carboxylate fragment, we extracted the highest of the two oxygen atomic charges and we compared it with the experimental p$K_\mathrm{a}$ of the corresponding molecule. Figure 1 shows the relationship between experimental p$K_\mathrm{a}$ and computed NPA charge for the training set. A linear equation is obtained by a least-square fit:

$$\mathrm{p}K_\mathrm{a} = a \cdot Q + b \quad \text{with} \quad Q = \max\{q(\mathrm{O}_1), q(\mathrm{O}_2)\} \tag{1}$$

where $a$ and $b$ are the fitted parameters and $Q = \max\{q(\mathrm{O}_1), q(\mathrm{O}_2)\}$ is the highest atomic charges of the two carboxylate oxygens, respectively. The parameters $a$ and $b$ and the squared Pearson correlation coefficient ($R^2$) are also illustrated in Figure 1. The predicted p$K_\mathrm{a}$'s are computed using Eq. 1 (i.e., by reporting $\max\{q(\mathrm{O}_1), q(\mathrm{O}_2)\}$ of a given molecule into the parametrized equation).

For carboxylate molecules, the $R^2$ value has been found to be 0.955. No strong outlier molecule was observed for the training set. The maximum difference between predicted and experimental p$K_\mathrm{a}$ among all the molecules was found as 0.60 units (see Table 1). These results indicate a strong correlation between experimental p$K_\mathrm{a}$'s and the oxygen charges.

In order to analyze the influence of the charge descriptor, charge model and solvent model on the quality of the fit, the same protocol was applied with four other charge descriptors, two other charge models, two other solvent models and gas phase calculations.

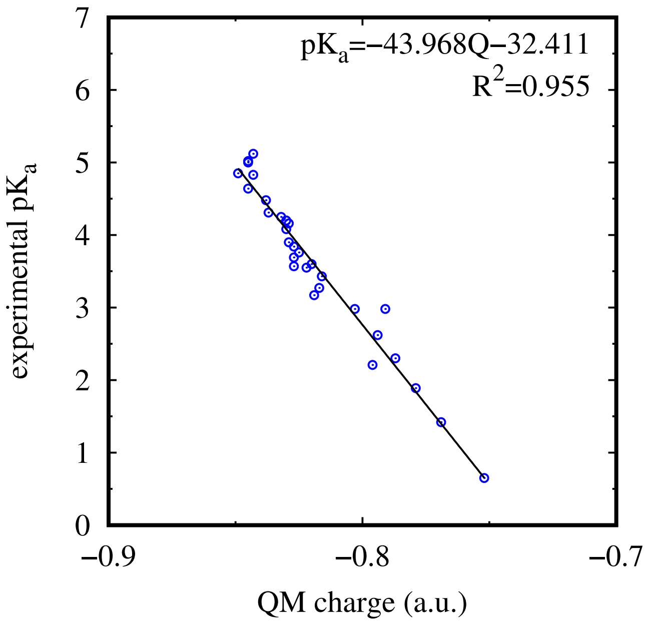


Figure 1: Linear regression between calculated NPA atomic charges and experimental p$K_a$. Calculations were done using M06L/6-311G(d,p)//SMD.

## Influence of the charge descriptor

Compared to alcohols and thiols that were analyzed by Ugur et al.,[35] the negative charge of the base form in the case of carboxylate can be shared between different atoms: the carbon and the two oxygen atoms of the carboxylate fragment. Thus, there are different ways to extract atomic charges for this fragment and then to compare them with experimental p$K_a$'s. We have analyzed different atomic extraction schemes for the negative charge $Q$ of the carboxylate fragment composed of atoms C, $O_1$ and $O_2$:

$$
\begin{aligned}
Q &= \max\{q(O_1), q(O_2)\} && (2)\\
Q &= \min\left\{q(O_1), q(O_2)\right\} && (3)\\
Q &= \frac{1}{2}\left[q(O_1) + q(O_2)\right] && (4)\\
Q &= q(C) + q(O_1) + q(O_2 && (5)\\
Q &= q(C) && (6)
\end{aligned}
$$

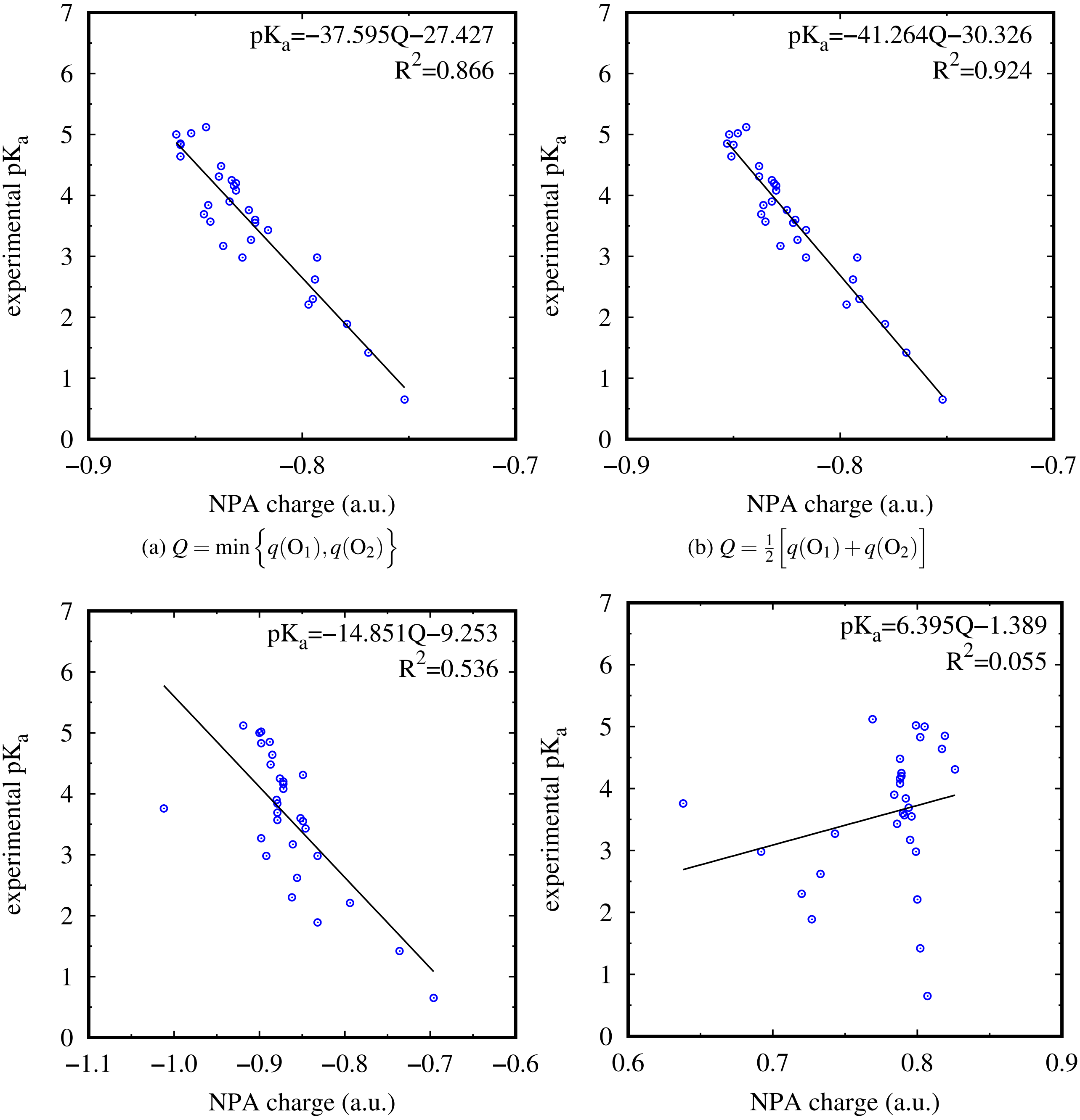
From the two oxygen atomic charges, it is possible to extract the highest value (Eq. 2), the lowest value (Eq. 3), or the average (Eq. 4). The carbon atomic charge can also be taken into account via the sum of all 3 atomic charges (Eq. 4) or by itself (Eq. 6).

Figure 2 shows the relationship between carboxylate atomic charges expressed by Eqs.3-6 and experimental p$K_\mathrm{a}$'s using M06L/6-311G(d,p)//SMD. When the lowest (i.e., the most negative) oxygen atomic charge is considered, the linear relationship is less accurate than with the highest oxygen atomic charge scheme: $R^2$ = 0.866 for the "min" scheme *vs.* $R^2 = 0.955$ for the "max" scheme, respectively. This is somewhat unexpected, since if one considers a proton, one could expect it to be more attracted by the most negative oxygen atoms. Therefore, one could expect that the $Q = \min\left\{q(\mathrm{O}_1), q(\mathrm{O}_2)\right\}$ scheme should better reflect the experimental p$K_\mathrm{a}$'s. In all our linear regressions with different density functionals, basis sets, etc., we have never found a better regression with the scheme $Q = \min\left\{q(\mathrm{O}_1), q(\mathrm{O}_2)\right\}$ than with its $Q = \max\{q(\mathrm{O}_1), q(\mathrm{O}_2)\}$ counterpart. As a consequence the scheme $Q = \frac{1}{2}\left[q(\mathrm{O}_1) + q(\mathrm{O}_2)\right]$ that computes the average of the two oxygen atomic charges is placed in between the two previous scheme with $R^2 = 0.924$.

Another possibility to search for a relationship between experimental p$K_\mathrm{a}$ and atomic charge is to take into account the atomic charge on the carboxylate carbon. Figure 2(d) shows the (lack of) relationship between the carbon atomic charges and experimental p$K_\mathrm{a}$'s. With a $R^2 = 0.055$, the carbon charge cannot be regarded as a descriptor of the experimental p$K_\mathrm{a}$. As a consequence, when the three atomic charges on the carboxylate fragment are considered together (Eq. 5), the correlation coefficient ($R^2 = 0.536$) is worse than when the carbon atom is not included.

## Influence of the charge model

In a p$K_\mathrm{a}$ prediction model, the variations in the p$K_\mathrm{a}$ during the dissociation process should be reflected precisely by the electronic changes. Three different charge schemes were tested for their predictivity power to generate charges that associate with the experimental p$K_\mathrm{a}$'s: NPA[53] as well as Mulliken[51] and Löwdin[52] population analysis. These methods are based on charge partition schemes and define the atomic orbitals by wave functions. In the Mulliken population analysis, the calculated electron density is equally shared through the adjacent atoms in a molecule. Löwdin population analysis is very similar to the Mulliken method with only difference in usage of orthogonal basis functions. Neither Löwdin or Mulliken schemes are



(a) $Q = \min\left\{q(\mathrm{O}_1), q(\mathrm{O}_2)\right\}$

(b) $Q = \frac{1}{2}\left[q(\mathrm{O}_1) + q(\mathrm{O}_2)\right]$

(c) $Q = q(\mathrm{C}) + q(\mathrm{O}_1) + q(\mathrm{O}_2)$

(d) $Q = q(\mathrm{C})$

Figure 2: Effect of the charge descriptor on the linear regression between calculated atomic charges and experimental p$K_a$'s. Calculations were done with M06L/6-311G(d,p)//SMD: (a) Minimum atomic charge on $\mathrm{O}_1$ and $\mathrm{O}_2$; (b) Average sum of atomic charges on $\mathrm{O}_1$ and $\mathrm{O}_2$; (c) Sum of atomic charges on C, $\mathrm{O}_1$ and $\mathrm{O}_2$; (d) Atomic charge on C.

able to reproduce the values of the dipole moments and they are both dependent on the basis set that is used. Natural population analysis localizes and classifies the orbitals into core, valence and Rydberg each of which contribute differently to the density. This partititoning of the atomic orbitals makes the NPA method less basis set dependent than its counterparts.

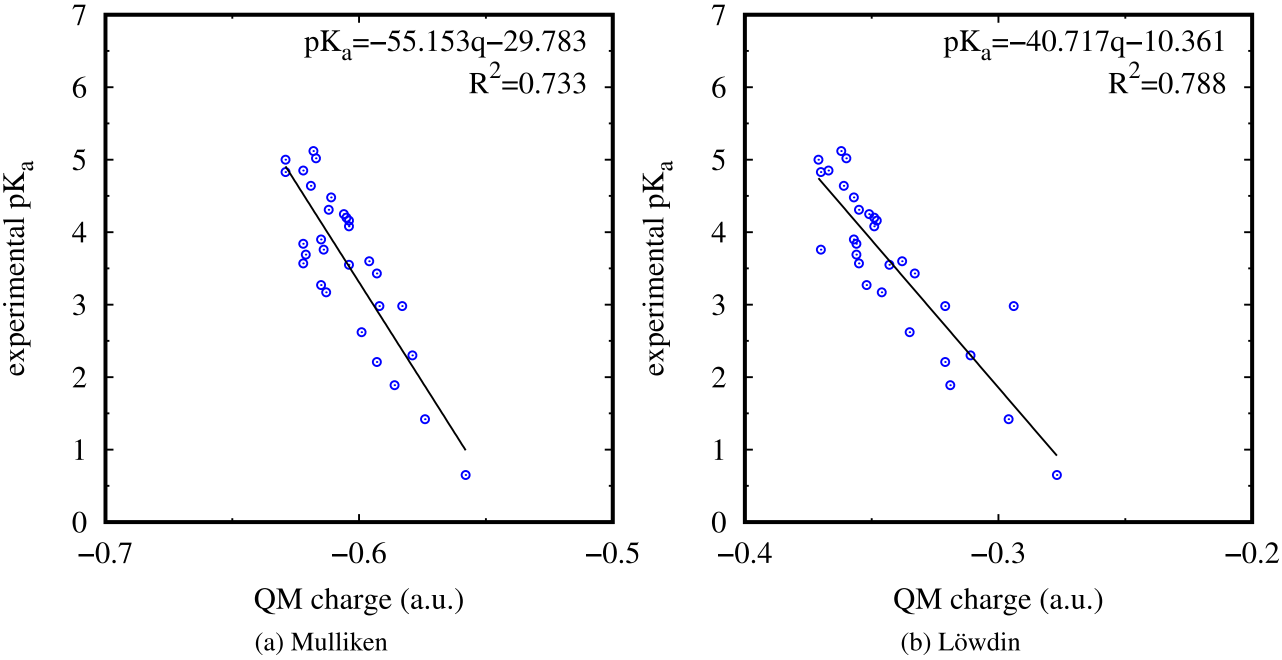


Figure 3: Effect of the charge model on the linear regression between calculated atomic charges and experimental p$K_a$'s. Calculations were done with M06L/6-311G(d,p)//SMD: (a) Mulliken atomic charge model; (b) Löwdin atomic charge model

The strength of a carboxylic acid is determined by the strength of its conjugate base and the strength of a base is proportional to the charge density on the carboxylate oxygens. The lesser the charge density on the oxygen atoms means more stability and thus it becomes a weaker base and finally a stronger acid. Figure 3 presents the linear regressions between the highest oxygen atomic charge and experimental p$K_a$ for the training set at the M06L/6-311G(d,p)//SMD using the Mulliken population analysis (Figure 3(a)) and the Löwdin population analysis (Figure 3(b)). The charge analysis shows that the oxygen charges become more negative with increasing p$K_a$, suggesting that an oxygen atom with more associated electron density readily accepts a proton; indication of a stronger conjugate base and thus a weaker acid. Mulliken and Löwdin charges give $R^2$ coefficients lower than that of NPA with values of 0.733 and 0.788 respectively. This result is similar to those obtained for alcohols and thiols by Ugur et al.:[35] atomic charges extracted from natural population analysis are more linearly correlated to p$K_a$'s than using the Mulliken's or Löwdin's schemes. Using Eq. 1, the calculated p$K_a$ of the strongest outlier is 1.35 unit different from the experimental p$K_a$ when Löwdin charges are used (Table S1). In case of Mulliken scheme, all predicted p$K_a$'s are within $\pm 1$ unit range, no strong outliers are observed (Table S1).

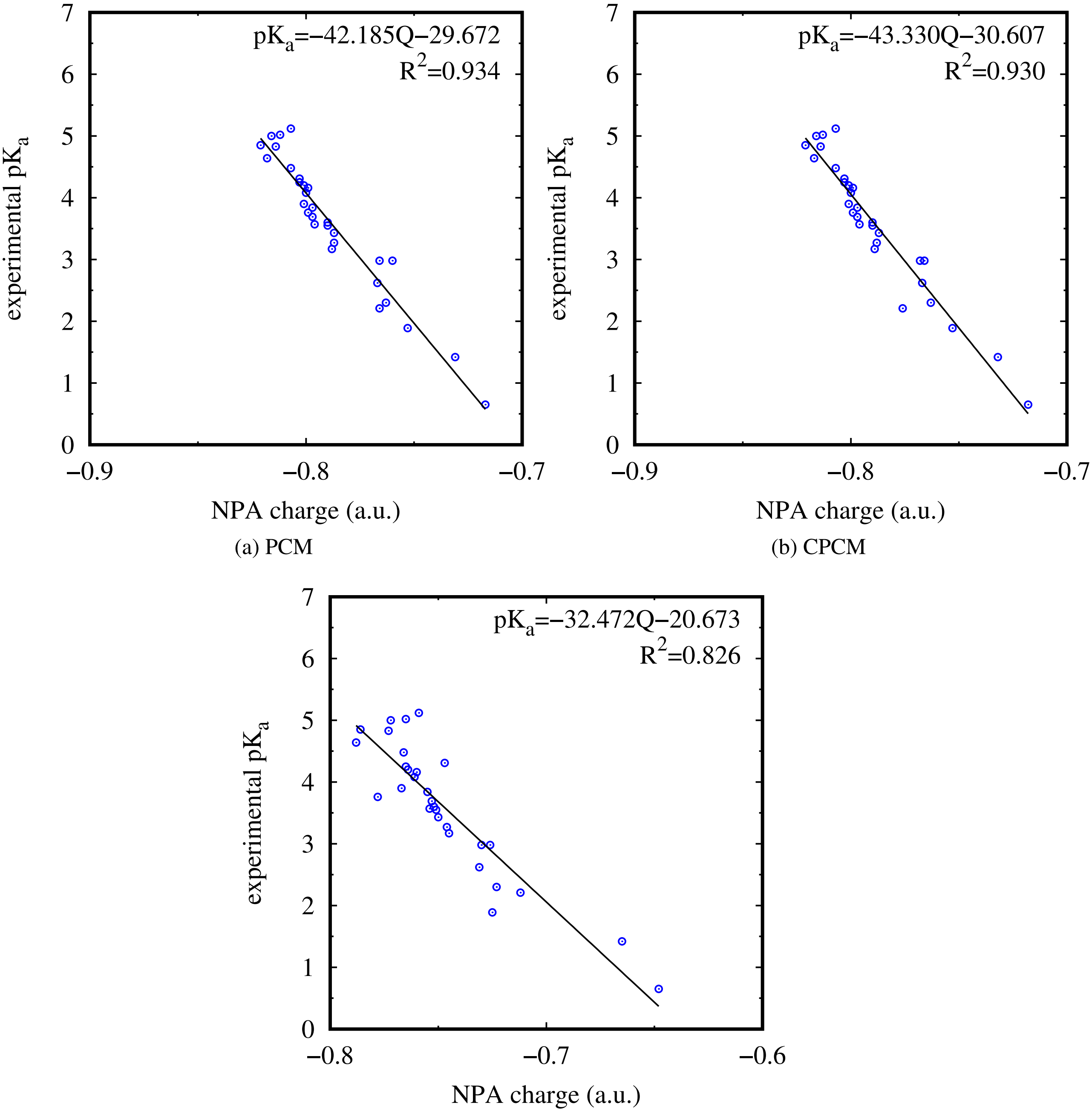
## Influence of the solvent model

The description of the surrounding environment that the charged species is exposed to accounts for the ideal charge derivation scheme. Implicit solvent models offer some advantages for modeling the interactions between the solute and the solvent. In this part of the study, we have tested the accuracy of PCM and CPCM implicit solvation models in addition to SMD model calculations. Besides, due to its smaller computational costs, gas phase calculations have also been taken into consideration. Figure 4 presents the linear regression fits of CPCM, PCM and gas phase calculations using NPA charges and the DFT method as discussed in the previous sections.

Both PCM and CPCM calculations are as accurate as SMD calculations with $R^2$=0.934 and $R^2$=0.930, respectively (Figure 4 (a) and (b)). The predictivity of gas phase model is poorer ($R^2$ = 0.826, Figure 4 (c)) compared to other models where PCM, CPCM and SMD solvation methods are applied since in this study we have extracted the water phase acidities rather than gas-phase proton affinities. SMD model is different from PCM and CPCM models in considering the dispersion-repulsion energies in addition to electronic energy. These additional terms seem to contribute in finding the global minimum in geometry optimizations and assigning the atomic charges. Maximum deviations of the predicted p$K_a$'s from the experimental p$K_a$'s are found to be 0.75, 0.80 and 1.13 units for PCM (Table S3), CPCM (Table S2) and gas phase calculations (Table S4), respectively.

## Density Functionals and Basis Set Benchmarks

A deep analysis of the influence of DFT functionals and basis sets on p$K_a$ prediction capability for carboxylic acids have been performed by applying the same protocol to the training set.



(a) PCM

(b) CPCM

(c) Gas Phase

Figure 4: Effect of the implicit solvent model on the linear regression between calculated atomic charges and experimental p$K_a$'s. Calculations were done with M06L/6-311G(d,p): (a) PCM model (b) CPCM model (c) gas phase.

Highest NPA charge on the oxygen atoms of carboxylate fragment calculated at various level of theories with SMD model were extracted to obtain $R^2$, $a$ and $b$ values in Eq. 1 from the linear fit with experimental p$K_a$'s. In Figure 5, for each combination of DFT functional and basis set, the Mean Absolute Deviations (MADs) are presented as box representations. The differences between the experimental and predicted p$K_a$'s ($\Delta$p$K_a$) have been calculated for each level of theory and the maximum value of this difference (MAX-$\Delta$p$K_a$) is represented as black colored lines in Figure 5.

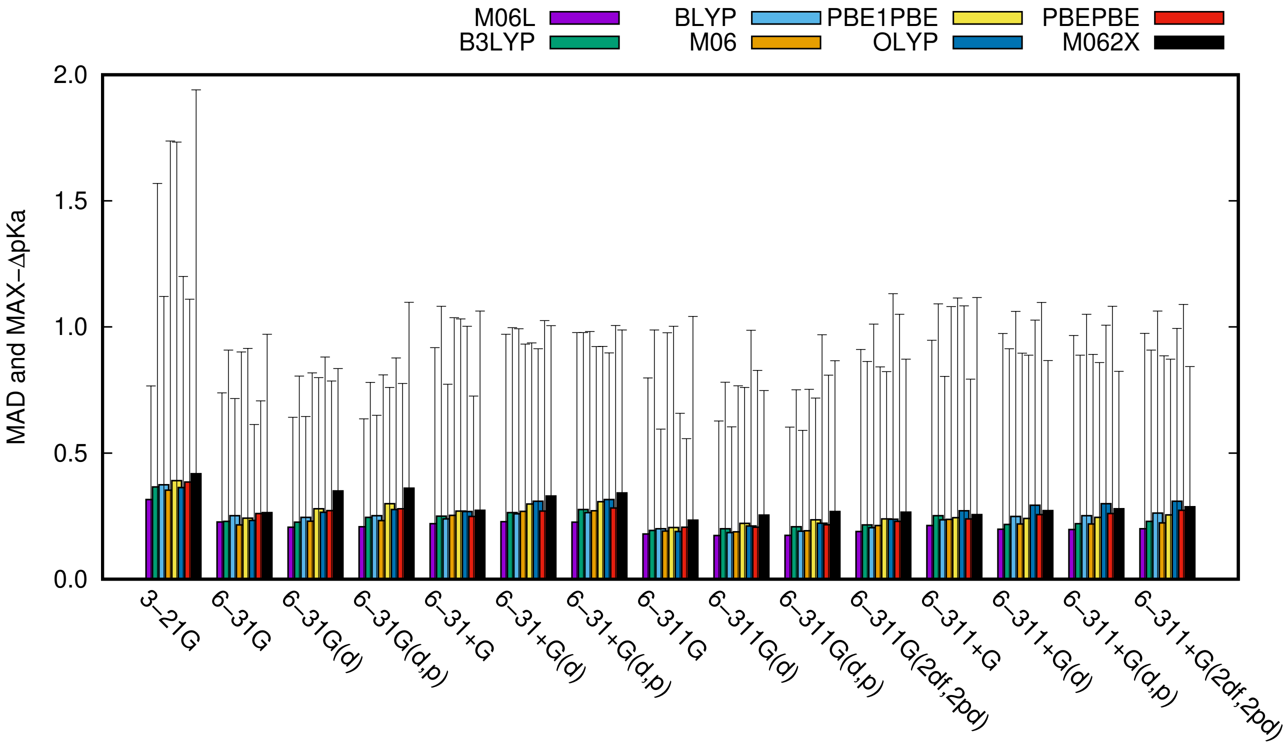


Figure 5: . Mean Absolute Deviation (MAD) and maximum difference between predicted and experimental pKa (MAX-$\Delta$p$K_a$) for eight different DFT functionals and fifteen different basis sets considered in this work. Geometry optimizations and NPA charge calculations were done using the SMD model.

All of the DFT methods gave strong correlations between calculated NPA atomic charges and experimental p$K_a$'s with $R^2$ range of $0.702 \leq R^2 \leq 0.955$. The largest MADs and MAX-$\Delta$p$K_a$'s were found for the combinations of 3-21G basis set with all the functionals except M06L. Removing the (small) 3-21G basis set combinations from the benchmark study, we obtained high accuracy range of MAD and $\Delta$p$K_a$ values ($0.17 \leq$ MAD $\leq 0.36$ and $0.56 \leq$ MAX-$\Delta$p$K_a \leq 1.13$). The power of the predictivity slightly diminishes with the addition of diffuse functions to the basis set for any of the DFT functionals (i.e. 6-31+G* has higher MAD and MAX-$\Delta pK_a$ compared to 6-31G*). On the other hand, polarization functions did not cause any significant improvement. Regarding the performance of the functionals, in all subsets the largest MADs were obtained with either M06-2X or OLYP functionals. The smallest MADs were found for the combinations of all basis sets with the M06L functional (except 6-31G) and among all the tested methods M06L/6-311G(d,p) gave the most accurate result with MAD value of 0.174. When we applied the Eq. 1 to the test set, the MAD value for the predicted $pK_a$'s was found to be 0.199 and the MAX-$\Delta pK_a$ was found to be 0.87.

The average predicted $pK_a$ over all the methods has been calculated in order to have an overview on the efficiency of the level of theory. The minimum and maximum predicted $pK_a$'s among all the methods (except 3-21G basis set due to its large MAD and MAX-$\Delta pK_a$) were added to the average predicted $pK_a$ of each molecule as error bars. The predicted $pK_a$ is plotted versus experimental values for both training and test sets (Figure 6). Minimum, maximum and average values of the predicted $pK_a$ were found to be within the range of ±1 unit compared to the experimental value.

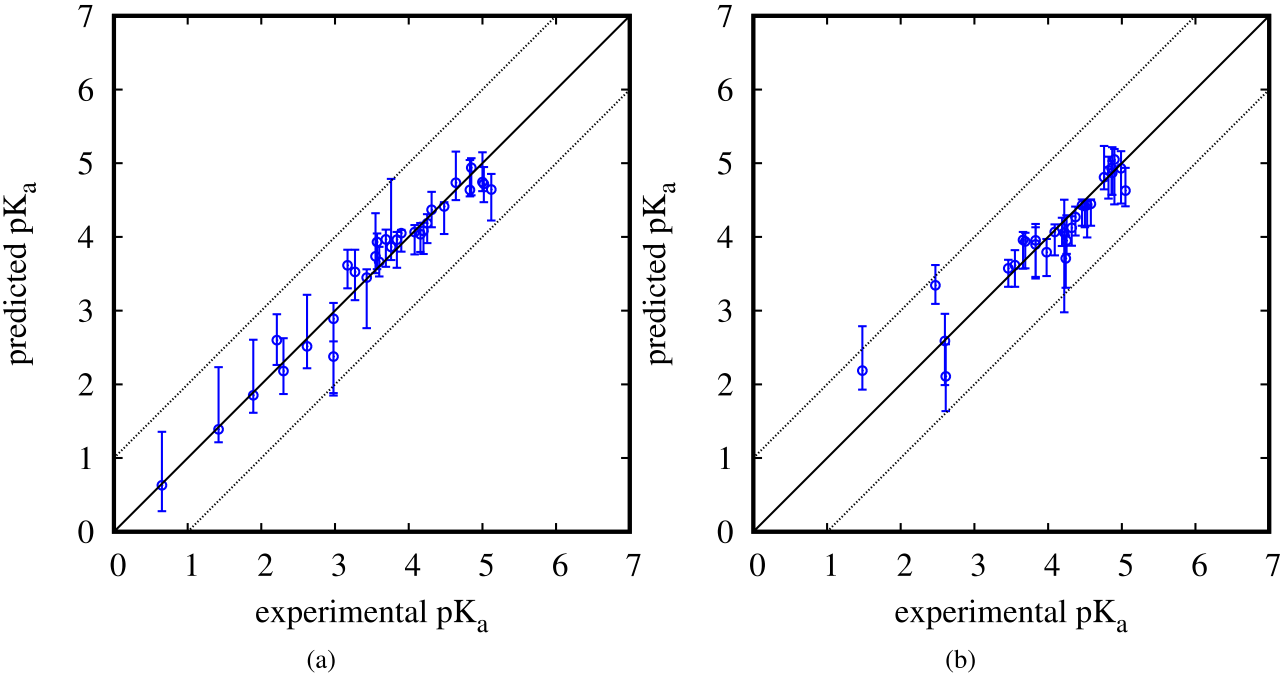


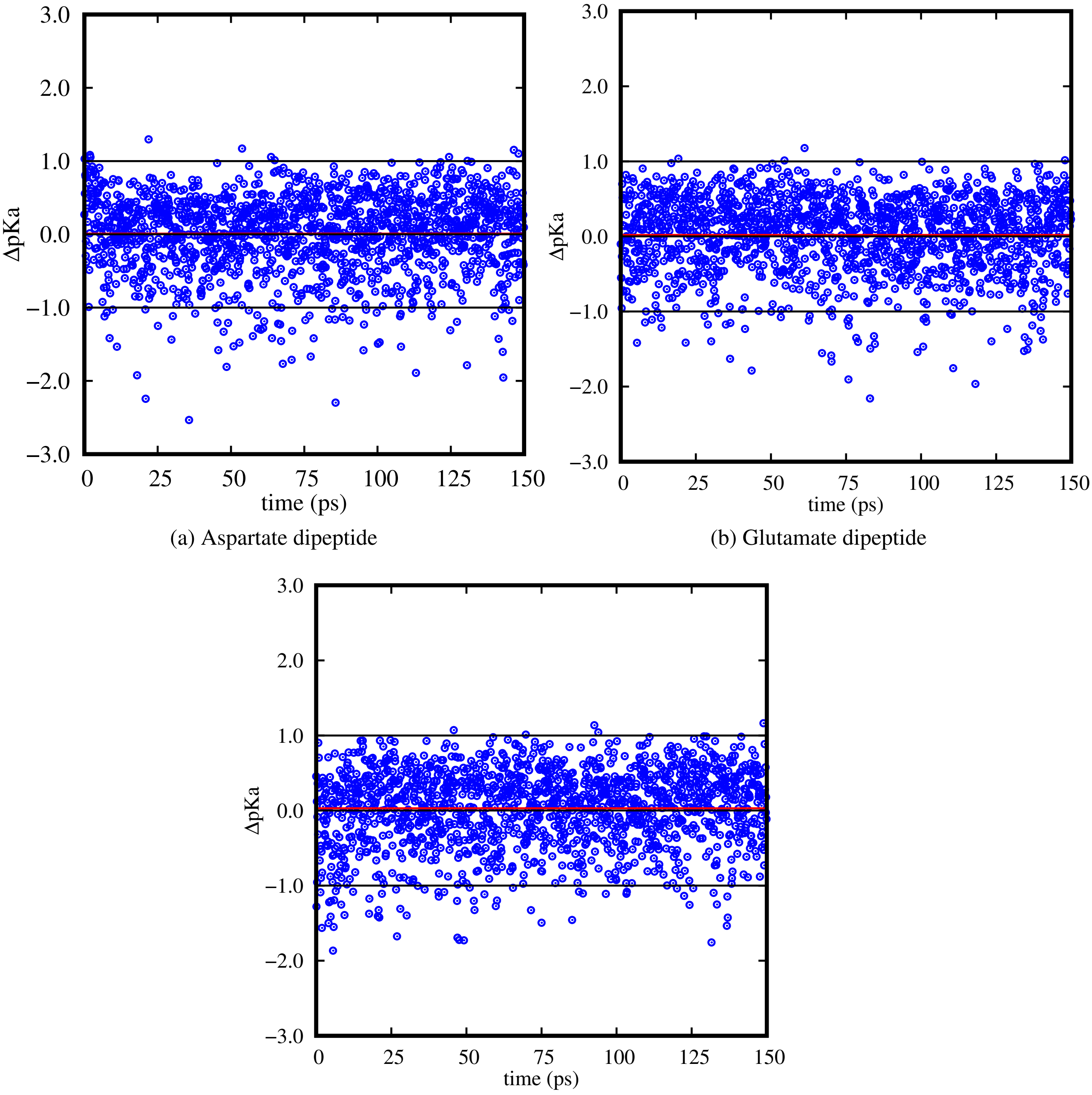
Figure 6: Predicted $pK_a$ over all the DFT functionals and basis sets (3-21G discarded) versus experimental $pK_a$ for the Training (a) and Test (b) sets (solvation model=SMD, charge model=NPA). Circles show the average $pK_a$, and the error bars denote minimum and maximum predicted $pK_a$.

### Stability of the Prediction Along Geometry Changes

The stability of the calculated p$K_\mathrm{a}$'s with respect to geometrical changes is crucial for the p$K_\mathrm{a}$ predictions of proteins. Short molecular dynamics simulations (150 ps) for N-acetyl alanine and dipeptide forms of aspartate and glutamate were performed in order to provide multiple geometries around the optimum structures and to establish the variability of the p$K_\mathrm{a}$ prediction with respect to geometrical changes. A total of 1500 frames were extracted from these MD simulations and single point NPA charge calculations were performed on these geometries by using SMD with the M06L/6-311G(d,p) method. The predicted p$K_\mathrm{a}$'s were obtained using *a* and *b* values derived from the fit. The experimental p$K_\mathrm{a}$'s (p$K_\mathrm{a}$ [aspartate]=3.94,[59] p$K_\mathrm{a}$ [glutamate]=4.25,[60] p$K_\mathrm{a}$ [alanine]=3.67[60]) were taken as a reference and the fluctuations of the calculated p$K_\mathrm{a}$'s with respect to geometrical changes were observed. The average value over all the frames were calculated and found to be in very good agreement with the experimental values for three of the peptides (red line in Figure 7). Almost 95% of the predictions are within ±1 p$K_\mathrm{a}$ unit. These results point out that the suggested protocol can accurately and efficiently predict p$K_\mathrm{a}$'s of aspartate, glutamate and alanine in solution, even when non-optimized geometries are considered.

## Conclusions

In this study, a protocol has been suggested in order to obtain a fast and accurate p$K_\mathrm{a}$ prediction for small carboxylic acids and its applicability to proteins has been tested with three amino acids. According to the suggested protocol, p$K_\mathrm{a}$'s are computed by using the equation derived from the linear regression of the experimental p$K_\mathrm{a}$'s with the atomic charges on the carboxylate fragment. Five charge descriptors, three charge models, three solvent models, gas phase calculations and several DFT methods (combination of eight DFT functionals and fifteen basis sets) were tested. Among those, NPA charge calculations performed with the SMD solvation model on optimized geometries gave the most accurate results. The best combination of DFT functionals and basis sets were found to be M06L/6-311G(d,p) ($R^2$ = 0.955). The strongest linearity is found by selecting the maximum atomic charge on carboxylic oxygen atoms and



(a) Aspartate dipeptide

(b) Glutamate dipeptide

(c) N-Acetyl-Alanine

Figure 7: Deviations of predicted p$K_\mathrm{a}$ with respect to geometrical changes. Geometries were obtained from aqueous phase MD calculations. M06L/6-311G(d,p) method was used for single point NPA calculations using SMD. The red line shows the numerical average of the p$K_\mathrm{a}$ deviations.

relating it to the experimental p$K_\mathrm{a}$. Molecular dynamics simulations have been performed for a set of aspartate, glutamate and alanine peptides in order to test the stability of the prediction. The protocol was applied to a randomly selected set of frames which were extracted from MD simulations and the calculations showed that the predicted p$K_\mathrm{a}$'s were scattered within ±1 unit from the experimental value. The ultimate goal would be to transfer the suggested protocol to the p$K_\mathrm{a}$ prediction of aspartate, glutamate and alanine within a protein environment. By reporting the calculated atomic charge of the carboxylate form into the linear relationship derived in this work, it should be possible to estimate the p$K_\mathrm{a}$'s of aspartate, glutamate and alanine residues inserted in a peptide or a protein sequence.

## Acknowledgement

CPU resources were allocated by the mésocentre EXPLOR of the Université de Lorraine and the Centre de Calcul ROMEO of the Université de Reims Champagne-Ardenne. Z.P.H. acknowledges the French Embassy in Ankara for the Joint PhD fellowship.

## Supporting Information Available

Cartesian coordinates and 2D drawings of the training and test sets of molecules, $R^2$, MAD and MAX-ΔpKa results for the training set for different DFT functionals ( B3LYP, BLYP, M06, M06L, M062X, OLYP, PBE0, and PBE ) and basis sets ( 3-21G, 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) ).

## References

(1) Brunton, L.; Lazo, J.; Parker, K. *Goodman & Gilman's The Pharmacological Basis of Therapeutics*, 11th ed.; Mc.Graw-Hill Medical Pub.: NewYork, 2005.

(2) Forsyth, W. R.; Antosiewicz, J. M.; Robertson, A. D. Empirical relationships between protein structure and carboxyl pKa values in proteins. *Proteins* **2002**, *48*, 388–403.

(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) Harris, T. K.; Turner, G. J. Structural Basis of Perturbed pKa Values of Catalytic Groups in Enzyme Active Sites. *IUBMB Life* **2002**, *53*, 85–98.

(5) Pace, C. N.; Grimsley, G. R.; Scholtz, J. M. Protein ionizable groups: pK values and their contribution to protein stability and solubility. *J. Biol. Chem.* **2009**, *284*, 13285–13289.

(6) Kim, J.; Mao, J.; Gunner, M. R. Are acidic and basic groups in buried proteins predicted to be ionized? *J. Mol. Biol.* **2005**, *348*, 1283–1298.

(7) Ji, C.; Mei, Y.; Zhang, J. Z. Developing polarized protein-specific charges for protein dynamics: MD free energy calculation of pKa shifts for Asp26/Asp20 in thioredoxin. *Biophys. J.* **2008**, *95*, 1080–1088.

(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) Anderson, D. E.; Becktel, W. J.; Dahlquist, F. W. pH-Induced Denaturation of Proteins: A Single Salt Bridge Contributes 3-5 kcal/mol to the Free Energy of Folding of T4 Lysozyme. *Biochemistry* **1990**, *29*, 2403–2408.

(11) Frericks Schmidt, H. L.; Shah, G. J.; Sperling, L. J.; Rienstra, C. M. NMR determination of protein pKa values in the solid state. *J. Phys. Chem. Lett.* **2010**, *1*, 1623–1628.

(12) Oksanen, E.; Chen, J. C.; Fisher, S. Z. Neutron crystallography for the study of hydrogen bonds in macromolecules. *Molecules* **2017**, *22*, 1–26.

(13) Seybold, P. G.; Shields, G. C. Computational estimation of pKa values. *WIREs Comput. Mol. Sci.* **2015**, *5*, 290–297.

(14) Liptak, M. D.; Shields, G. C. Accurate pKa calculations for carboxylic acids using Complete Basis Set and Gaussian-n models combined with CPCM continuum solvation methods. *J. Am. Chem. Soc.* **2001**, *123*, 7314–7319.

(15) Liptak, M. D.; Gross, K. C.; Seybold, P. G.; Feldgus, S.; Shields, G. C. Absolute p$K_a$ Determinations for Substituted Phenols. *J. Am. Chem. Soc.* **2002**, *124*, 6421–6427.

(16) Rebollar-Zepeda, A. M.; Galano, A. First principles calculations of pKa values of amines in aqueous solution: Application to neurotransmitters. *Int. J. Quantum Chem.* **2012**, *112*, 3449–3460.

(17) Thapa, B.; Schlegel, H. B. Calculations of pKa's and redox potentials of nucleobases with explicit waters and polarizable continuum solvation. *J. Phys. Chem. A* **2015**, *119*, 5134–5144.

(18) Casasnovas, R.; Ortega-Castro, J.; Frau, J.; Donoso, J.; Muñoz, F. Theoretical pKa calculations with continuum model solvents, alternative protocols to thermodynamic cycles. *Int. J. Quantum Chem.* **2014**, *114*, 1350–1363.

(19) Jinhua, Z.; Kleinöder, T.; Gasteiger, J. Prediction of pKa values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors. *J. Chem. Inf. Model.* **2006**, *46*, 2256–2266.

(20) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and original pKa prediction method using grid molecular interaction fields. *J. Chem. Inf. Model.* **2007**, *47*, 2172–2181.

(21) Xing, L.; Glen, R. C. Novel Methods for the Prediction of logP, pKa, and logD. *J. Chem. Inform. Comput. Sci.* **2002**, *42*, 796–805.

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

(23) Parthasarathi, R.; Padmanabhan, J.; Elango, M.; Chitra, K.; Subra-manian, V.; Chattaraj, P. K. pKa Prediction Using Group Philicity. *J. Phys. Chem. A* **2006**, *110*, 6540–6544.

(24) Tao, L.; Han, J.; Tao, F. M. Correlations and predictions of carboxylic acid pKa values using intermolecular structure and properties of hydrogen-bonded complexes. *J. Phys. Chem. A* **2008**, *112*, 775–782.

(25) Abkowicz-Bieñko, A. J.; Latajka, Z. Density Functional Study on Phenol Derivative-Ammonia Complexes in the Gas Phase. *J. Phys. Chem. A* **2000**, *104*, 1004–1008.

(26) Caballero-García, G.; Mondragón-Solórzano, G.; Torres-Cadena, R.; Díaz-García, M.; Sandoval-Lira, J.; Barroso-Flores, J. Calculation of Vs,Max and its use as a descriptor for the theoretical calculation of pKa values for carboxylic acids. *Molecules* **2019**, *24*.

(27) Grüber, C.; Buß, V. Quantum-mechanically calculated properties for the development of quantitative structure-activity relationships (QSAR'S). pKA-values of phenols and aromatic and aliphatic carboxylic acids. *Chemosphere* **1989**, *19*, 1595–1609.

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

(29) Clarke, F. H.; Cahoon, N. M. Ionization Constants by Curve Fitting: Determination of Partition and Distribution Coefficients of Acids and Bases and Their Ions. *J. Pharm. Sci.* **1987**, *76*, 611–620.

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

(31) Gross, K. C.; Seybold, P. G.; Hadad, C. M. Comparison of Different Atomic Charge Schemes for Predicting pKa Variations in Substitued Anilines and Phenols. *Int. J. Quantum Chem.* **2002**, *90*, 445–458.

(32) Hollingsworth, C. A.; Seybold, P. G.; Hadad, C. M. Substituent Effects on the Electronic Structure and pKa of Benzoic Acid. *Int. J. Quantum Chem.* **2002**, *90*, 1396–1403.

(33) Citra, M. J. Estimating the pKa of phenols, carboxylic acids and alcohols from semi-empirical quantum chemical methods. *Chemosphere* **1999**, *38*, 191–206.

(34) Svobodová Vařeková, R.; Geidl, S.; Ionescu, C. M.; Skřehota, O.; Kudera, M.; Sehnal, D.; Bouchal, T.; Abagyan, R.; Huber, H. J.; Koča, J. Predicting pKa 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.

(35) Ugur, I.; Marion, A.; Parant, S.; Jensen, J. H.; Monard, G. Rationalization of the pKa values of alcohols and thiols using atomic charge descriptors and its application to the prediction of amino acid pKa's. *J. Chem. Inf. Model.* **2014**, *54*, 2200–2213.

(36) Lide, D. *CRC Handbook of Chemistry and Physics*, 91st ed.; CRS Press, 2009.

(37) 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.

(38) Frisch, M. J. et al. Gaussian 09 Revision B.01. Gaussian Inc.

(39) 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.

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

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

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

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

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

(45) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent 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.

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

(47) 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.

(48) 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.

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

(50) 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.

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

(52) 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.

(53) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural population analysis. *J. Chem. Phys.* **1985**, *83*, 735–746.

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

(55) 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.

(56) Case, D. A. et al. AMBER 2018. University of California: San Francisco, 2018.

(57) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving The Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(58) 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.

(59) 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.

(60) 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.

# Graphical TOC Entry

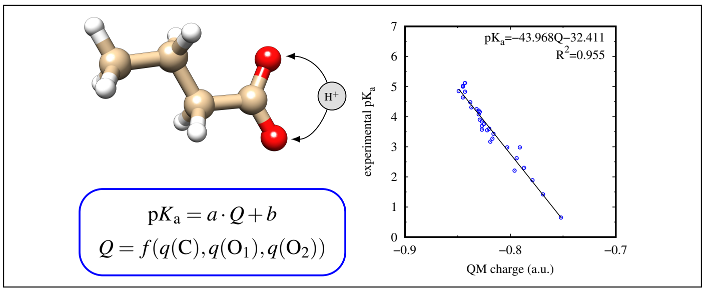


Table 1: Carboxylic Acid Training Set: CAS Number, Molecule Name, Experimental p$K_\mathrm{a}$, Predicted p$K_\mathrm{a}$, and Differences between Experimental and Predicted p$K_\mathrm{a}$ values

| CAS Number | Molecule Name | p$K_\mathrm{a}$ (exp.) | p$K_\mathrm{a}$ (pred.) [a] | $\Delta$ p$K_\mathrm{a}$ |
|---|---|---:|---:|---:|
| 129-66-8 | 2,4,6-Trinitrobenzoic acid | 0.65[36] | 0.63 | -0.02 |
| 610-30-0 | 2,4-Dinitrobenzoic acid | 1.42[37] | 1.39 | -0.03 |
| 471-25-0 | Propiolic acid | 1.89[37] | 1.85 | -0.04 |
| 552-16-9 | 2-Nitrobenzoic acid | 2.21[37] | 2.60 | 0.39 |
| 1460-34-0 | $\alpha$-Keto-$\beta$-methylvaleric acid | 2.30[37] | 2.18 | -0.12 |
| 590-93-2 | 2-Butynoic acid | 2.62[36] | 2.52 | -0.10 |
| 298-12-4 | 2-Oxoacetic acid | 2.98[37] | 2.38 | -0.60 |
| 69-72-7 | 2-Hydroxybenzoic acid | 2.98[36] | 2.89 | -0.09 |
| 122-59-8 | Phenoxyacetic acid | 3.17[36] | 3.62 | 0.45 |
| 88-14-2 | 2-Furoic acid | 3.27[37] | 3.52 | 0.25 |
| 62-23-7 | 4-Nitrobenzoic acid | 3.43[36] | 3.45 | 0.02 |
| 480-63-7 | 2,4,6-Trimethylbenzoic acid | 3.55[37] | 3.74 | 0.19 |
| 625-45-6 | Methoxyacetic acid | 3.57[37] | 3.93 | 0.36 |
| 1877-72-1 | 3-Cyanobenzoic acid | 3.60[36] | 3.66 | 0.06 |
| 33445-07-7 | Isopropoxyacetic acid | 3.69[37] | 3.97 | 0.28 |
| 64-18-6 | Formic acid | 3.76[37] | 3.86 | 0.10 |
| 627-03-2 | Ethoxyacetic acid | 3.84[37] | 3.96 | 0.12 |
| 488-93-7 | 3-Furoic acid | 3.90[36] | 4.05 | 0.15 |
| 99-06-9 | 3-Hydroxybenzoic acid | 4.08[36] | 4.07 | -0.01 |
| 93-09-4 | 2-Naphtoic acid | 4.16[36] | 4.04 | -0.12 |
| 190965-42-5 | 3-Propoxybenzoic acid | 4.20[37] | 4.08 | -0.12 |
| 99-04-7 | 3-Methylbenzoic acid | 4.25[36] | 4.18 | -0.07 |
| 103-82-2 | Phenylacetic acid | 4.31[36] | 4.37 | 0.06 |
| 99-50-3 | 3,4-Dihydroxybenzoic acid | 4.48[36] | 4.41 | -0.07 |
| 79-31-2 | Isobutyric acid | 4.64[37] | 4.74 | 0.10 |
| 1759-53-1 | Cyclopropanecarboxylic acid | 4.83[36] | 4.64 | -0.19 |
| 142-62-1 | Hexanoic acid | 4.85[36] | 4.94 | 0.09 |
| 6202-94-4 | trans-2-Methylcyclopropanecarboxylic acid | 5.00[37] | 4.75 | -0.25 |
| 6142-57-0 | cis-2-Methylcyclopropanecarboxylic acid | 5.02[37] | 4.72 | -0.30 |
| 541-47-9 | 3-Methyl-2-butenoic acid | 5.12[37] | 4.64 | -0.48 |

[a] p$K_\mathrm{a}$ values are computed for each molecule on the anionic form, optimized with M06L/6-311G(d,p) and SMD, using the highest NPA atomic charge of the two oxygen atoms of the carboxylate fragment (see text).

Table 2: Monocarboxylic Acid Test Set: CAS Number, Molecule Name, Experimental p$K_a$, Predicted p$K_a$, and Differences between Experimental and Predicted p$K_a$ values

| CAS Number | Molecule Name | p$K_a$ (exp.) | p$K_a$ (pred.) [a] | $\Delta$ p$K_a$ |
|---|---|---|---|---|
| 625-75-2 | Nitroacetic acid | 1.48[36] | 2.19 | 0.71 |
| 372-09-8 | Cyanoacetic acid | 2.47[36] | 3.34 | 0.87 |
| 127-17-3 | Pyruvic acid | 2.60[37] | 2.59 | -0.01 |
| 5699-58-1 | Acetopyruvic acid | 2.61[37] | 2.11 | -0.50 |
| 121-92-6 | 3-Nitrobenzoic acid | 3.46[36] | 3.57 | 0.11 |
| 619-65-8 | 4-Cyanobenzoic acid | 3.55[36] | 3.62 | 0.07 |
| 2516-93-0 | Butoxyacetic acid | 3.66[37] | 3.96 | 0.30 |
| 54497-00-6 | Propoxyacetic acid | 3.69[37] | 3.94 | 0.25 |
| 50-21-5 | 2-Hydroxypropanoic acid | 3.83[37] | 3.95 | 0.12 |
| 79-14-1 | Hydroxyacetic acid | 3.83[36] | 3.90 | 0.07 |
| 118-90-1 | 2-Methylbenzoic acid | 3.98[37] | 3.79 | -0.19 |
| 586-38-9 | 3-Methoxybenzoic acid | 4.09[37] | 4.07 | -0.02 |
| 65-85-0 | Benzoic acid | 4.19[37] | 4.12 | -0.07 |
| 2529-39-7 | 2,3,4,5-Tetramethylbenzoic acid | 4.22[37] | 4.06 | -0.16 |
| 86-55-5 | 1-Naphtoic acid | 4.24[37] | 3.71 | -0.53 |
| 79-10-7 | Acrylic acid | 4.25[36] | 3.95 | -0.30 |
| 1077-07-2 | 3-Allylbenzoic acid | 4.32[37] | 4.12 | -0.20 |
| 99-94-5 | 4-Methylbenzoic acid | 4.37[36] | 4.27 | -0.10 |
| 5438-19-7 | 4-Propoxybenzoic acid | 4.46[37] | 4.43 | -0.03 |
| 100-09-4 | 4-Methoxybenzoic acid | 4.50[36] | 4.42 | -0.08 |
| 1498-96-0 | 4-Butoxybenzoic acid | 4.53[37] | 4.43 | -0.10 |
| 99-96-7 | 4-Hydroxybenzoic acid | 4.58[37] | 4.45 | -0.13 |
| 64-19-7 | Acetic acid | 4.76[37] | 4.81 | 0.05 |
| 107-92-6 | Butyric acid | 4.82[37] | 4.90 | 0.08 |
| 109-52-4 | Pentanoic acid | 4.86[37] | 4.93 | 0.07 |
| 79-09-4 | Propanoic acid | 4.87[36] | 4.87 | -0.00 |
| 98-89-5 | Cyclohexanecarboxylic acid | 4.90[37] | 5.05 | 0.15 |
| 3400-45-1 | Cyclopentanecarboxylic acid | 4.99[36] | 4.93 | -0.06 |
| 75-98-9 | Trimethylacetic acid | 5.05[37] | 4.63 | -0.42 |

[a] p$K_a$ values are computed for each molecule on the anionic form, optimized with M06L/6-311G(d,p) and SMD, using the highest NPA atomic charge of the two oxygen atoms of the carboxylate fragment (see text).