



Development and assessment of a ChemInformatics model for accurate pK_a prediction in aqueous medium

Alec J. Sanchez¹ · Krishnan Raghavachari¹

Received: 15 April 2023 / Accepted: 13 July 2023 / Published online: 18 August 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

The accurate prediction of the acid dissociation constants (pK_a) of organic and drug molecules is known to be a challenging problem in computational quantum chemistry. Specifically, density functional theory-based predictions suffer from a high dependence on the nature of the functional group as well as the underlying exchange–correlation functional. Additionally, the introduction of explicit solvent molecules is known to be important for accurate prediction of the pK_a values for many functional groups in water, making it a particularly challenging problem. The inclusion of only implicit solvation effects, though highly efficient, is often inadequate for the prediction of pK_a s. In this paper, we have considered a data set of 303 molecules containing 13 different functional groups to assess the predictability of DFT for the calculation of pK_a s. Using just implicit solvation models with DFT, each functional group shows a linear correlation with experiment, though with different slopes for different functional groups. Using simple linear regression-based corrections for systematic errors of different functional groups, we show that DFT including implicit solvation can be used to make reliable predictions of pK_a s with a mean absolute deviation of only 0.397 pK_a units. For a test set of 100 larger and more complex drug molecules, the performance of our model is very good, though with a slightly larger mean absolute deviation of 0.629 pK_a units. More importantly, our pK_a protocol is general and applicable to any underlying density functional, making it an effective computational tool for pK_a predictions.

1 Introduction

Accurately predicting the acid dissociation constant (pK_a) for organic and bio-organic molecules containing different functional groups in solution has been an ongoing challenge in computational quantum chemistry [1–5]. Many of the popular protocols involving highly correlated levels of theory for calculating pK_a are limited in applicability due to the accuracy–cost trade-off inherent to all quantum mechanical methods, as well as the added complexity of accurately modeling solvation effects [2, 5–9]. To overcome this bottleneck, approximate methods, such as density functional theory (DFT) paired with an implicit solvation model (e.g., polarizable continuum model (PCM), COSMO), are often the most practical options to represent the solute–solvent interactions [4, 5, 10–15]. While these tools can be readily used to estimate pK_a s of complex systems, such implicit

solvation models fail to give an adequate description of the chemical environment of the functional group being (de) protonated since they ignore explicit interactions with the solvent (typically, H_2O) via hydrogen bonding [10]. Recent reports from the Raghavachari group have introduced a more feasible pK_a calculation protocol which uses the connectivity-based hierarchy (CBH) in conjunction with a recommended number of explicit water molecules depending on the functional group [16]. Although this protocol has been used to achieve chemical accuracy for a set of bio-organic molecules, modeling solvent interactions explicitly is much more computationally demanding and may not be practical to study larger biomolecular systems [16]. Furthermore, state-of-the-art pK_a calculations with both implicit and explicit solvation models suggest that the significance of explicit solvation is not uniform among different molecular systems [6, 10, 17–19]. Thus, a broadly applicable procedure, not requiring explicit solvation, would greatly benefit the field and remove the need to analyze individual systems. Herein, we propose an alternative method for pK_a prediction to circumvent the need for explicit solvation entirely. In our method, we exploit the local nature of the acid dissociation

✉ Krishnan Raghavachari
kraghava@indiana.edu

¹ Department of Chemistry, Indiana University, Bloomington, IN, USA

constant to perform systematic error corrections via a simple chemical informatics-based linear regression model to achieve high accuracy. Our initial goal was to make reliable predictions with a target accuracy of < 1 pK_a unit for a wide variety of functional groups, though, as demonstrated below, we achieve much higher accuracy.

2 Background

The negative logarithm of the acid dissociation constant (pK_a) plays an essential role in chemical and biological processes related to solvation, protein–ligand binding, and protein structure [1, 5, 20–22]. In drug design, the physicochemical properties that are screened for in ADME (absorption, distribution, metabolism, and excretion) are directly correlated with the protonated and deprotonated forms of the molecule [1, 23, 24]. In materials science, the charge state of a molecule can influence properties of nanomaterials such as dispersibility, catalysis, 3-D structure, and the tautomeric form [25]. Since the equilibrium of the protonated and deprotonated states is dependent on the change in Gibbs free energy (ΔG), pK_a can be described via a well-defined thermodynamic process [5] [26].

Quantum mechanical methods including higher-order electron correlation effects can typically be used to predict the acid dissociation constant of small molecules [16, 27, 28, 28, 29]. Due to the steep computational scaling, the applicability of these methods toward larger molecules (i.e., drug molecules) may not be practical, particularly when a large number of molecules have to be screened. One example of a popular QM-based ab initio program is Jaguar pK_a , which utilizes linear fit equations for 1 K molecules along with functional group-specific parameters to predict the pK_a [30]. Although the performance is strong, the shell model that is used for pK_a prediction is part of a commercial software package that is not openly available and may require frequent updates involving the latest literature data [30, 31]. QSAR (quantitative structure–activity relationship) algorithms are also commonly used and are typically faster and more accurate compared to ab initio predictors for common functional groups [32, 33]. Unfortunately, many of these algorithms perform poorly for functional groups that are not well represented, or with molecules (typically found in materials science) containing multiple conformations or resonance forms. Finally, there has been a growing interest and excitement toward artificial intelligence, specifically machine learning (ML), which stems from the low computational cost, and ability to model complex real-world problems [32, 34–37].

In this manuscript, we develop and compare 3 simple chemical informatics-based models that overcome the

accuracy–cost trade-off inherent to all quantum mechanical methods. In Model 1, raw pK_a s are calculated using density functional theory (DFT) and implicit solvation. For Model 2, the systematic error associated with DFT in the raw pK_a s is corrected using a single linear regression over the entire data set. In Model 3, we exploit the local nature of pK_a s through functional group-specific linear fits that are applicable for large drug molecules. In future work, we plan to represent the local nature of pK_a s as molecular descriptors for machine learning models and assess the performance against Model 3.

3 Computational models

3.1 Raw pK_a evaluation procedure

3.1.1 Model 1

Calculating the raw pK_a of the deprotonation reaction, e.g., $AH \leftrightarrow A^- + H^+$, can be described using the following protocol,

$$\text{Raw } pK_a = \frac{\Delta G_{aq}^*}{2.303RT} \quad (1)$$

where ΔG_{aq}^* is the aqueous free energy change for the deprotonation reaction, R is the molar gas constant, and T is the temperature (298.15 K). ΔG_{aq}^* can be calculated as

$$\Delta G_{aq}^* = G_{A^-,aq}^* + G_{H^+,aq}^* - G_{AH,aq}^* \quad (2)$$

where $G_{A^-,aq}^*$ and $G_{AH,aq}^*$ are the free energies associated with the deprotonated (A^-) and protonated (AH) species in aqueous phase using SMD [38] (solvation model based on density) implicit solvation. $G_{H^+,aq}^*$ is the free energy of a proton in aqueous phase and is given as,

$$G_{H^+,aq}^* = G_{H^+,gas}^o + \Delta G_{H^+,solv}^* + \Delta G^{1atm \rightarrow 1M} \quad (3)$$

where $\Delta G_{H^+,solv}^* = -265.9$ [39–42] kcal/mol is the change in free energy of a solvated proton, $\Delta G^{1atm \rightarrow 1M} = 1.89$ kcal/mol is the change in free energy associated with converting from 1 atm in the standard state to 1 molarity in aqueous media, and $G_{H^+,gas}^o = H_{gas}^o - TS_{gas}^o$ is the free energy of a proton in the gas phase. $H_{gas}^o = \left(\frac{5}{2}\right)RT$ is the enthalpic contribution of hydrogen gas while $S_{gas}^o = 26.05$ cal/(mol • K) is the entropic contribution of hydrogen gas. As recommended from previous studies [13, 43, 44], a thermodynamic cycle was not used to calculate ΔG_{aq}^* .

3.2 Systematic error correction associated with DFT

3.2.1 Models 2 and 3

From the linear regression fit of the entire pK_a set in our model, the loss function was minimized to correct for the systematic errors. This can be calculated as

$$Y_I = m(\text{raw } pK_a) + b \quad (4)$$

where Y_I is the corrected calculated pK_a , m the slope, and b the y-intercept of the linear regression equations. Given the corrected calculated pK_a , the mean absolute deviation (MAD) can be calculated as

$$MAD = \frac{\sum_{i=1}^n |Y_I - Y|}{n} \text{ where } |Y_I - Y| = \text{absolute deviation, } n = \text{number of } pK_a\text{'s}$$

In Model 2, a single regression was used for the entire set of molecules. In Model 3, the linear regression was carried out for individual functional groups to correct for the systematic errors for each functional group (vide infra).

3.3 Computational details

Geometry optimizations for each molecule in the model and test set were obtained with the B3LYP [45–47] functional, 6–311++G(d,p) [48–52] basis set, and SMD [4] universal solvation model for implicit solvation. To explore the dependence of the computed results on the density functional used, the deprotonation free energy was obtained using the following levels of theory: B3LYP/6–311++G(d,p), B3LYP-D3(BJ)/6–311++G(d,p), and ω B97X-D

[53]/6–311++G(d,p). The treatment of solvent was done implicitly through the SMD continuum solvation model. In B3LYP-D3(BJ), Grimme's empirical dispersion model, D3(BJ) [54, 55], was used in conjunction with the B3LYP functional. In addition, all three density functionals were used with the 6–31G(d) basis set to explore whether a much smaller basis set is adequate for pK_a studies, but the results were substantially worse, and will not be discussed further. The deficiency of the 6–31G(d) basis is likely due to the absence of diffuse functions that are known to be important for the treatment of anions. All computational work was performed using the Gaussian 16 program suite [56].

3.4 Training set

To encompass a wide chemical space, 13 functional groups with a total of 303 molecules and 330 raw pK_a s were used as the framework for our chemical informatics model (Table 1). Not only do functional groups identify the regions of a molecule where the chemical reaction occurs, but also can be used as a general descriptor for a set of structurally similar set of molecules (e.g., amino acids). Out of the 13 functional groups, tertiary amine, secondary amine, aliphatic alcohol, aromatic alcohol, and carboxylic acid were cited in a list of 10 most frequent functional groups in bioactive molecules found in the medicinal chemistry literature [57]. In addition, since chloro and fluoro groups were also frequently present in medicinal chemistry, molecules that contained either of them, regardless of the deprotonation site, were used to create separate linear fits to explore their behavior (SI Fig. 1).

Table 1 List of 13 functional groups and mean absolute error (MAE) associated with calculated and experimental pK_a at the B3LYP/6–311++G(d,p) level of theory with implicit solvation

Functional group	Number of molecules	Number of pK_a	MAE $ \Delta pK_a $
Nitrogen containing	24	24	0.717
Aromatic groups			
Aliphatic alcohols	20	20	7.842
Aliphatic thiols	20	20	8.120
Primary amines	21	21	0.714
Secondary amines	16	16	0.568
Tertiary amines	14	14	0.883
Carboxylic acids	38	38	1.831
Thiophenols	13	13	4.933
Phenols	41	41	3.680
Anilines	36	36	3.733
Benzoic acids	26	26	1.757
Carbon acids	14	14	4.171
Amino acids	20	47	2.000 ^a , 2.008(COOH), 1.744(NH3+), 2.728(R-group)
Total	303	330	3.000

^aTotal MAE for functional group

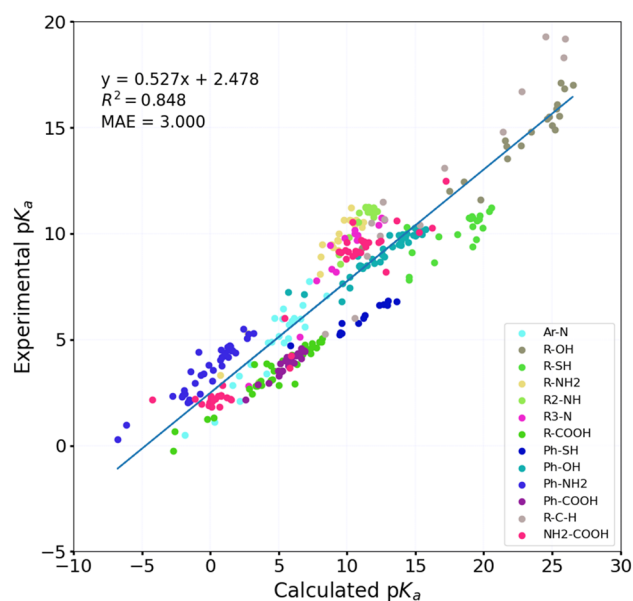


Fig. 1 Relationship between experimental and calculated pK_a for 303 molecules (330 total pK_a s) calculated at B3LYP/ 6-311++G(d,p) level of theory with SMD solvation. Set of molecules are found in SI Table 1

However, these two fits were separate from any of the final models to avoid introducing duplicate data.

4 Results and discussion

Each model utilizes computationally inexpensive DFT methods to calculate the pK_a values. Of the three functionals tested, B3LYP/6-311++G(d,p) was slightly better than the other two for the raw errors and was chosen to illustrate the performance for the remainder of this study. However, the conclusions are similar for all three functionals.

4.1 Model 1

In Model 1, a DFT calculation (B3LYP/6-311++G(d,p)) with implicit solvation was performed on 303 molecules for 330 pK_a s and when compared to experimental values, had a large MAE of 3.000 pK_a units. Since pK_a is measured on the logarithmic scale, the chemical insight provided by Model 1 is minimal.

4.2 Model 2

The same DFT-calculated pK_a s were then fit to the corresponding experimental values using a standard linear regression, shown in Fig. 1.

The observed linear regression is then used to derive Model 2. Each pK_a was plugged into the linear fit equation

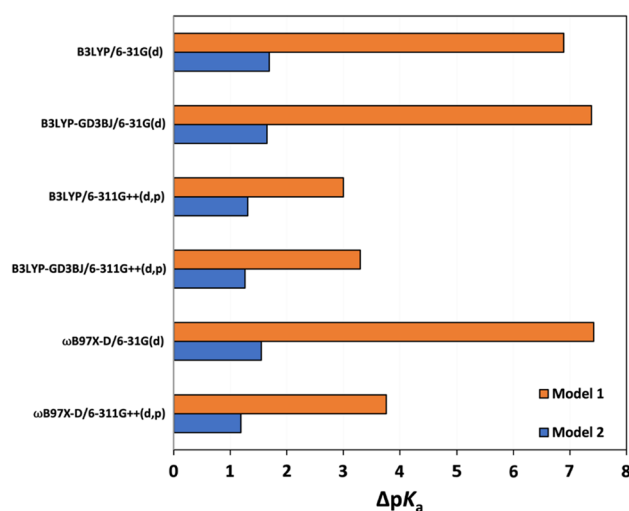


Fig. 2 Performance of Model 1 and Model 2 for each functional/basis set

from Fig. 1 to correct for the systematic error associated with DFT. This resulted in a reduction of error from 3.000 to 1.281 pK_a units going from Model 1 to Model 2. The results from Model 1 and Model 2 are shown in Fig. 2.

While Model 2 is a significant improvement compared to Model 1 (uncorrected DFT-calculated pK_a), it is still not within the target accuracy of <1 pK_a unit. For several of the functional groups, e.g., aliphatic thiol, primary amine, secondary amine, and thiophenol, the MAE of Model 2 was quite far from target accuracy (over 2 pK_a units) though the coefficient of determination $R^2 > 0.92$ was quite high for most of the groups. This indicates that QM calculated deprotonation energies are systematic for each functional group. Furthermore, for primary, secondary, and tertiary amines, the MAE was worse using Model 2 than Model 1, indicating that a global correction does not accurately represent all functional groups in the data set, leading to Model 3.

4.3 Model 3

For Model 3, the full set was divided into functional groups and separate linear fit equations were used to correct for the systematic errors associated with each group (Fig. 3).

Thus, in Model 3, the DFT-calculated pK_a s of each functional group were fitted separately to the corresponding experimental pK_a s. This correction lowered the error for the full set of molecules from 1.281 to 0.396 pK_a units, which is well within our target accuracy. The results for all three models are shown for the individual functional groups are shown in Fig. 4.

As shown in Fig. 4, aliphatic thiols had the largest drop in error (8.120 to 0.476 pK_a units), while aliphatic alcohols were the second most improved group with a reduction from

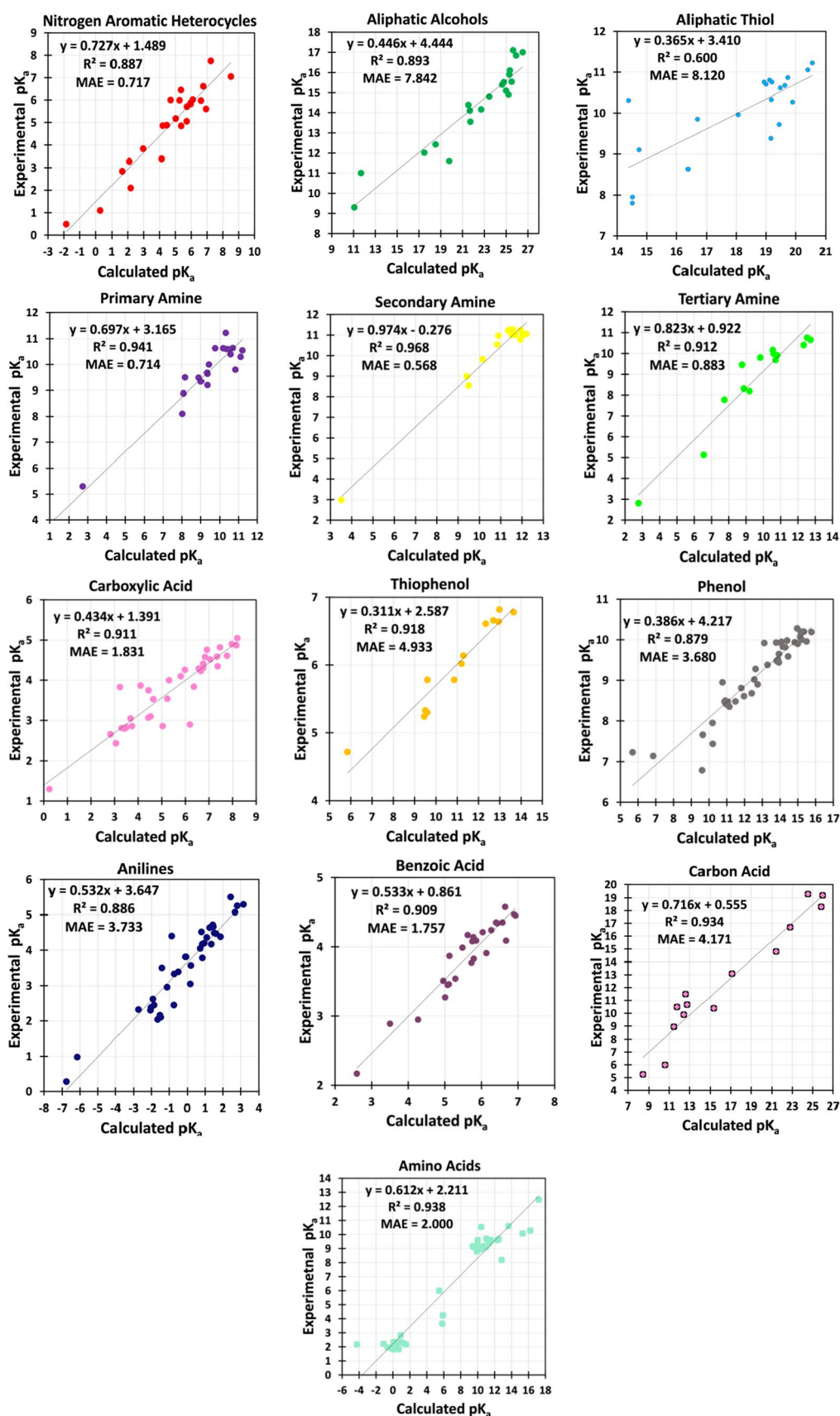
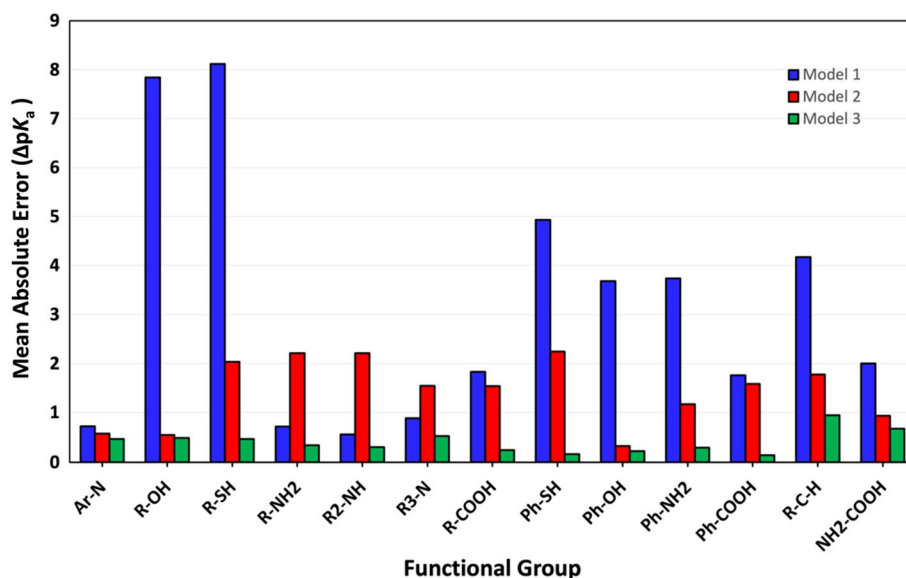


Fig. 3 Functional group-specific linear fits used to correct for systematic errors

Fig. 4 Effect of the model used on the absolute difference between the calculated and experimental pK_a for each functional group. On x-axis, Ar-N corresponds to nitrogen containing aromatic, R-OH is primary alcohol, R-SH is thiol, R-NH2 is primary amine, etc



7.842 to 0.498. The smallest improvement was with nitrogen containing heterocycle (0.717 to 0.475) and secondary amines (0.568 to 0.311) since they were already quite accurate compared to experimental values. Similar errors were obtained in a different study [16] which utilized an optimal number of water molecules for 12 of the functional groups studied. However, our work suggests that explicit solvation may not be required to derive chemically accurate pK_a s, but can be obtained instead through systematic error correction. There were a few groups (primary, secondary, tertiary amines) that when used in Model 2, MAE increased compared to Model 1, clearly demonstrating that a single linear regression for all systems is inadequate for cases where the raw performance in Model 1 is fortuitously very good. As expected, Model 3 with individual linear regressions performs very well on these systems with accuracy within 0.5 pK_a units. The largest errors in Model 3 can be seen in more complex groups such as carbon acids which cover a large range of pK_a s (14 units) and amino acids which can have a doubly charged or zwitterionic form. Nevertheless, after the Model 3 correction, the MAE decreased for all 13 functional groups and resulted in all of them falling well within the target accuracy of 1 pK_a unit (Fig. 4). As mentioned earlier, Model 3 corrections for chloro- and fluoro-containing molecules were also separately fitted (SI Fig. 1) to explore their behavior, regardless of the site of deprotonation. They both show a good linear trend ($R^2 > 0.80$) and reasonable performance (0.921 and 1.022 pK_a unit MAD, respectively), suggesting that perturbations from highly electron withdrawing groups can also be systematically corrected and may play a role in pK_a calculations. Nevertheless, separating the molecules into their corresponding functional group-specific fit, as we have done

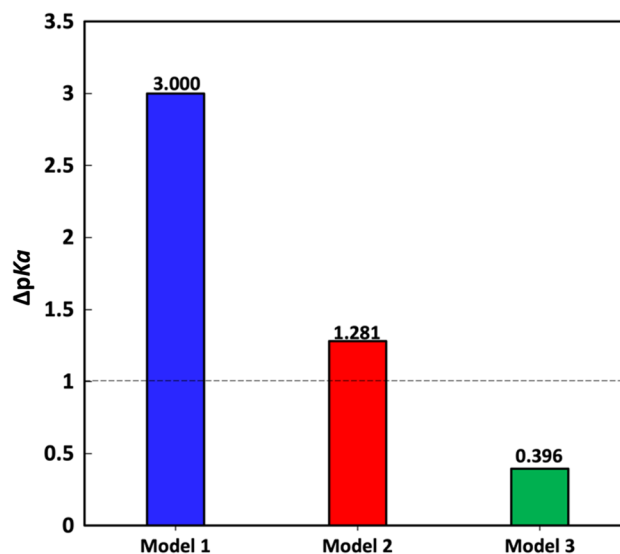
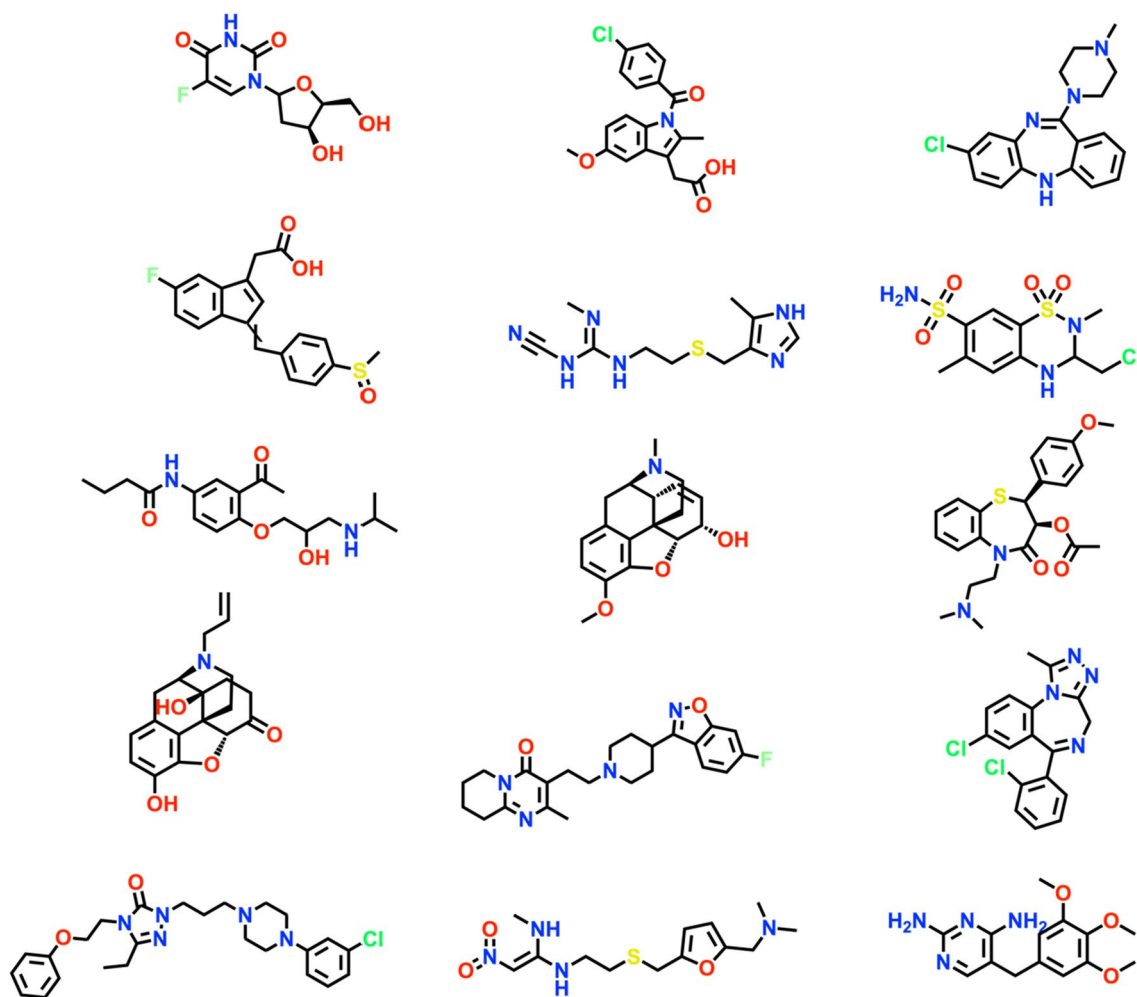


Fig. 5 Error between calculated and experimental pK_a for entire 303 molecule set. Model 1 is the raw pK_a , Model 2 includes systematic error correction on raw pK_a using linear fit for all molecules (Fig. 2). Model 3 includes functional group-specific linear fit equations to correct for the systematic error (Figs. 3, 4). Dashed line indicates target accuracy < 1 pK_a unit

in this work, is still most important and performs best. For simplicity and to avoid introducing duplicate molecules into the training set, chloro- and fluoro-specific fits were not used in the final models.

Figure 5 summarizes the performance of all three models for the training set used.



Scheme 1 Sample of drug molecules in test set (15 out of 100)

5 Rigorous assessment of model 3 for more complex systems

To gauge the chemical span and applicability of our model, a set of 100 drug molecules with multiple functional groups were randomly chosen. No constraints were made on the characteristics (i.e., size, stereochemistry) of the drug molecules studied, and a sample from the test set of molecules can be found in Scheme 1.

With the same framework as before, the B3LYP/6–311++G(d,p) functional and basis set were used in conjunction with implicit solvation (SMD) to obtain QM calculated deprotonation energies that can be used to determine the acid dissociation constant. Since literature values present only one pK_a , the drug molecules in our study were deprotonated individually at all functional group sites. While it may seem intuitive to compare the closest calculated raw pK_a to the experimental value, as we saw earlier, because of

the systematic error inherent for each functional group, one must correct each pK_a using the functional group-specific linear fit equations before comparing with the literature values. The corrected values closest to the experimental pK_a were used for comparison. In most cases, the assignments were unambiguous.

A list of the drug molecules along with their associated raw pK_a s can be found in the supplemental information (SI Table 2). After obtaining the correct raw pK_a , each drug molecule was then corrected using the corresponding functional group-specific linear fit (Model 3) to obtain the new predicted pK_a . As expected, the correction dropped the overall error for all 100 drug molecules from 1.381 to 0.629 pK_a units, highlighting the predictive strength of Model 3 on large molecules with multiple functional groups. The final error is slightly larger than the value for the training set, but not entirely surprising since the drug molecules are both larger and more complex. However, it was somewhat

surprising that the uncorrected (raw) MAE was only 1.381 pK_a units. While the drug molecules were arbitrarily chosen, many of them were basic, which may be one possible cause for the unexpectedly low uncorrected MAE of 1.381. In addition, the functional groups with the largest raw errors were less represented in this test set. To get a more complete gauge of the Model 3 performance, a broader distribution of acidic and basic drugs may be needed. It is also worth noting that in the test set, there was one molecule (triazolam) which had a deviation of > 3 pK_a units. The molecule contained a triazole group attached to a 7-membered ring with phenyl and chloro groups. For this rare case, very different from any of the molecules in the training set, the protonated amine was unable to be fully represented by the functional group-specific fit. In a separate instance, there was a drug molecule (methadone) which had a poor initial structure and optimized to a geometry that had a phenyl group hovering directly over a protonated amine. The electrostatic attraction between the two groups resulted in an overestimated pK_a and required a rotation of the amine away from the ring to reach a more stable conformation that performed better. A conformer search should be used in future work to avoid such issues. Nevertheless, the overall performance of Model 3 is well within the target accuracy of 1 pK_a unit even for such complex systems.

One of the major advantages with Model 3 is that the computational cost for creating and testing the model was very inexpensive, making it feasible to quickly expand the chemical space. To improve Model 3, it would be beneficial to include a conformational search for molecules with many rotatable bonds, ensuring that the optimized structure is described by the local minimum on the potential energy surface. Developing an algorithm that can choose the correct micro- pK_a would also be a huge improvement and decrease the chance of assigning the wrong deprotonation site. Furthermore, it should be interesting to see whether we could formulate a similar model or correction term that could predict the pK_a in different solvents (e.g., DMF).

6 Conclusions

To summarize, we derived an accurate predictive model, Model 3, by starting from an inexpensive method such as DFT with an implicit solvation model and taking advantage of the locality of the acid dissociation constant. Using simple linear regression-based corrections for systematic errors of different functional groups, Model 3 was able to obtain pK_a s with a mean absolute deviation of only 0.397 pK_a units. For a test set of 100 larger and more complex drug molecules, the performance of our model is still very good, though with a slightly larger mean absolute deviation of 0.629 pK_a units.

More importantly, our pK_a protocol is general and applicable to any underlying density functional, making it an effective computational tool for pK_a predictions.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00214-023-03024-6>.

Acknowledgements We acknowledge financial support from the National Science Foundation Grant CHE-2102583 at Indiana University. The Big Red 3 supercomputing facility at Indiana University was used for most of the calculations in this study.

Author contributions A.J.S. carried out the research project under the supervision of Prof. K.R. The initial draft of the manuscript was prepared by A.J.S. and was reviewed and revised by K.R.

Declarations

Conflict of interest The authors declare no competing financial interest.

References

1. Manallack DT (2007) The PK(a) distribution of drugs: application to drug discovery. *Perspect Med Chem* 1:25–38
2. Alongi KS, Shields GC (2010) Theoretical calculations of acid dissociation constants: a review article. *Annual Rep Comput Chem* 6:113–138. [https://doi.org/10.1016/S1574-1400\(10\)06008-1](https://doi.org/10.1016/S1574-1400(10)06008-1)
3. Liao C, Nicklaus MC (2009) Comparison of nine programs predicting pK_a values of pharmaceutical substances. *J Chem Inf Model* 49(12):2801–2812. <https://doi.org/10.1021/ci900289x>
4. Ho J, Coote ML (2010) A universal approach for continuum solvent PK a calculations: are we there yet? *Theor Chem Acc* 125(1–2):3–21. <https://doi.org/10.1007/s00214-009-0667-0>
5. Fujiki R, Matsui T, Shigeta Y, Nakano H, Yoshida N (2021) Recent developments of computational methods for p K a prediction based on electronic structure theory with solvation models. *J* 4(4): 849–64. <https://doi.org/10.3390/j4040058>
6. Zhang S (2012) A reliable and efficient first principles-based method for predicting pK_a values. 4. Organic bases. *J Comput Chem* 33(31):2469–2482. <https://doi.org/10.1002/jcc.23068>
7. Zhang S, Baker J, Pulay P (2010) A reliable and efficient first principles-based method for predicting pK_a values. 2. Organic acids. *J Phys Chem A* 114(1):432–442. <https://doi.org/10.1021/jp9067087>
8. Shields GC, Seybold PG (2014) Computational approaches for the prediction of PKa values: QSAR in environmental and health sciences; CRC press. Taylor & Francis Group, Boca Raton
9. Mangold M, Rolland L, Costanzo F, Sprik M, Sulpizi M, Blumberger J (2011) Absolute pK_a values and solvation structure of amino acids from density functional based molecular dynamics simulation. *J Chem Theory Comput* 7(6):1951–1961. <https://doi.org/10.1021/ct100715x>
10. Ho J (2014) Predicting PKa in implicit solvents: current status and future directions. *Aust J Chem* 67(10):1441–1460
11. Klamt A (2011) The COSMO and COSMO-RS solvation models. *WIREs Comput Mol Sci* 1(5):699–709. <https://doi.org/10.1002/wcms.56>
12. Klamt A, Eckert F, Diedenhofen M, Beck ME (2003) First principles calculations of aqueous pK_a values for organic and inorganic acids Using COSMO–RS reveal an inconsistency in the slope of

- the pK_a scale. *J Phys Chem A* 107(44):9380–9386. <https://doi.org/10.1021/jp034688o>
13. Ho J, Ertem MZ (2016) Calculating free energy changes in continuum solvation models. *J Phys Chem B* 120(7):1319–1329. <https://doi.org/10.1021/acs.jpcc.6b00164>
 14. Eckert F, Klamt A (2006) Accurate prediction of basicity in aqueous solution with COSMO-RS. *J Comput Chem* 27(1):11–19. <https://doi.org/10.1002/jcc.20309>
 15. Eckert F, Diedenhofen M, Klamt A (2010) Towards a first principles prediction of pK_a : COSMO-RS and the cluster-continuum approach. *Mol Phys* 108(3–4):229–241. <https://doi.org/10.1080/00268970903313667>
 16. Thapa B, Raghavachari K (2019) Accurate PKa evaluations for complex bio-organic molecules in aqueous media. *J Chem Theory Comput* 15(11):6025–6035. <https://doi.org/10.1021/acs.jctc.9b00606>
 17. Kelly CP, Cramer CJ, Truhlar DG (2006) Adding explicit solvent molecules to continuum solvent calculations for the calculation of aqueous acid dissociation constants. *J Phys Chem A* 110(7):2493–2499
 18. Pliego JR, Riveros JM (2002) Theoretical calculation of pK_a using the cluster–continuum model. *J Phys Chem A* 106(32):7434–7439. <https://doi.org/10.1021/jp025928n>
 19. Adam KR (2002) New density functional and atoms in molecules method of computing relative pK_a values in solution. *J Phys Chem A* 106(49):11963–11972. <https://doi.org/10.1021/jp026577f>
 20. Charifson PS, Walters WP (2014) Acidic and basic drugs in medicinal chemistry: a perspective. *J Med Chem* 57(23):9701–9717. <https://doi.org/10.1021/jm501000a>
 21. Bell RP (2013) The proton in chemistry. Springer Science & Business Media, USA
 22. Stewart R (2012) The proton: applications to organic chemistry. Elsevier, USA
 23. Comer J, Box K (2003) High-throughput measurement of drug PKa values for ADME screening. *JALA J Assoc Lab Autom* 8(1):55–59. <https://doi.org/10.1016/S1535-5535-04-00243-6>
 24. Cruciani G, Milletti F, Storch L, Sforza G, Goracci L (2009) *In Silico* pK_a prediction and ADME profiling. *Chem Biodivers* 6(11):1812–1821. <https://doi.org/10.1002/cbdv.200900153>
 25. Orth ES, Ferreira JGL, Fonsaca JES, Blaskiewicz SF, Domingues SH, Dasgupta A, Terrones M, Zarbin AJG (2016) PKa determination of graphene-like materials: validating chemical functionalization. *J Colloid Interface Sci* 467:239–244. <https://doi.org/10.1016/j.jcis.2016.01.013>
 26. Pliego JR (2003) Thermodynamic cycles and the calculation of PKa. *Chem Phys Lett* 367(1–2):145–149. [https://doi.org/10.1016/S0009-2614\(02\)01686-X](https://doi.org/10.1016/S0009-2614(02)01686-X)
 27. Liptak MD, Gross KC, Seybold PG, Feldgus S, Shields GC (2002) Absolute pK_a determinations for substituted phenols. *J Am Chem Soc* 124(22):6421–6427
 28. Liptak MD, Shields GC (2001) Accurate pK_a calculations for carboxylic acids using complete basis set and Gaussian- n models combined with CPCM continuum solvation methods. *J Am Chem Soc* 123(30):7314–7319
 29. Klicic JJ, Friesner RA, Liu S-Y, Guida WC (2002) Accurate prediction of acidity constants in aqueous solution via density functional theory and self-consistent reaction field methods. *J Phys Chem A* 106(7):1327–1335. <https://doi.org/10.1021/jp012533f>
 30. Bochevarov AD, Harder E, Hughes TF, Greenwood JR, Braden DA, Philipp DM, Rinaldo D, Halls MD, Zhang J, Friesner RA (2013) Jaguar: a high-performance quantum chemistry software program with strengths in life and materials sciences. *Int J Quantum Chem* 113(18):2110–2142. <https://doi.org/10.1002/qua.24481>
 31. Bochevarov AD, Watson MA, Greenwood JR, Philipp DM (2016) Multiconformation, density functional theory-based PKa prediction in application to large, flexible organic molecules with diverse functional groups. *J Chem Theory Comput* 12(12):6001–6019. <https://doi.org/10.1021/acs.jctc.6b00805>
 32. Mansouri K, Cariello NF, Korotcov A, Tkachenko V, Grulke CM, Sprankle CS, Allen D, Casey WM, Kleinstreuer NC, Williams AJ (2019) Open-source QSAR models for PKa prediction using multiple machine learning approaches. *J Cheminformatics* 11(1):1–20
 33. Sproul DG, Palmer RK, Swanson JT, Lawless M (2010) QSAR in the pharmaceutical research setting: QSAR models for broad Large problems. *Curr Top Med Chem* 10(6):619–637
 34. Wu J, Kang Y, Pan P, Hou T (2022) Machine learning methods for PKa prediction of small molecules: advances and challenges. *Drug Discov. Today* 103372
 35. Lawler R, Liu Y-H, Majaya N, Allam O, Ju H, Kim JY, Jang SS (2021) DFT-machine learning approach for accurate prediction of pK_a . *J Phys Chem A* 125(39):8712–8722
 36. Marcel B, Czodrowski P (2020) Machine learning meets PK a. *F1000 Research* 9
 37. Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. *Science* 349(6245):255–260
 38. Marenich AV, Cramer CJ, Truhlar DG (2009) 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* 113(18):6378–6396. <https://doi.org/10.1021/jp810292n>
 39. Camaioni DM, Schwerdtfeger CA (2005) Comment on “accurate experimental values for the free energies of hydration of H^+ , OH^- , and H_3O^+ .” *J Phys Chem A* 109(47):10795–10797
 40. Kelly CP, Cramer CJ, Truhlar DG (2006) Aqueous solvation free energies of ions and ion–water clusters based on an accurate value for the absolute aqueous solvation free energy of the proton. *J Phys Chem B* 110(32):16066–16081
 41. Isse AA, Gennaro A (2010) Absolute potential of the standard hydrogen electrode and the problem of interconversion of potentials in different solvents. *J Phys Chem B* 114(23):7894–7899
 42. Marenich AV, Ho J, Coote ML, Cramer CJ, Truhlar DG (2014) Computational electrochemistry: prediction of liquid-phase reduction potentials. *Phys Chem Chem Phys* 16(29):15068–15106
 43. Ho J (2015) Are thermodynamic cycles necessary for continuum solvent calculation of PK_a s and reduction potentials? *Phys Chem Chem Phys* 17(4):2859–2868. <https://doi.org/10.1039/C4CP04538F>
 44. Thapa B, Schlegel HB (2016) Density functional theory calculation of pK_a ’s of Thiols in aqueous solution using explicit water molecules and the polarizable continuum model. *J Phys Chem A* 120(28):5726–5735. <https://doi.org/10.1021/acs.jpca.6b05040>
 45. Becke AD (1992) Density-functional thermochemistry. I. The effect of the exchange-only gradient correction. *J Chem Phys* 96(3):2155–2160
 46. Becke AD (1997) Density-functional thermochemistry. V. Systematic optimization of exchange-correlation functionals. *J Chem Phys* 107(20):8554–8560
 47. Lee C, Yang W, Parr RG (1988) Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys Rev B* 37(2):785
 48. Clark T, Chandrasekhar J, Spitznagel GW, Schleyer PVR (1983) Efficient diffuse function-augmented basis sets for anion calculations. III. The 3–21+ G basis set for first-row elements, Li–F. *J Comput Chem* 4(3):294–301
 49. Ditchfield R, Hehre WJ, Pople JA (1971) Self-consistent molecular-orbital methods. IX. An extended gaussian-type basis for molecular-orbital studies of organic molecules. *J Chem Phys* 54(2):724–728

50. Franci MM, Pietro WJ, Hehre WJ, Binkley JS, Gordon MS, DeFrees DJ, Pople JA (1982) Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *J Chem Phys* 77(7):3654–3665
51. Hariharan PC, Pople JA (1973) The influence of polarization functions on molecular orbital hydrogenation energies. *Theor Chim Acta* 28:213–222
52. Hehre WJ, Ditchfield R, Pople JA (1972) Self-consistent molecular orbital methods. XII. Further extensions of Gaussian-type basis sets for use in molecular orbital studies of organic molecules. *J Chem Phys* 56(5):2257–2261
53. Chai J-D, Head-Gordon M (2008) Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys Chem Chem Phys* 10(44):6615–6620
54. Grimme S, Antony J, Ehrlich S, Krieg H (2010) A consistent and accurate Ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J Chem Phys* 132(15):154104
55. Grimme S, Ehrlich S, Goerigk L (2011) Effect of the damping function in dispersion corrected density functional theory. *J Comput Chem* 32(7):1456–1465. <https://doi.org/10.1002/jcc.21759>
56. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Petersson GA, Nakatsuji H, Li X, Caricato M, Marenich AV, Bloino J, Janesko BG, Gomperts R, Mennucci B, Hratchian HP, Ortiz JV, Izmaylov A F, Sonnenberg JL, Williams DJ, Lipparini F, Egidi F, Goings J, Peng B, Petrone A, Henderson T, Ranasinghe D, Zakrzewski V G, Gao J, Rega N, Zheng G, Liang W, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Throssell K, Montgomery Jr JA, Peralta JE, Ogliaro F, Bearpark MJ, Heyd JJ, Brothers EN, Kudin KN, Staroverov VN, Keith TA, Kobayashi R, Normand J, Raghavachari K, Rendell AP, Burant JC, Iyengar SS, Tomasi J, Cossi M, Millam JM, Klene M, Adamo C, Cammi R, Ochterski JW, Martin RL, Morokuma K, Farkas O, Foresman JB, Fox DJ Gaussian 16 Rev C 01
57. Ertl P, Altmann E, McKenna JM (2020) The most common functional groups in bioactive molecules and how their popularity has evolved over time. *J Med Chem* 63(15):8408–8418. <https://doi.org/10.1021/acs.jmedchem.0c00754>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.