JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

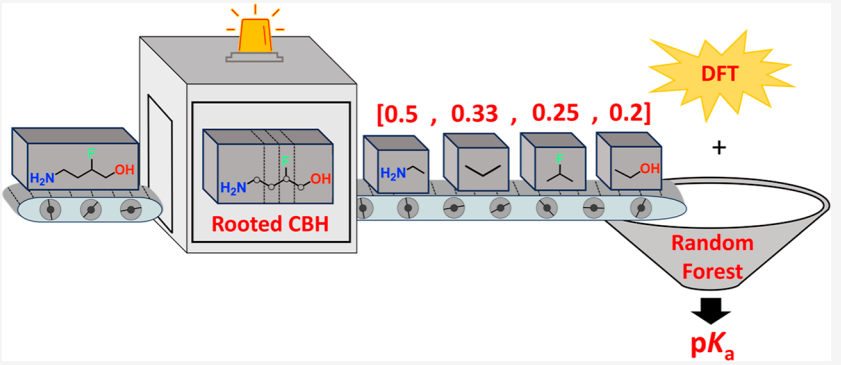
# Leveraging DFT and Molecular Fragmentation for Chemically Accurate p$K_a$ Prediction Using Machine Learning

Alec J. Sanchez, Sarah Maier, and Krishnan Raghavachari*

Cite This: J. Chem. Inf. Model. 2024, 64, 712−723 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information






**ABSTRACT:** We present a quantum mechanical/machine learning (ML) framework based on random forest to accurately predict the p$K_a$s of complex organic molecules using inexpensive density functional theory (DFT) calculations. By including physics-based features from low-level DFT calculations and structural features from our connectivity-based hierarchy (CBH) fragmentation protocol, we can correct the systematic error associated with DFT. The generalizability and performance of our model are evaluated on two benchmark sets (SAMPL6 and Novartis). We believe the carefully curated input of physics-based features lessens the model's data dependence and need for complex deep learning architectures, without compromising the accuracy of the test sets. As a point of novelty, our work extends the applicability of CBH, employing it for the generation of viable molecular descriptors for ML.

## 1. INTRODUCTION

The acid dissociation constant ($K_a$) and its logarithmic equivalent (p$K_a$) are valuable quantitative tools for assessing the strength of an acid or the stability of its conjugate base in solution. p$K_a$s are employed as a useful metric in numerous fields, including total synthesis, medicinal chemistry, and catalysis.[1−5] Experimental measurements of p$K_a$s are often complicated by complex solvent effects, synthetic challenges, as well as difficulties associated with compound isolation and purification.[6] Due to such complications, theoretical methods are frequently used to corroborate or even replace experimentally determined p$K_a$s. Computationally, p$K_a$ determination involves the evaluation of the free energy change for the deprotonation reaction. p$K_a$ is calculated as

$$pK_a = \frac{\Delta G^*_{aq}}{2.303RT} \quad (1)$$

where $\Delta G_{aq}$* is the aqueous free energy change for the deprotonation reaction, $R$ is the gas constant, and $T$ is the absolute temperature.

The efficient calculation of p$K_a$s for complex drug-like molecules remains a challenging task for computational chemists. Highly accurate, correlated methods like coupled-cluster theory including single and double excitations with perturbative triples (CCSD(T)[7]) are capable of chemical accuracy (<1 p$K_a$), though such methods come with steep computational costs. Due to the computational expense, the application of such methods has been limited.[7] Accurate composite theories such as the Gaussian-$n$ and complete basis set (CBS) methods are associated with decreased CPU time, though they are limited to systems with no more than 20 heavy atoms.

To tackle more sizable systems, quantum chemists typically employ faster density functional theory (DFT) methods. Despite DFT's relative speed, its accuracy often proves inadequate. For example, the absolute error of DFT-derived p$K_a$ predictions of alcohols and anilines can exceed 3 p$K_a$ units compared to experimental results.[8] These functional groups are central to many biochemical processes and are among the most

Received: December 4, 2023
Revised: January 17, 2024
Accepted: January 18, 2024
Published: February 1, 2024

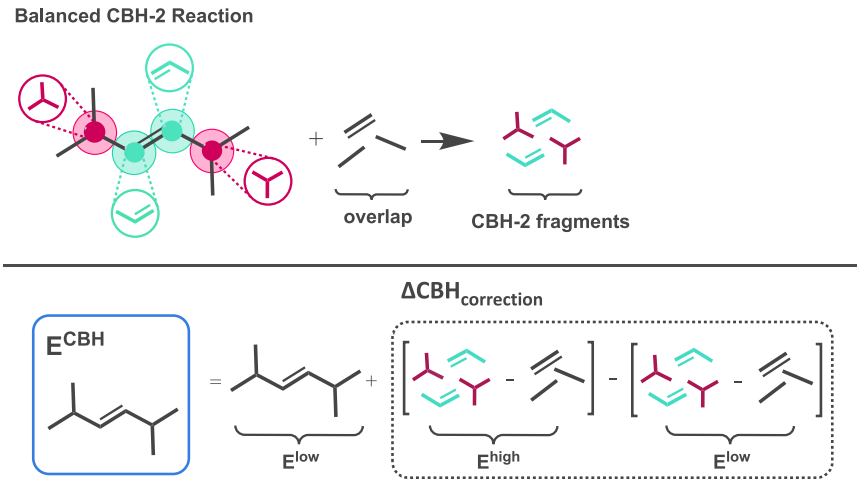



**Figure 1.** CBH-2 reaction scheme where $\Delta CBH_{correction}$ represents the correction to the low-level energy $E^{low}$ of the full system.

commonly encountered when studying bioactive molecules.[9] Thus, efficient computational models that maintain high accuracy across a broad range of chemical groups are in high demand.

To address this demand, several recent studies have demonstrated the successful integration of quantum mechanical (QM) calculations and machine learning (ML) techniques for highly accurate physicochemical property prediction. The adaptation of ML as a viable tool in the quantum chemist's toolbox has led to numerous applications in materials discovery, catalysis, drug design, etc.[10−15] When it comes to p$K_a$ prediction, one QM/ML model by Hunt et al.[16] used semiempirical features along with radial basis functions to obtain commendable performance on the SAMPL6 and Jensen data sets.[17] Similarly, Lawler et al.[18] used features derived from DFT with a kernel ridge regression model to achieve a low mean absolute error (MAE) of 0.60 on oxoacids.

In the current study, we introduce a random forest (RF)-based QM/ML framework for the prediction of highly accurate p$K_a$s. Specifically, we illustrate the development of a QM/ML p$K_a$ prediction model for use on complex drug-like molecules. The current work achieves chemically accurate p$K_a$ predictions by leveraging an ML model to correct low-level DFT. The RF model is trained using a modest data set of 2147 experimental p$K_a$s, originally published by Hunt et al.[16] Despite the modest training set size, our model can achieve high accuracy (MAE <1 p$K_a$ unit) and performs well in a benchmark against several state-of-the-art models found in the literature. Such accuracy is achieved in part through a carefully curated input of chemically relevant features. We employ physics-based features from DFT, along with several descriptors derived from molecular structure.[8]

As a point of novelty, we introduce a new class of ML descriptors for p$K_a$ prediction, the RootedCBH fingerprint.[19,20] This fingerprint acts as a basis for representing molecular substructure and its effect on p$K_a$.[21] Inspired by the ECFP[22] fingerprint as well as the class of so-called "rooted fingerprints",[23] the RootedCBH fingerprint is a new molecular descriptor that addresses the importance of chemical substructure in p$K_a$ prediction. RootedCBH provides a concise description of the chemical units that constitute a molecule as well as their proximity to a site of (de)protonation.

## 2. METHODS

**2.1. p$K_a$ Calculation Using DFT.** Given a general deprotonation reaction, e.g., AH ↔ $A^-$ + $H^+$, the corresponding logarithmic equivalent (p$K_a$) of the acid dissociation constant is calculated as

$$pK_a = \frac{\Delta G^*_{aq}}{2.303RT} \tag{2}$$

where $\Delta G_{aq}$* is the aqueous free energy change for the deprotonation reaction, $R$ is the gas constant ($1.985 \times 10^{-3}$ kcal/mol·K), and $T$ is the absolute temperature (298.15 K). $\Delta G_{aq}$* is calculated as

$$\Delta G^*_{aq} = G^*_{A^-,aq} + G^*_{H^+,aq} - G^*_{AH,aq} \tag{3}$$

where $G^*_{A^-,aq}$ and $G^*_{AH,aq}$ are the free energies associated with the conjugate base ($A^-$) and conjugate acid (AH) species, respectively, in aqueous phase using SMD[24] (solvation model based on density) implicit solvation. $G^*_{H^+,aq}$ is the free energy of a proton and is calculated via

$$G^*_{H^+,aq} = G^\circ_{H^+,gas} + \Delta G^*_{H^+,solv} + \Delta G^{1\text{ atm} \rightarrow 1\text{ M}} \tag{4}$$

where $\Delta G^*_{H^+,solv}$ (−265.9 kcal/mol)[25−27] 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 $G^\circ_{H^+,gas}$ is the gas phase proton free energy.

$$G^\circ_{H^+,gas} = H^\circ_{gas} - TS^\circ_{gas} \tag{5}$$

$H^\circ_{gas} = \left(\frac{5}{2}\right)RT$ is the enthalpic contribution of hydrogen gas, $S_{gas}{}^\circ$ (26.05 cal/mol·K) is the entropic contribution of hydrogen gas, and $T$ is the absolute temperature (298.15 K).

**2.2. Connectivity-Based Hierarchy + QM Descriptors.** Previous reports from the Raghavachari group provide an extensive review of the connectivity-based hierarchy (CBH), an error cancelation protocol based on a generalization of the isodesmic bond separation scheme.[28,29] CBH provides error corrections to low-level theory by generating reaction schemes with high degrees of bond-type matching and error cancellation. First, a molecule is broken down into its


https://doi.org/10.1021/acs.jcim.3c01923
J. Chem. Inf. Model. 2024, 64, 712−723

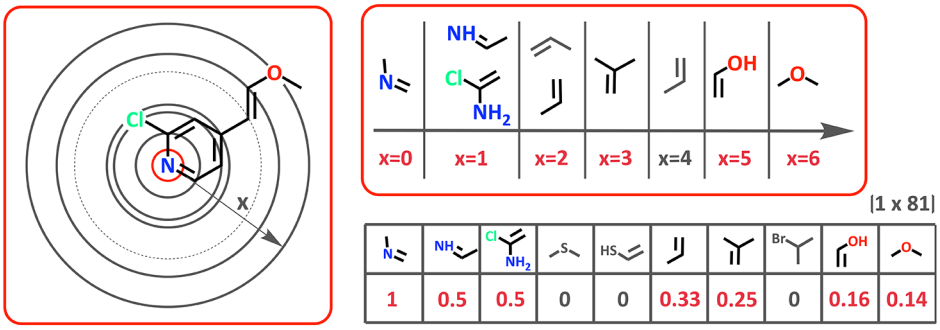



**Figure 2.** Example of RootedCBH where nitrogen is the site of protonation. Fragments in gray are either not present in the molecule or do not correspond to the fragment with the minimum path length.

constituent fragments based on a chosen rung of the hierarchy (i.e., CBH-*n*). While advancing through the rungs of the hierarchy, fragment size increases systematically, with CBH-0 units consisting of a single heavy (i.e., non-hydrogen) atom, CBH-1 units consisting of two adjacent heavy atoms (i.e., a single heavy atom bond), and CBH-2 units consisting of one heavy atom along with all heavy atoms in its immediate bonding environment. Explicit hydrogens are added to maintain the original hybridization.

Resultant fragments are then used to generate a correction term, which may be added to the low-level energy to extrapolate to the high-level energy. To illustrate, the difference in the sum of fragment energies for the low and high levels of theory is computed and then added to the low-level full system calculation to obtain the extrapolated high-level total energy. Energies of the overlapping fragments are subtracted to prevent overcounting. An example of a CBH-2 energy correction scheme is given in Figure 1. For a thorough explanation of CBH, please see Ramabhadran and Raghavachari.[28]

The success of CBH and its fragmentation schemes in achieving chemical accuracy at DFT cost for various thermodynamic properties, and specifically for p$K_a$, is well established.[28−37] Such work underscores how representative fragment reactions can be used to approximate complex chemical transformations. Through CBH, we observe the molecular substructures that form the basis of chemical space.

More recently, molecular descriptors based on CBH fragments were used as inputs to an ML model. In a 2020 study[19] by our group, vectors containing the coefficients of CBH-2 reactants and products (overlapping fragments) were used as novel ML fingerprints, coupled with a simple neural network framework, in order to accurately predict the heats of formation of small molecules. CBH fingerprints indicate the presence and count of various molecular substructures in each molecule. In this way, CBH fingerprints resemble the widely used ECFP fingerprint. The scope of our previous work was limited to the heats of formation of small molecules. The current study extends the applicability of CBH descriptors to a problem of increased complexity: the accurate determination of p$K_a$s for drug-like molecules. The following paragraphs discuss the generation of our newly developed RootedCBH fingerprints, a new class of CBH-based ML descriptors, as implemented in the current study.

First, CBH-2 product fragments were generated for all molecules in the training set (vide infra) using an in-house Python program and *xyz* structures as input. All resultant fragments were combined, and duplicate fragments removed, generating a set of unique CBH fragments. Due to the presence of sulfone and nitro groups, double bonds between oxygen and sulfur/nitrogen were not cut during fragmentation. Additionally, bonds to phosphorus were not cut during fragmentation due to the presence of phosphate groups. To limit the complexity of the feature space, those fragments that appeared in less than 10 unique molecules were removed. This procedure resulted in 81 unique CBH-2 product fragments. Higher rungs of CBH (i.e., CBH-3, 4, etc.) generate larger fragments and thus increase the complexity of the chemical descriptor space. CBH-2-based fragment descriptors were adopted for this study, striking a balance between two goals: maximizing the information content of the feature vectors and minimizing the feature space. The CBH feature vector length was set to 81, with one dimension reserved for each fragment.

Once a molecule is fragmented into its corresponding CBH-2 reactant fragments, the path-length (number of bonds) between the center of each fragment and the site of (de)protonation is determined ($x$ in Figure 2). This path-length, $x$, between each fragment and the site of (de)-protonation is passed to the function $1/(x + 1)$ and embedded in the feature vector at the appropriate index. In the limit that a fragment is infinitely far from the (de)protonation site, the weight is zero. Fragments absent from a molecule likewise carry a weight of zero. If a fragment is present more than once, only the minimum path-length fragment is kept.

The procedure for the generation of a RootedCBH fingerprint is illustrated in Figure 2. Here, concentric circles represent steps in the path, centered at the site of (de)protonation, with a gray arrow indicating increasing path-length. In Figure 2, gray substructures correspond to CBH fragments mapped to zero in the feature vector. These fragments are either not present in the molecule or do not correspond to the fragment with the minimum path length.

In addition to the CBH features, we include a description of the functional group involved in the (de)protonation reaction. Functional groups covered in this study include the following: phenol (Ph−OH), carboxylic acid (R−COOH), benzoic acid (Ph−COOH), thiol (R−SH), aliphatic alcohol (R−OH), primary amine (R−NH2), secondary amine with deprotonation ($R_2-N^-$), secondary amine with protonation ($R_2-NH^+$), tertiary amine ($R_3-N$), heterocyclic phenol (Het−OH), and alanine (Ph−NH2). Functional groups were identified using SMARTS strings and then one-hot encoded. Additional features specific to the (de)protonation site (e.g., hybridization and aromaticity) were generated with RDKit. We refer to the features from SMARTS and RDKit as "RDKit" for the remainder of the paper.


https://doi.org/10.1021/acs.jcim.3c01923
J. Chem. Inf. Model. 2024, 64, 712−723

Finally, a key point of our model is the use of physics-based DFT descriptors. DFT features were obtained from the M06-2X calculations described in the Methods Section and are enumerated in pink in Table 1. Thus, calculated p$K_a$s along with additional DFT-, CBH-, and RDKit-derived features were used as input features in the RF model. The model was trained to reproduce experimentally derived p$K_a$s. Feature permutation (Table S1) was performed to identify the most important features for our model.

**Table 1. Features Used in the RF Model**

| feature | length | type | source |
|---|---|---|---|
| Rooted-CBH2 | 81 | float64 | in-house script |
| H-heteroatom bond length | 1 | float64 | DFT (Gaussian 16) |
| Δ HOMO-LUMO gap | 1 | float64 | DFT (Gaussian 16) |
| Δ SCF energy | 1 | float64 | DFT (Gaussian 16) |
| Δ total electronic extent | 1 | float64 | DFT (Gaussian 16) |
| Δ enthalpy correction | 1 | float64 | DFT (Gaussian 16) |
| p$K_a$ | 1 | float64 | DFT (Gaussian 16) |
| heteroatom hybridization | 1 | integer | rdkit |
| is heteroatom aromatic? | 1 | 0 or 1 | rdkit |
| is heteroatom in ring? | 1 | 0 or 1 | rdkit |
| Δ charge | 1 | 1 or -1 | rdkit |
| Δ TPSA | 1 | float64 | rdkit |
| functional group identity | 11 | one-hot | SMARTS |

The final feature vector was of length 103. All features are listed in Table 1. Δ indicates the difference between the conjugate acid and conjugate base. In the table, the label "heteroatom" signals the heavy atom being (de)protonated. The enthalpy correction from DFT mentioned in Table 1 is obtained as the sum of the correction to the internal thermal energy (translational, rotational, vibrational, and electronic contributions) and $k_BT$.

**2.3. Data Sets.** A carefully curated data set[16] of 2386 molecules published by Hunt et al.[16] was used, featuring a diverse set of elements (C, N, O, S, F, Cl, Br, Si, P, and I). Experimental p$K_a$s for this data set range from −5.5 to 16.0. The data set covers a wide chemical space and includes molecules with up to 49 heavy atoms. Additionally, the data set features zwitterions and encompasses a diverse set of functional groups. The data set, as it was originally published, includes dications and dianions. These molecules are excluded from this study, noting that they make up less than 2% of the entire data set. The authors of the data set provide an "unambiguous and clear"[16] assignment of the (de)protonation site for each molecule, which we adopted. Further analysis of the data set (i.e., ring count and number of heteroatoms) can be found in Figure 3.

To evaluate the generalizability of our model and to compare its performance against popular ML models and commercial programs mentioned in the literature, we included the SAMPL6[38] and Novartis[39] data sets as external test sets. The SAMPL6 data set, shown in Figure 4, features 24 unique drug-like molecules with 31 unique experimental p$K_a$s and is often used to benchmark p$K_a$ predictive tools. Sites of (de)protonation were chosen based on the work by Xiong et al.,[40] where 29 of the 31 experimental p$K_a$s were labeled as belonging to either the most acidic or basic site. If the site was labeled most acidic, then each functional group was deprotonated and the lowest calculated p$K_a$ from DFT was used. If the site was labeled the most basic, each functional group was protonated, and the highest calculated p$K_a$ from DFT was used. For the two p$K_a$s not labeled (SM14: experimental p$K_a$ of 2.58 and SM18: experimental p$K_a$ of 11.02), the aniline group of SM14 was protonated, and the amide nitrogen not in a ring of SM18 was deprotonated.

In addition to the SAMPL6 data set, we included an additional test set of 101 molecules, which represent a subset of the total Novartis data set. For simplicity, only those molecules that contained a single reported site of (de)-protonation were selected. For these molecules, sites of (de)protonation were taken from the work by Liao and Nicklaus.[39] Molecules from SAMPL6 and Novartis were excluded from the training set (2147).

**2.4. Computational Details.** For p$K_a$ calculations using DFT, a conformational search was first performed on the neutral form of each molecule using the LowModeMD method, as implemented in the Molecular Operating Environ-

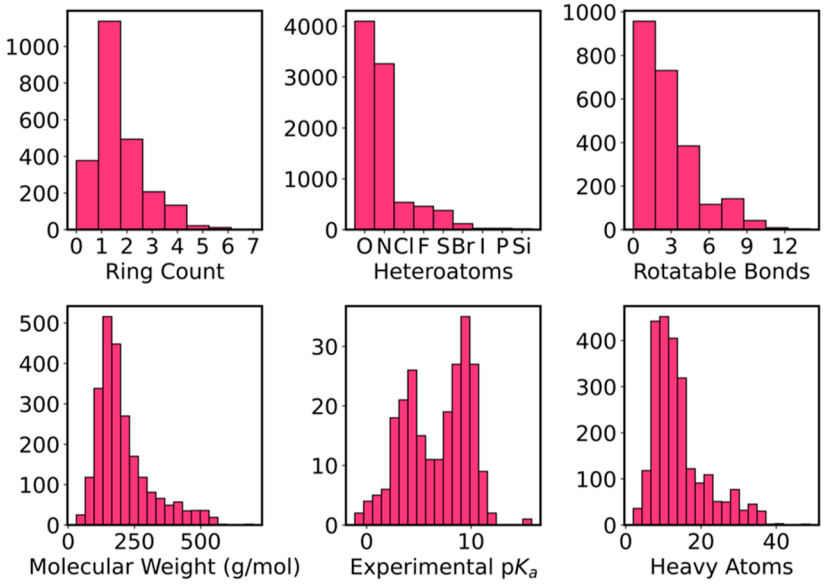



**Figure 3.** Properties of the total data set of 2386 molecules used to train, validate, and test the ML model.



https://doi.org/10.1021/acs.jcim.3c01923
J. Chem. Inf. Model. 2024, 64, 712−723

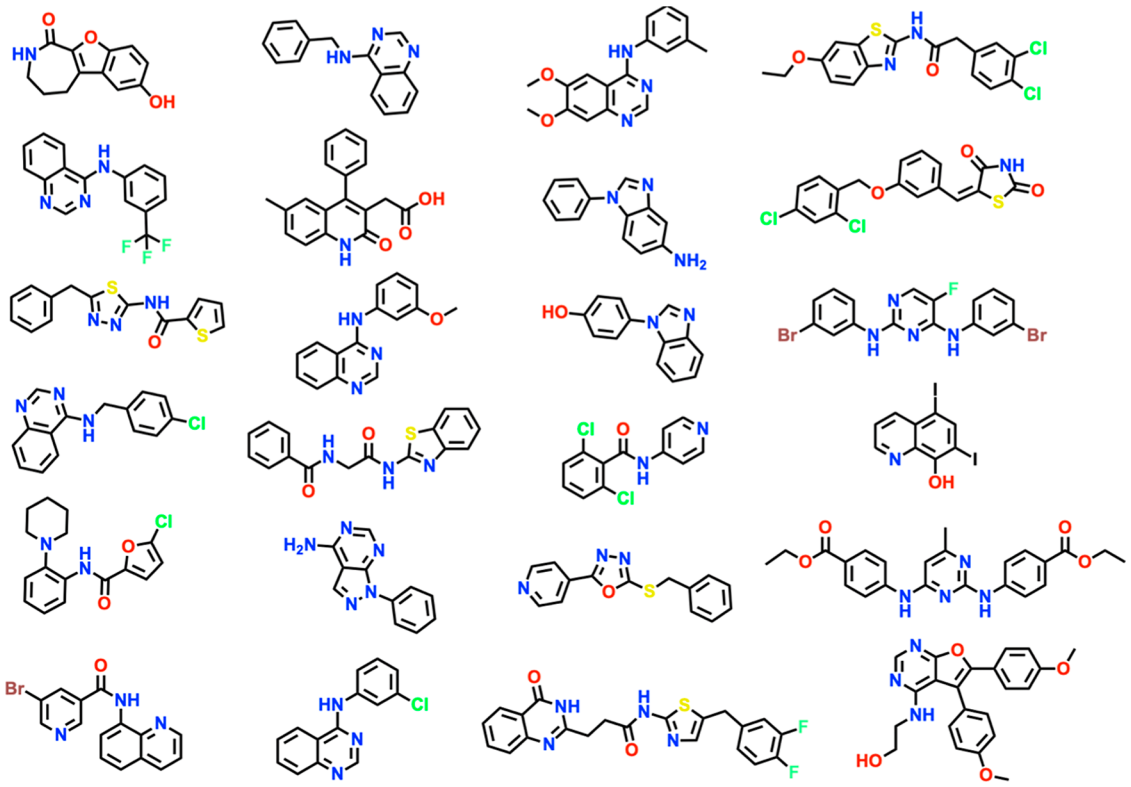


**Figure 4.** SAMPL6 data set featuring 24 unique molecules with 31 unique p$K_a$s.

ment (MOE) software (version 2019.01).[41] Structural complexity, the presence of many rotatable bonds, and intramolecular hydrogen bonding with the (de)protonation site underline the need for a conformational search in this study. An energy window of 7 kcal/mol was used. Two conformations were considered identical if their root-mean-square deviations (RMSDs) were less than 0.25 Å.

Following the search, the 10 lowest energy conformations as identified by MOE were further refined via a protocol described by Zeng et al.[42] The protocol prescribes geometry optimization with the M06-2X functional using tight convergence and an ultrafine integration grid. The protocol employs the SMD[24] implicit solvation model in water to account for solvent effects. If the reaction corresponding to each experimental p$K_a$ (neutral as reactant) is a protonation, the 6-31G(d) basis set is used for both the neutral and charged species; otherwise, 6-31+G(d) is used. The use of different basis sets accounts for the diffuse nature of anions. For molecules containing iodine, the LANL2DZ pseudopotential was used.

Frequency calculations were performed on the lowest energy conformer obtained from DFT and scaled by 0.9465 and 0.9500 for calculations employing 6-31G(d) and 6-31+G(d), respectively. With the optimized geometry of the neutral molecule, (de)protonation was carried out, and optimization and frequency calculations were performed again using the same level of theory as the neutral molecule. All geometries were confirmed to be local minima. In the case of zwitterions, 6-31+G(d) was used, and frequencies were scaled by 0.9500. All DFT calculations were performed using the Gaussian 16 program suite.[43]

**2.5. RF Methods.** For the ML portion of the study, the data set was split 90:10 for training and testing, respectively. The data was split using the stratified_continuous_split function of the verstack[44] Python library to ensure that the distribution of experimental p$K_a$s in the training set resembled that of the entire data set. This split type was chosen to ensure that experimental p$K_a$s near the extrema were included in training. Figure 3 clearly illustrates the bimodal distribution of experimental p$K_a$s, with peaks near 5 and 10, with experimental p$K_a$s below 0 and above 12 being far less represented in the data set.

RF is one of the most popular supervised learning algorithms and consists of an ensemble of decision trees that can be used for classification or regression. Decision trees are made up of decision nodes and have criteria (e.g., if-else statements) that partition and pass the data to subsequent decision nodes. When there are no more decision nodes, one has reached a root node, and a prediction is made based on an average of the data points in that node. Since RF utilizes bagging (sampling with replacement), each decision tree is trained using only a subset of the data. Additionally, each tree sees a subset of the feature set, thereby mitigating the issue of overfitting.

To tune the hyperparameters of the RF model, we submitted the training set to $k$-fold cross validation ($k$ = 3) as implemented in scikitlearn. Hyperparameters of the RF model were tuned using a randomized grid search in scikitlearn (Table S2). The parameters which gave the lowest validation error are given as n_estimators = 2000, max_features = 0.5, max_depth = 40, min_samples_split = 2, and min_samples_leaf = 2. The full hyperparameter grid is given in Table S2.

## 3. RESULTS AND DISCUSSION

For the total data set of 2386 molecules, M06-2X calculations produced an MAE of 1.82 p$K_a$ units and an RMSE of 2.38 p$K_a$ units compared to those from experiment. These results are commendable, given that DFT often produces much higher



https://doi.org/10.1021/acs.jcim.3c01923
J. Chem. Inf. Model. 2024, 64, 712−723

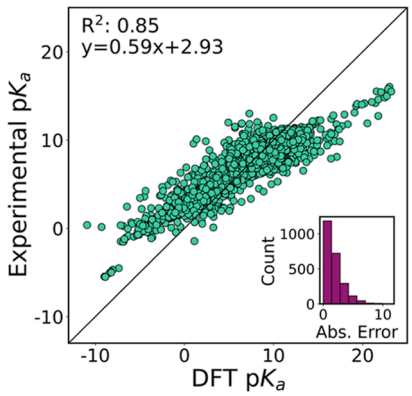



**Figure 5.** Correlation and distribution of absolute error between the calculated p$K_a$ from DFT and experimental p$K_a$ in the full data set (2386 molecules).

errors.[8,37,45−50] Nonetheless, chemical accuracy is regarded as <1 p$K_a$ unit. Thus, DFT alone does not achieve chemical accuracy. In fact, nearly 40% of the data had an absolute deviation larger than two p$K_a$ units, as seen in the histogram of Figure 5. When compared to the $x = y$ line (black diagonal in Figure 5), the DFT results are skewed significantly, with a slope of 0.59. Nevertheless, the high correlation ($R^2 = 0.85$) in the linear fit suggests that the error in the calculated values is systematic, a common observation with QM calculations. ML models are often effective at removing such a systematic error.

Recognizing that the DFT error is systematic, we used a RF framework to correct DFT-calculated p$K_a$s. Calculated p$K_a$s, along with additional DFT-, CBH-, and RDKit(+SMARTS)-derived features, were used as input features in our RF model. The model was trained to reproduce experimentally derived p$K_a$s. The data set (2386 molecules) was split 90:10 for training (2147) and testing (239), respectively. Figure 6 highlights the distribution of functional groups in the training and test splits. The figure shows a similar distribution in the two sets, which may be a result of the stratified split. Aliphatic alcohols and amines dominate the data set, while heterocyclic alcohols and thiols are not well-represented.

The model was first evaluated on the test split (~240 molecules). It achieved MAE and RMSE values of 0.51 and 0.76 p$K_a$ units, respectively. After application of the RF correction, nearly 90% of calculated p$K_a$ absolute errors fall within 1 p$K_a$ unit, and ~95% fall within 2 p$K_a$ units. The maximum error of the test set is 3.02 p$K_a$ units. By comparison,

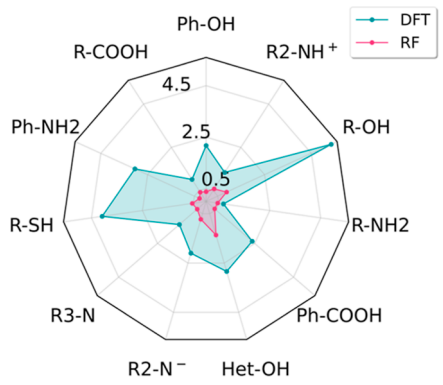



**Figure 7.** Spider plot of the functional-group-specific MAEs associated with DFT and RF model on the test split.

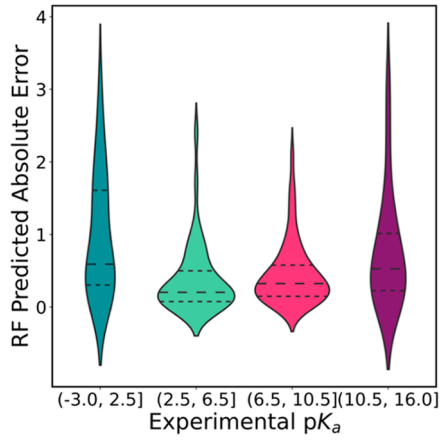



**Figure 8.** Violin plot of the RF absolute errors of the test split at different ranges of experimental p$K_a$s. From bottom to top, the dashed lines indicate the 25th percentile, median, and 75th percentile. Distributions less than zero are a fabrication of the kernel density estimation function.

using DFT alone, ~35% of absolute errors fall within 1 p$K_a$ unit, ~65% fall within 2 p$K_a$ units, and the maximum error is 9.45 p$K_a$ units.

An in-depth look at the functional group dependence of systematic error, considering DFT and RF-corrected values, shows that aliphatic alcohols, R−OH, thiols, R−SH, and anilines, Ph−$NH_2$ have the largest DFT MAEs (5.35, 4.08, and 3.04 p$K_a$ unit error, respectively, with respect to experiment)

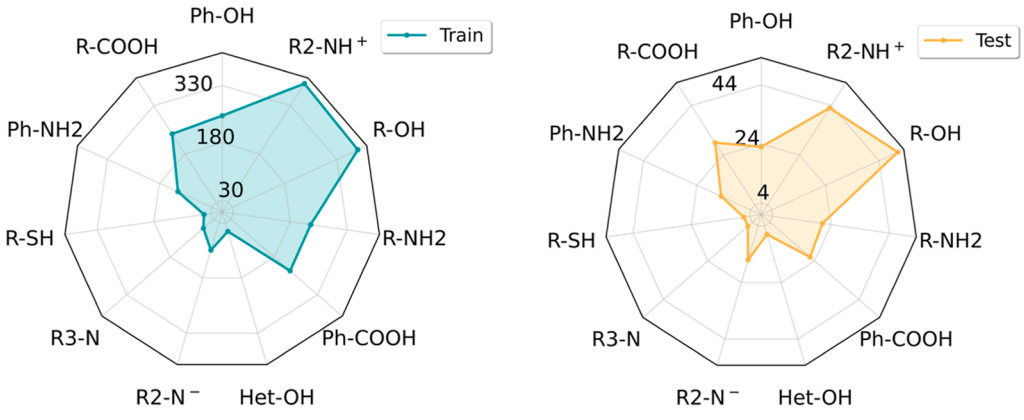



**Figure 6.** Frequency of deprotonated/protonated functional groups in the train and test split.



https://doi.org/10.1021/acs.jcim.3c01923
J. Chem. Inf. Model. 2024, 64, 712−723

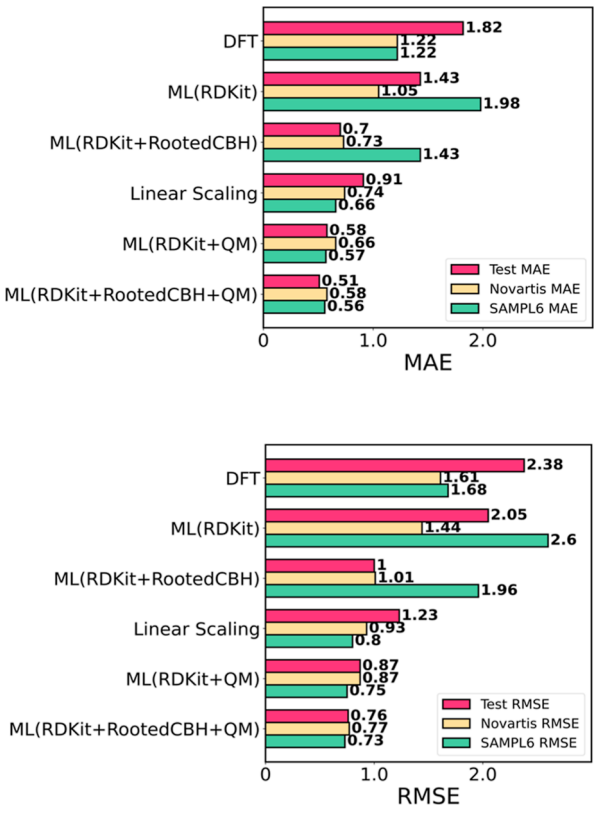



**Figure 9.** MAE and RMSE of predicted p$K_a$s from test split, SAMPL6, and Novartis and the effect of features on the RF model. DFT corresponds to the raw theoretical value, without ML. Linear scaling refers to the correction from linear regression fit to the training data (DFT-calculated p$K_a$ vs experimental p$K_a$).

but were effectively corrected (0.88, 0.54, and 0.29 p$K_a$ unit error, respectively, with respect to those of experiment) using the RF model with RDKit + RootedCBH + QM features. This fact is illustrated in Figure 7. In fact, the error for every functional group was improved by using our model. After application of the RF correction, all functional-group-specific MAEs fall within chemical accuracy except for heterocyclic alcohols. Interestingly, heterocyclic alcohols were one of the most underrepresented groups in the training set (53 molecules) and test (7 molecules) set. This is illustrated in Figure 6. In addition, five of the seven heterocyclic alcohol test compounds contained two or more heterocyclic atoms, increasing the chemical complexity of these systems.

We also explored the model performance across the full range of experimental p$K_a$s featured in the test set. Figure 8 illustrates the performance of the method across four ranges of experimental p$K_a$s. The middle two violins exhibit relatively distinct peaks. They show a narrow distribution of errors and a 75th percentile mark well below 1 p$K_a$. Poorer model performance is observed in the first (considering experimental p$K_a$s between −3 and 2.5) and last (considering experimental p$K_a$s between 10.5 and 16) violins, which represent the extrema of experimental values. A similar observation was noted by Hunt et al.[16] As mentioned previously, the distribution of experimental p$K_a$s is bimodal with peaks at 5 and 10. p$K_a$s below 0 and above 12 are generally underrepresented in the data set. This shortage of data at the extrema may lead to a degradation of model performance. Incorporating additional data points with experimental p$K_a$s falling in the most acidic and most basic ranges should help alleviate this weakness.

To evaluate the generalizability and performance against state-of-the-art academic and commercial models, we conducted additional testing on two benchmark test sets: SAMPL6 and Novartis. The importance of our chosen features is stressed in Figure 9. Site-specific features (e.g., hybridization, aromaticity, and functional group) from RDKit with RF result in improved RMSE values for test (2.05 p$K_a$ units) and Novartis (1.44 p$K_a$ units) compared to those with using DFT alone (2.38 and 1.61 p$K_a$ units). However, RDKit features worsen the model performance for SAMPL6. This may be due to the structural and chemical complexity of the molecules in the SAMPL6 test set. Indeed, these molecules exhibit complex interactions (i.e., donating/withdrawing groups and hydrogen/halogen bonding) that may not be well-accounted for by RDKit-derived features alone. With the inclusion of structural features from RootedCBH, the RMSE and MAE fall to roughly 1 p$K_a$ unit for test and Novartis. SAMPL6 sees less of an improvement in performance from the addition of RootedCBH features; however, the RMSE and MAE are still not within acceptable accuracy (1.96 RMSE and 1.43 MAE).

With the inclusion of physics-based features from QM, the RMSE and MAE for all three test sets fall within 1 p$K_a$ for ML(RDKit + QM). The result illustrates the use of low-level DFT calculations as a viable foundation for learning. This is supported further by the results from a simple linear scaling

**Table 2. SAMPL6 Performance of the Top Commercial and Academic Models Listed in the Literature[a]. MAE and RMSE are Given in Units of p$K_a$**

| author/model | class | MAE | RMSE | $R^2$ | comments | model |
|---|---|---|---|---|---|---|
| Hunt's Model[16] | academic | | 0.85 | | | ML: radial basis function + semiempirical QM |
| MolGp$K_a$[51,52] | academic | 0.522 | 0.773 | 0.907 | removed 5 p$K_a$ values (SM11, SM22_1, SM22_2, SM14, SM18) | ML: graph neural network trained on 1.1 million calculated p$K_a$ using ACD/p$K_a$ |
| ACD Laboratories[16] | commercial | | 0.77 | 0.92 | | ACD/p$K_a$ classic |
| MF-SuP-p$K_a$[52] | academic | 0.687 | 0.751 | 0.912 | removed 5 p$K_a$ values (SM11, SM22_1, SM22_2, SM14, SM18) | ML: graph neural network trained on 1.1 million calculated p$K_a$ using ChemAxon |
| S + p$K_a$[53] | commercial | 0.59 | 0.73 | | | ensemble of neural networks |
| Graph-p$K_a$[40] | academic | 0.594 | 0.726 | 0.918 | removed 2 p$K_a$ values (SM14 & SM18) | ML: multi-instance graph neural network trained on 17K experimental p$K_a$ |
| Pracht et al[54] | academic | 0.58 | 0.68 | 0.937 | removed 2 p$K_a$ values (SM14 & SM18) | LFER + QM + conformer sampling |
| Epik v 7 emsemble[55] | commercial | 0.48 | 0.61 | | | ensemble of atomic GCNN's trained on 42K p$K_a$ |

[a]Full table containing twenty-one literature models can be found in Table S3.



https://doi.org/10.1021/acs.jcim.3c01923
J. Chem. Inf. Model. 2024, 64, 712−723

Table 3. Novartis Performance across Popular Commercial Programs. MAE and RMSE are Given in Units of p$K_a$

| author | class | MAE | RMSE | comments | model |
|---|---|---|---|---|---|
| Schrodinger[57,58] | commercial | 1.02 | 1.56 | program failed to predict 12 p$K_a$ | Jaguar |
| ChemAxon[59] | commercial | 1.06 | 1.55 | | Marvin |
| SciTegic[60] | commercial | 0.73 | 1.35 | program failed to predict 1 p$K_a$ | Pipeline Pilot |
| CompuDrug[61] | commercial | 0.59 | 1.14 | program failed to predict 1 p$K_a$ | Pallas |
| Schrodinger[62,63] | commercial | 0.78 | 1.03 | program failed to predict 2 p$K_a$ | Epik |
| University of Georgia/Environmental Protection Agency[64,65] | commercial | 0.73 | 1.01 | program failed to predict 2 p$K_a$ | SPARC |
| SimulationsPlus[66] | commercial | 0.53 | 1.00 | | ADMET Predictor |
| ACD Laboratories[67] | commercial | 0.36 | 0.56 | | ACD/p$K_a$ |
| Pharma algorithms[68] | commercial | 0.33 | 0.52 | program failed to predict 1 p$K_a$ | ADME Boxes |

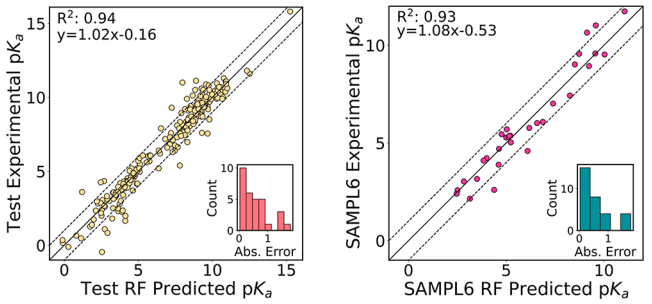



**Figure 10.** Correlation and distribution of errors from the RF-predicted p$K_a$ and experimental p$K_a$ on the test (left) and SAMPL6 set (right). Dashed lines indicate a ±1 p$K_a$ unit.

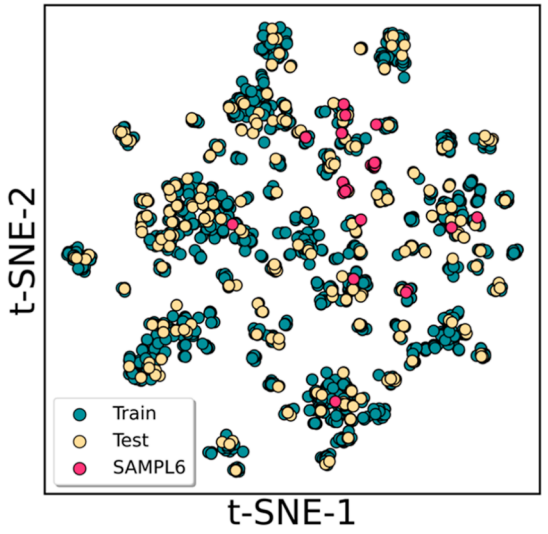



**Figure 11.** t-SNE plot showing the train, test, and SAMPL6 latent space in two dimensions.

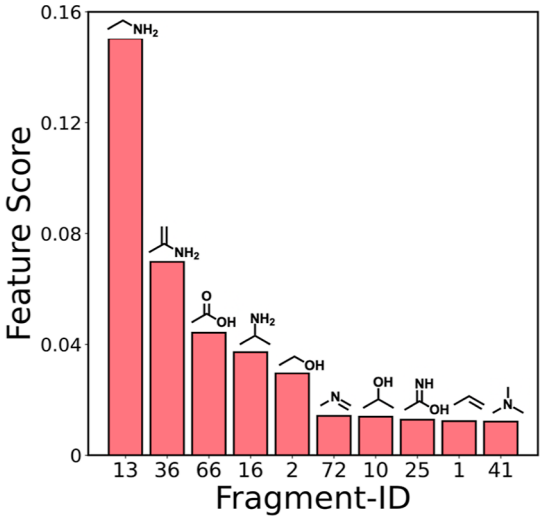



**Figure 12.** CBH fragments with the highest feature score based on feature permutation.

model, fit on the training data. Using the best fit line, the MAE and RMSE of DFT are halved for all three test sets. This aligns with the results published by Sanchez and Raghavachari.[8] The observation is crucial because it underscores the importance of DFT as a feature. Simple scaling produces a low MAE for SAMPL6. This result may be due to the comparatively high number of amines in the training set. DFT error for amines is relatively low. In contrast, the test and Novartis sets are composed of molecules featuring a wider range of functional groups, and their DFT errors are higher in comparison to those of SAMPL6. By adding RootedCBH, the model ML(RDKit + RootedCBH + QM) is able to achieve one of the best performances in the literature: test (0.51 MAE), Novartis (0.58 MAE), and SAMPL6 (0.56 MAE). Hyperparameter tuning was performed for all feature sets shown.

Importantly, our model was able to achieve results rivaling many of the current models from the literature for the SAMPL6 data set and the subset of the Novartis data set, Tables 2 and 3. We give less focus to the model's performance on the Novartis set since we only consider a subset of the total data set; however, all performances reported for Novartis correspond to performances obtained on the same 101 molecules used in this study (Table 3). Our model achieves an MAE of 0.56 p$K_a$ units and an RMSE of 0.73 p$K_a$ units on the SAMPL6 molecules. For the subset of Novartis structures containing a single deprotonation site, we achieve an MAE of 0.58 p$K_a$ units and an RMSE of 0.77 p$K_a$ units. Compared to nine commercial programs in the literature, our model achieves a respectable performance, as shown in Table 3. All absolute errors for the SAMPL6 data set fall within 2 p$K_a$ units, and ∼85% of the errors fall within 1 p$K_a$ unit, Figure 10.

The effect of including RootedCBH fragment features improves model performance overall on the test split and both benchmark sets (Novartis and SAMPL6). For the test split, the RMSE decreases from 2.05 to 1 p$K_a$ and from 1.44 to 1 p$K_a$ units for Novartis. RootedCBH fragments do not lead to a substantial average improvement in accuracy when considering SAMPL6. We suggest a few possible explanations for this observation. First, CBH fragments may be unable to capture important conformational effects present in some SAMPL6 molecules. Second, the use of "rooted" fingerprints which identify a single site of (de)protonation may be


https://doi.org/10.1021/acs.jcim.3c01923
J. Chem. Inf. Model. 2024, 64, 712−723

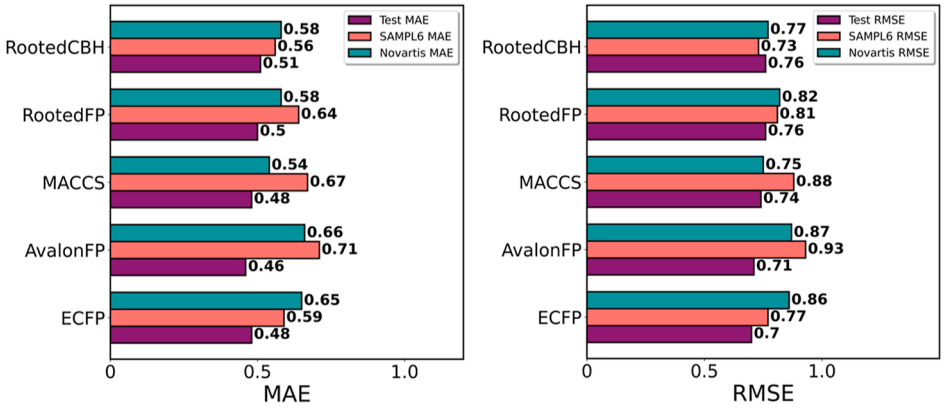



**Figure 13.** Comparison of our RootedCBH fingerprint against popular fingerprints in the literature. RootedFP, AvalonFP, ECFP: 2048-bit size, radius = 3 for ECFP. MACCS: 166 bit size.

inappropriate when many close lying p$K_a$s are involved (the case for several SAMPL6 molecules). It is important to note that more than 70% of SAMPL6 absolute errors (23 molecules) improved or remained the same upon the addition of CBH features. Only two molecules experienced an increase in absolute error greater than 1 p$K_a$ unit upon the addition of CBH features and seven by more than 0.5 p$K_a$ units.

For our QM/ML model, the largest prediction error on the SAMPL6 set comes from SM24, which has an absolute error of 1.8 p$K_a$ units and squared error of 3.18 p$K_a$ units compared to that of experiment. For this molecule, the most basic site from DFT, a heterocyclic secondary amine, was chosen, following the protocol by Xiong et al.[40] However, in a publication by Hunt et al.,[16] the authors protonate the secondary amine in the alkyl chain. If we compare the DFT-calculated p$K_a$s, we notice a large discrepancy between the amine in the alkyl chain (−6 p$K_a$) and the amine in the heterocyclic ring (2 p$K_a$). Since there is a discrepancy in the correct protonation site of SM24, we also report the SAMPL6 performance with this outlier removed (0.52 MAE and 0.67 RMSE). With the single outlier removed, our model approaches state-of-the-art accuracy on the SAMPL6 data set, as compared to that of the Epik v 7 ensemble model (0.48 MAE and 0.61 RMSE), which has the best performance in the literature to the best of our knowledge. Reduced model performance was also seen in SM05 (1.50 p$K_a$ unit absolute error). Similar results were reported by Yang et al.[56] who hypothesized that strong intra/intermolecular hydrogen bonding may be at play.

For Novartis, the largest outliers include id 88 and 102, which have an absolute error of 2.42 and 2.16 p$K_a$ units, respectively. While we believe that the protonation site is correct for id 88, the site is part of a bridged bicyclic ring, a rather uncommon group. Regarding id 102, the deprotonation site is a heterocyclic alcohol that can undergo intramolecular hydrogen bonding, and as mentioned previously, the model yields the worst performance for heterocyclic alcohols.

One point of interest is the model's commendable performance despite the use of a small training set (∼2100). Many of the models listed in Table 2 make use of extensive data sets, with training sets ranging from 17,000 to 1.1 million molecules. We believe that our use of physics-based descriptors lessens our model's dependence on big data. Additionally, the results of a 1-nearest-neighbor model, considering all RDKit + QM + RootedCBH descriptors, can be found in Table S4. The results of this model suggest that the RF model has not simply memorized the training data.

Site-specific features from RDKit, structural features from CBH fragments, and physics-based features from DFT calculations were used as chemically relevant features. In order to visualize the overlap in feature space among the train, test, and SAMPL6 sets, we include a t-distributed stochastic neighbor embedding (t-SNE) plot in Figure 11. Here, the 103-dimensional feature space has been condensed to 2 dimensions for visualization. From the plot, it is clear that a majority of the molecules in the train and test split, as well as the SAMPL6 set, occupy similar spaces. Based on feature importance calculated using scikitlearn, we observe that ΔSCF and calculated p$K_a$ from DFT are the two most important features in the model. This result underscores the importance of using DFT-derived features.

Figure 12 shows a bar plot of the 10 most important CBH fragments, ranked by their feature permutation score. It is interesting to note the prevalence of fragments featuring amines (13, 36, 16, 72, 25, 41), alcohols, and carboxylic acids (2, 10, 25). This aligns with our expectations since N and O are the most abundant heavy atoms (excluding carbon) in the data set (see Figure 3). Furthermore, nitrogen and oxygen are also present in 10 out of the 11 ionizable sites within the data set. Based on these observations, it is easy to understand why these fragments are important for model learning. This result is informative as it provides us with a cheminformatics-based understanding of our model.

Lastly, we evaluated the performance of our RootedCBH fingerprints against commonly used structural fingerprints in the literature (RootedFP,[69] MACCS,[70] AvalonFP,[71] and ECFP[22]) (Figure 13). Our motivation for developing the RootedCBH was inspired by two of these feature types, ECFP and RootedFP. For our fingerprint comparison, we kept the RDKit and QM features consistent, replacing only RootedCBH. Hyperparameter tuning was done for each set of features. The bar plots in Figure 13 illustrate that our RootedCBH fingerprints perform well, with a consistent MAE (0.51−0.58) and RMSE (0.73−0.77) across all three test sets. It is worth noting that RootedFP, MACCS, AvalonFP, and ECFP displayed a slightly better performance on the test split compared to that of RootedCBH; however, these fingerprints tend to perform significantly worse on at least one of the test sets. For instance, AvalonFP had the lowest MAE (0.46) on the test split compared with that of the other


https://doi.org/10.1021/acs.jcim.3c01923
J. Chem. Inf. Model. 2024, 64, 712−723

fingerprints, but conversely, it had the largest MAE on SAMPL6 (0.71). This observation emphasizes the generalizability of RootedCBH across various chemical spaces.

It is important to note that the length of our RootedCBH (81 bits) is only a fraction of the size of ECFP, RootedFP, AvalonFP (2048 bits), and MACCS (166 bits). This result is significant because it demonstrates that RootedCBH fragments provide a concise yet informative description of a molecule's structure, which in turn play a pivotal role in p$K_a$ prediction.

## 4. CONCLUSIONS

To summarize, we developed a QM/ML framework with RF to accurately predict p$K_a$s of complex molecules using physics-based features from DFT and structural features from our CBH fragmentation protocol. Notably, this work extends the applicability of RootedCBH fragmentation and QM/ML frameworks as viable tools for predicting accurate physiochemical properties. Our model corrects functional-group-specific deficiencies associated with DFT and achieves impressive accuracy on two external test sets, the SAMPL6 and Novartis data sets. If we exclude SM21 from SAMPL6, we can achieve a near-state-of-the-art performance (0.52 MAE and 0.67 RMSE).

Despite the small training set size, our model achieves a high accuracy. We believe the use of physics-based descriptors and carefully curated input of chemically relevant features lessens the model's data dependence and need for complex deep learning architectures. One drawback our model experiences is the need to identify a single site of deprotonation, especially for complex molecules with many ionizable sites and close lying p$K_a$s. In these cases, a Boltzmann weighing of various microstates may be more appropriate.

## ■ ASSOCIATED CONTENT

### Data Availability Statement

Model and data sets used in this publication are freely available at https://github.com/sarmaier/RootedCBH_pka.

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c01923.

20 most important features as determined by feature permutation, *t*-SNE plots of the train and SAMPL6 data sets using CBH, full hyperparameter grid in the randomized grid search, and SAMPL6 performances found in the literature (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Krishnan Raghavachari** − *Department of Chemistry, Indiana University?, Bloomington, Indiana 47405, United States;* orcid.org/0000-0003-3275-1426; Email: kraghava@indiana.edu

### Authors

**Alec J. Sanchez** − *Department of Chemistry, Indiana University?, Bloomington, Indiana 47405, United States*

**Sarah Maier** − *Department of Chemistry, Indiana University?, Bloomington, Indiana 47405, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c01923


### Author Contributions

A.J.S. and S.M. contributed equally to the work.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We acknowledge the 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.

## ■ REFERENCES


(1) Manallack, D. T. The pK(a) Distribution of Drugs: Application to Drug Discovery. *Perspect. Med. Chem.* **2007**, *1*, 25−38.

(2) Charifson, P. S.; Walters, W. P. Acidic and Basic Drugs in Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2014**, *57*, 9701−9717.

(3) Bell, R. P. *The Proton in Chemistry*; Springer Science & Business Media, 2013.

(4) Stewart, R. *The Proton: Applications to Organic Chemistry*; Elsevier, 2012; Vol. *46*.

(5) Manallack, D. T.; Prankerd, R. J.; Yuriev, E.; Oprea, T. I.; Chalmers, D. K. The Significance of Acid/Base Properties in Drug Discovery. *Chem. Soc. Rev.* **2013**, *42*, 485−496.

(6) Ho, J.; Coote, M. L. A Universal Approach for Continuum Solvent pK a Calculations: Are We There Yet? *Theor. Chem. Acc.* **2010**, *125*, 3−21.

(7) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. A Fifth-Order Perturbation Comparison of Electron Correlation Theories. *Chem. Phys. Lett.* **1989**, *157*, 479−483.

(8) Sanchez, A. J.; Raghavachari, K. Development and Assessment of a ChemInformatics Model for Accurate pKa Prediction in Aqueous Medium. *Theor. Chem. Acc.* **2023**, *142*, 86.

(9) Ertl, P.; Altmann, E.; McKenna, J. M. The Most Common Functional Groups in Bioactive Molecules and How Their Popularity Has Evolved over Time. *J. Med. Chem.* **2020**, *63*, 8408−8418.

(10) Maier, S.; Collins, E. M.; Raghavachari, K. Quantitative Prediction of Vertical Ionization Potentials from DFT via a Graph-Network-Based Delta Machine Learning Model Incorporating Electronic Descriptors. *J. Phys. Chem. A* **2023**, *127*, 3472−3483.

(11) Muller, C.; Rabal, O.; Diaz Gonzalez, C. Artificial Intelligence, Machine Learning, and Deep Learning in Real-Life Drug Design Cases. In *Artificial Intelligence in Drug Design*; Heifetz, A., Ed.; Methods in Molecular Biology; Springer US: New York, NY, 2022; Vol. *2390*, pp 383−407.

(12) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K. R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chem. Rev.* **2021**, *121*, 9816−9872.

(13) Han, Y.; Ali, I.; Wang, Z.; Cai, J.; Wu, S.; Tang, J.; Zhang, L.; Ren, J.; Xiao, R.; Lu, Q.; Hang, L.; Luo, H.; Li, J. Machine Learning Accelerates Quantum Mechanics Predictions of Molecular Crystals. *Phys. Rep.* **2021**, *934*, 1−71.

(14) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 2336−2347.

(15) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *npj Comput. Mater.* **2016**, *2*, 16028.

(16) Hunt, P.; Hosseini-Gerami, L.; Chrien, T.; Plante, J.; Ponting, D. J.; Segall, M. Predicting p $K_a$ Using a Combination of Semi-Empirical Quantum Mechanics and Radial Basis Function Methods. *J. Chem. Inf. Model.* **2020**, *60*, 2989−2997.

(17) Jensen, J. H.; Swain, C. J.; Olsen, L. Prediction of p $K_a$ Values for Druglike Molecules Using Semiempirical Quantum Chemical Methods. *J. Phys. Chem. A* **2017**, *121*, 699−707.

(18) Lawler, R.; Liu, Y.-H.; Majaya, N.; Allam, O.; Ju, H.; Kim, J. Y.; Jang, S. S. DFT-Machine Learning Approach for Accurate Prediction of p K a. *J. Phys. Chem. A* **2021**, *125*, 8712−8722.
(19) Collins, E. M.; Raghavachari, K. Effective Molecular Descriptors for Chemical Accuracy at DFT Cost: Fragmentation, Error-Cancellation, and Machine Learning. *J. Chem. Theory Comput.* **2020**, *16*, 4938−4950.
(20) Collins, E. M.; Raghavachari, K. A Fragmentation-Based Graph Embedding Framework for QM/ML. *J. Phys. Chem. A* **2021**, *125*, 6872−6880.
(21) Wu, J.; Kang, Y.; Pan, P.; Hou, T. Machine Learning Methods for pKa Prediction of Small Molecules: Advances and Challenges. *Drug Discovery Today* **2022**, *27*, 103372.
(22) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(23) Lu, Y.; Anand, S.; Shirley, W.; Gedeck, P.; Kelley, B. P.; Skolnik, S.; Rodde, S.; Nguyen, M.; Lindvall, M.; Jia, W. Prediction of p $K_a$ Using Machine Learning Methods with Rooted Topological Torsion Fingerprints: Application to Aliphatic Amines. *J. Chem. Inf. Model.* **2019**, *59*, 4706−4719.
(24) 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.
(25) Camaioni, D. M.; Schwerdtfeger, C. A. Comment on "Accurate Experimental Values for the Free Energies of Hydration of H+, OH- and H3O+. *J. Phys. Chem. A* **2005**, *109*, 10795−10797.
(26) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. 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* **2006**, *110*, 16066−16081.
(27) Isse, A. A.; Gennaro, A. Absolute Potential of the Standard Hydrogen Electrode and the Problem of Interconversion of Potentials in Different Solvents. *J. Phys. Chem. B* **2010**, *114*, 7894−7899.
(28) Ramabhadran, R. O.; Raghavachari, K. Theoretical Thermochemistry for Organic Molecules: Development of the Generalized Connectivity-Based Hierarchy. *J. Chem. Theory Comput.* **2011**, *7*, 2094−2103.
(29) Raghavachari, K.; Maier, S.; Collins, E. M.; Debnath, S.; Sengupta, A. Approaching Coupled Cluster Accuracy with Density Functional Theory Using the Generalized Connectivity-Based Hierarchy. *J. Chem. Theory Comput.* **2023**, *19*, 3763−3778.
(30) Maier, S.; Thapa, B.; Raghavachari, K. G4 Accuracy at DFT Cost: Unlocking Accurate Redox Potentials for Organic Molecules Using Systematic Error Cancellation. *Phys. Chem. Chem. Phys.* **2020**, *22*, 4439−4452.
(31) Sengupta, A.; Raghavachari, K. Prediction of Accurate Thermochemistry of Medium and Large Sized Radicals Using Connectivity-Based Hierarchy (CBH). *J. Chem. Theory Comput.* **2014**, *10*, 4342−4350.
(32) Liu, J.; Wang, R.; Tian, J.; Zhong, K.; Nie, F.; Zhang, C. Calculation of Gas-Phase Standard Formation Enthalpy via Ring-Preserved Connectivity-Based Hierarchy and Automatic Bond Separation Reaction Platform. *Fuel* **2022**, *327*, 125203.
(33) Ramabhadran, R. O.; Sengupta, A.; Raghavachari, K. Application of the Generalized Connectivity-Based Hierarchy to Biomonomers: Enthalpies of Formation of Cysteine and Methionine. *J. Phys. Chem. A* **2013**, *117*, 4973−4980.
(34) Debnath, S.; Sengupta, A.; Raghavachari, K. Eliminating Systematic Errors in DFT via Connectivity-Based Hierarchy: Accurate Bond Dissociation Energies of Biodiesel Methyl Esters. *J. Phys. Chem. A* **2019**, *123*, 3543−3550.
(35) Collins, E. M.; Sengupta, A.; AbuSalim, D. I.; Raghavachari, K. Accurate Thermochemistry for Organic Cations via Error Cancellation Using Connectivity-Based Hierarchy. *J. Phys. Chem. A* **2018**, *122*, 1807−1812.
(36) Chan, B.; Collins, E.; Raghavachari, K. Applications of Isodesmic-type Reactions for Computational Thermochemistry. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *11*, No. e1501.
(37) Thapa, B.; Raghavachari, K. Accurate p $K_a$ Evaluations for Complex Bio-Organic Molecules in Aqueous Media. *J. Chem. Theory Comput.* **2019**, *15*, 6025−6035.
(38) Işık, M.; Levorse, D.; Rustenburg, A. S.; Ndukwe, I. E.; Wang, H.; Wang, X.; Reibarkh, M.; Martin, G. E.; Makarov, A. A.; Mobley, D. L.; Rhodes, T.; Chodera, J. D. pKa Measurements for the SAMPL6 Prediction Challenge for a Set of Kinase Inhibitor-like Fragments. *J. Comput. Aided Mol. Des.* **2018**, *32*, 1117−1138.
(39) Liao, C.; Nicklaus, M. C. Comparison of Nine Programs Predicting p $K_a$ Values of Pharmaceutical Substances. *J. Chem. Inf. Model.* **2009**, *49*, 2801−2812.
(40) Xiong, J.; Li, Z.; Wang, G.; Fu, Z.; Zhong, F.; Xu, T.; Liu, X.; Huang, Z.; Liu, X.; Chen, K.; Jiang, H.; Zheng, M. Multi-Instance Learning of Graph Neural Networks for Aqueous p *K* a Prediction. *Bioinformatics* **2022**, *38*, 792−798.
(41) *Molecular Operating Environment (MOE), 2022.02*; Chemical Computing Group ULC: 910−1010 Sherbrooke St. W., Montreal, QC H3A 2R7, Canada, 2023.
(42) Zeng, Q.; Jones, M. R.; Brooks, B. R. Absolute and Relative pKa Predictions via a DFT Approach Applied to the SAMPL6 Blind Challenge. *J. Comput. Aided Mol. Des.* **2018**, *32*, 1179−1189.
(43) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; 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, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16*, Revision C.01; Gaussian, Inc., 2016.
(44) Zherebtsov, D. *Verstack*; GitHub, Inc., 2024.Https://Github.Com/DanilZherebtsov/Verstack.
(45) Thapa, B.; Schlegel, H. B. Density Functional Theory Calculation of p $K_a$'s of Thiols in Aqueous Solution Using Explicit Water Molecules and the Polarizable Continuum Model. *J. Phys. Chem. A* **2016**, *120*, 5726−5735.
(46) Thapa, B.; Schlegel, H. B. Improved p $K_a$ Prediction of Substituted Alcohols, Phenols, and Hydroperoxides in Aqueous Medium Using Density Functional Theory and a Cluster-Continuum Solvation Model. *J. Phys. Chem. A* **2017**, *121*, 4698−4706.
(47) Thapa, B.; Schlegel, H. B. Theoretical Calculation of p $K_a$'s of Selenols in Aqueous Solution Using an Implicit Solvation Model and Explicit Water Molecules. *J. Phys. Chem. A* **2016**, *120*, 8916−8922.
(48) Ho, J. Predicting pKa in Implicit Solvents: Current Status and Future Directions*. *Aust. J. Chem.* **2014**, *67*, 1441−1460.
(49) Seybold, P. G.; Shields, G. C. Computational Estimation of pK $_a$ Values. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2015**, *5*, 290−297.
(50) Alongi, K. S.; Shields, G. C. Theoretical Calculations of Acid Dissociation Constants: A Review Article. In *Annual Reports in Computational Chemistry*; Elsevier, 2010; Vol. *6*, pp 113−138.
(51) Pan, X.; Wang, H.; Li, C.; Zhang, J. Z. H.; Ji, C. MolGpka: A Web Server for Small Molecule p $K_a$ Prediction Using a Graph-Convolutional Neural Network. *J. Chem. Inf. Model.* **2021**, *61*, 3159−3165.
(52) Wu, J.; Wan, Y.; Wu, Z.; Zhang, S.; Cao, D.; Hsieh, C.-Y.; Hou, T. MF-SuP-pKa: Multi-Fidelity Modeling with Subgraph Pooling Mechanism for pKa Prediction. *Acta Pharm. Sin. B* **2023**, *13*, 2572−2584.


https://doi.org/10.1021/acs.jcim.3c01923
*J. Chem. Inf. Model.* 2024, 64, 712−723

(53) Fraczkiewicz, R.; Lobell, M.; Göller, A. H.; Krenz, U.; Schoenneis, R.; Clark, R. D.; Hillisch, A. Best of Both Worlds: Combining Pharma Data and State of the Art Modeling Technology To Improve *in Silico* p $K_a$ Prediction. *J. Chem. Inf. Model.* **2015**, *55*, 389−397.
(54) Pracht, P.; Wilcken, R.; Udvarhelyi, A.; Rodde, S.; Grimme, S. High Accuracy Quantum-Chemistry-Based Calculation and Blind Prediction of Macroscopic pKa Values in the Context of the SAMPL6 Challenge. *J. Comput. Aided Mol. Des.* **2018**, *32*, 1139−1149.
(55) Johnston, R. C.; Yao, K.; Kaplan, Z.; Chelliah, M.; Leswing, K.; Seekins, S.; Watts, S.; Calkins, D.; Chief Elk, J.; Jerome, S. V.; Repasky, M. P.; Shelley, J. C. Epik: P $K_a$ and Protonation State Prediction through Machine Learning. *J. Chem. Theory Comput.* **2023**, *19*, 2380−2388.
(56) Yang, Q.; Li, Y.; Yang, J.; Liu, Y.; Zhang, L.; Luo, S.; Cheng, J. Holistic Prediction of the p $K_a$ in Diverse Solvents Based on a Machine-Learning Approach. *Angew. Chem.* **2020**, *132*, 19444−19453.
(57) *Jaguar, Version 7.5*; Schrödinger, LLC: New York, 2008.
(58) Bochevarov, A. D.; Harder, E.; Hughes, T. F.; Greenwood, J. R.; Braden, D. A.; Philipp, D. M.; Rinaldo, D.; Halls, M. D.; Zhang, J.; Friesner, R. A. Jaguar: A High-performance Quantum Chemistry Software Program with Strengths in Life and Materials Sciences. *Int. J. Quantum Chem.* **2013**, *113*, 2110−2142.
(59) pKa Plugin Ionization Equilibrium Partial Charge Distribution. https://docs.chemaxon.com/display/docs/pka-plugin.md (accessed 09 17, 2009).
(60) Dassault Systèmes. *Pipeline Pilot Data Analysis and Reporting Platform*, 2023. http://accelrys.com/products/scitegic/(accessed 09 17, 2009).
(61) CompuDrug. Latest Upgrades. http://www.compudrug.com (accessed 09 17, 2009).
(62) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A Software Program for pK a Prediction and Protonation State Generation for Drug-like Molecules. *J. Comput. Aided Mol. Des.* **2007**, *21*, 681−691.
(63) *Epik*, Schrodinger, LLC: New York, 2008.
(64) Sparc On-Line Calculator. http://ibmlc2.chem.uga.edu/sparc/ (accessed 09 17, 2009).
(65) Lee, P. H.; Ayyampalayam, S. N.; Carreira, L. A.; Shalaeva, M.; Bhattachar, S.; Coselmon, R.; Poole, S.; Gifford, E.; Lombardo, F. In Silico Prediction of Ionization Constants of Drugs. *Mol. Pharmacol.* **2007**, *4*, 498−512.
(66) Simulations Plus. http://www.simulations-plus.com (accessed 09 17, 2009).
(67) Predict Ionization Constant, Acid-Base Dissociation Constant, pKa, Experimental pKa. https://www.acdlabs.com/products/percepta-platform/physchem-suite/pka/#product_demo/ (accessed 09 17, 2009).
(68) Physicochemical and ADMET Laboratory. http://pharma-algorithms.com/ionization.htm (accessed 09 17, 2009).
(69) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.
(70) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.
(71) Gedeck, P.; Rohde, B.; Bartels, C. QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.* **2006**, *46*, 1924−1936.


https://doi.org/10.1021/acs.jcim.3c01923
*J. Chem. Inf. Model.* 2024, 64, 712−723