

# Epik: $pK_a$ and Protonation State Prediction through Machine Learning

*Ryne C. Johnston,<sup>\*,†</sup> Kun Yao,<sup>‡</sup> Zachary Kaplan,<sup>‡</sup> Monica Chelliah,<sup>‡</sup> Karl Leswing,<sup>‡</sup> Sean  
Seekins,<sup>†</sup> Shawn Watts,<sup>†</sup> David Calkins,<sup>†</sup> Jackson Chief Elk,<sup>†</sup> Steven V. Jerome,<sup>§</sup> Matthew P.  
Repasky,<sup>†</sup> John C. Shelley<sup>†</sup>*

<sup>†</sup>Schrödinger, Inc., 101 SW Main St., Suite 1300, Portland OR 97204, USA

<sup>‡</sup>Schrödinger, Inc., 1540 Broadway St., 24th Floor, New York, NY 10036, USA

<sup>§</sup>Schrödinger, Inc., 9171 Towne Centre Drive, San Diego, CA 92122, USA

**KEYWORDS**  $pK_a$ ; Epik; Acid dissociation; Protonation state; Machine learning; Graph  
Convolutional Neural Network

**ABSTRACT** Epik version 7 is a software program that uses machine learning for predicting the  $pK_a$  values and protonation state distribution of complex, drug-like molecules. Using an ensemble of atomic graph convolutional neural networks (GCNNs) trained on over 42,000  $pK_a$  values across broad chemical space from both experimental and computed origins, the model predicts  $pK_a$  values with 0.42 and 0.72  $pK_a$  unit median absolute and RMS errors, respectively, across seven test sets. Epik version 7 also generates protonation states and recovers 95% of the most populated protonation states compared to previous versions. Requiring on average only 47 ms per ligand,

Epik version 7 is rapid and accurate enough to evaluate protonation states for crucial molecules and prepare ultra-large libraries of compounds to explore vast regions of chemical space. The simplicity of and time required for the training allows for the generation of highly accurate models customized to a program's specific chemistry.

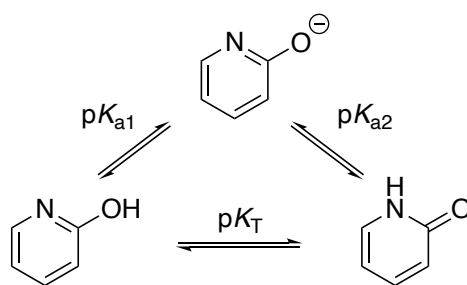
## INTRODUCTION

In solution, many molecules undergo ionization where a proton associates or dissociates. The equilibrium between associated (HA) and dissociated states ( $[H^+]$  and  $[A^-]$ ) is measured by the quantities  $pK_a$  and  $K_a$ .

$$pK_a = -\log\left(\frac{[A^-][H^+]}{[HA]}\right) = -\log K_a \quad (1)$$

In the associated state, the proton is tightly bound to specific titratable sites around the molecule. If a molecule has more than one titratable site, a proton can associate with any of them through prototropy with a specific microscopic  $pK_a$ , or  $\mu pK_a$  value. Together the  $\mu pK_a$  values from multiple sites contribute to the apparent, macroscopic  $pK_a$  or  $pK_a^{macro}$  values of the molecule.

**Scheme 1.** Thermodynamic cycle used to calculate  $pK_T$ .



$$pK_T = pK_{a2} - pK_{a1} \quad (2)$$

The various protonated forms of a molecule with the same net charge are referred to as tautomers or, more specifically, protomers. Some sites can dissociate more easily than others, and this influences the distribution of protomers which are all in equilibrium with one another. The

equilibrium between two prototropic tautomers ( $pK_T$ ) that are related by a common ionization can be calculated from the difference in  $\mu pK_a$  values (Scheme 1 & Eqn 2). In this way  $pK_a$  and  $pK_T$  values can be used to estimate the populations of different states at any pH value. This is important because it allows prediction of the most populated protonation states for structure-based drug design. There is a free energy cost associated with conversion of protonation states that isn't included explicitly in most structure-based methods such as docking and naive free energy perturbation (FEP) calculations which assume input of the lowest energy state.<sup>1</sup> So input of a higher energy ligand state will result in misprediction of binding affinity and/or prediction of a binding pose that isn't reflected by experimental observation.<sup>2</sup>

The pH dependence of the state distribution can affect a number of molecular properties<sup>3,4</sup>, including solubility, membrane permeability, and protein binding affinity. Because  $pK_a$  affects key molecular properties it is a critical quantity to know for drug design. Several software applications utilizing different approaches have been developed to address the need of predicting  $pK_a$  values for molecules, and some also predict state populations. One popular approach is the empirical Hammett-Taft (HT) method using linear free energy relationships to predict  $pK_a$ . Because this trained method can have thousands of finely-tuned parameters, high accuracy ( $< 0.5$   $pK_a$  units) can be achieved.<sup>5</sup> Popular programs using the HT method include ACD/Labs' ACD/ $pK_a$  module<sup>6</sup> and earlier versions of Schrödinger's Epik<sup>7,8</sup>. The  $pK_a$  Plugin<sup>9</sup> from ChemAxon uses empirically calculated partial charges and other parameters to predict  $pK_a$  values<sup>10,11</sup>.  $pK_a$  values may also be calculated to high accuracy through quantum mechanical (QM) computation coupled with empirical fitting, as is done by Schrödinger's Jaguar<sup>12</sup>  $pK_a$ ,<sup>13,14</sup> which has a mean absolute deviation of 0.49  $pK_a$  units<sup>15</sup>. MoKa<sup>16</sup> from Molecular Discovery uses a novel QSPR method based on molecular interaction fields to predict  $pK_a$  values. The COSMO-RS  $pK_a$  method<sup>17,18</sup> uses QM

calculations coupled with statistical thermodynamics to more realistically include solvation effects than traditional implicit solvation models.

Within the past few years, machine learning (ML) has found use in predicting  $pK_a$  values. An example is S+p $K_a$ <sup>19</sup> from SimulationsPlus, which uses a series of models constructed from artificial neural network ensembles to predict both  $pK_a$  values to within 0.5  $pK_a$  units and the distribution of microstates. Li, et al.<sup>20</sup> used radial basis function artificial neural networks with a particle swarm optimization algorithm to predict  $pK_a$  values for a small number of neutral and basic drugs with MAE of 0.31. Roszak et al.<sup>21</sup> estimated  $pK_a$  values in DMSO and reactivities of carbon acids using atomic graph convolutional neural networks<sup>22</sup> (GCNNs). In their report, they use only the local topological neighborhood out to 4 bonds away, as well as Gasteiger partial charges to reach accuracies of 2.1  $pK_a$  units. Baltruschat and Czodrowski published<sup>23</sup> an RDKit descriptor-based random forest method for predicting  $pK_a$  values of monoprotated molecules and reported an accuracy of 0.68  $pK_a$  units.

We report the new program Epik v 7, ML software for predicting  $pK_a$  values and populations of protonation states for molecules. It is the successor to Epik, hereafter referred to as Epik Classic. Epik v 7 is based on atomic GCNNs that use a minimal set of rapidly computed atomic descriptors and the local topological neighborhood for short-range chemical perturbation. Epik v 7 can predict  $pK_a$  values and the populations of states at a target pH in aqueous solution. We demonstrate that Epik v 7 is more accurate than its progenitor at up to double the speed. In addition to accuracy, speed is an important factor for  $pK_a$  and state prediction software, as the current demand is growing to include screening billions of compounds.<sup>24,25</sup> This volume is expected to grow rapidly in the coming years as larger enumerated libraries come online. Epik v 7 must be both rapid and generalizable enough to handle the diversity of chemical space in such vast enumerations.

In addition to being generalizable, we also recognize the desire for customizability. Many organizations have large internal libraries of  $pK_a$  data on proprietary compounds, which we hope to leverage to generate highly-accurate local models around chemistries more pertinent to the user.

In the following sections we will discuss the Epik v 7 methodology, training, accuracy, and performance.

## METHODS

Epik v 7 can perform two main types of calculations,  $pK_a$  query and state prediction. In the query calculation,  $pK_a$  values are predicted for all titratable sites of the input molecule. In the state prediction calculation, Epik v 7 enumerates potential states, predicts their  $pK_a$  values, and estimates the populations for all highly populated microstates at a supplied pH. There are three stages to any Epik v 7 calculation: 1) ionization state and tautomer network construction, 2)  $pK_a$  prediction, 3) microstate population estimation.

**Microstates Network Construction.** Upon entering Epik v 7, an input molecule is first converted to an RDKit molecular graph and then has its titratable sites identified by matching against a list of 74 acidic and basic SMARTS patterns. These rules were manually curated to cover most of the traditionally ionizable sites under aqueous conditions near standard temperature and pressure. Any hydrogens directly bound to these sites are considered both labile and equivalent and will be the ones that are redistributed during tautomer enumeration. The initial list of tautomers at the input ionization state ( $q_0$ ) are generated from the combination of the titratable sites and the labile hydrogens. Because the number of tautomers considered grows quickly with the number of titratable sites, each titratable site is assigned a priority to filter which tautomers are processed. High priority sites are always taken, medium priority sites are taken if the total number of high

and medium sites is 10 or below, and low priority sites are taken only if the total number of high, medium and low sites is 10 or below.

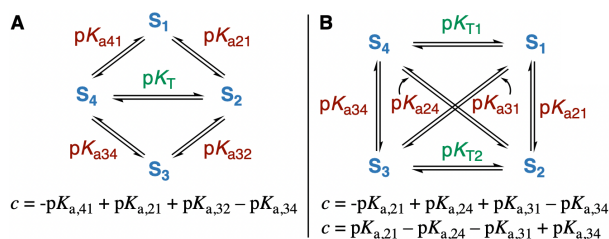
Tautomers are generated by moving explicit labile hydrogens to different sites around the molecular graph of heavy atoms stripped of hydrogens. Each tautomer then has its Lewis structure adjusted to accommodate the new hydrogens and then the relevant resonance structures are enumerated by charge cancellation and resonance adjustment, if needed. Finally, high-energy tautomers, e.g., structures with multiple separated opposite charges, are filtered out, reducing the overall number of structures whose  $pK_a$  are evaluated. To generate different ionization states, hydrogens are either added to ( $q_i > q_0$ ) or removed from ( $q_i < q_0$ ) consideration as labile. In practice, we find that the default  $q_0 \pm 2$  usually captures the vast majority of the experimentally relevant ionizations, although these bounds may be modified from the command line or Maestro panel. Thus, the tautomers across the available ionization levels form the microstates network.

**$pK_a$  Prediction.**  $pK_a$  values are predicted site-wise given the atom indices of the donors and the RDKit molecule to which they belong. Predictions may be made with either the single best ML model based on its 95% confidence interval  $t^*$ -scores or an ensemble of the three best models. To improve the stability of our predictions, we utilize a five-fold cross-validation ensemble of the top three ML models. Predictions were made by evaluating each of the five folds with the top models and taking the Olympic mean. The standard deviation of the ensemble is taken as the  $pK_a$  uncertainty. The predicted  $pK_a$  values are finally mapped back onto the parent states in the microstates network. Although using the single best model is faster the results are usually less accurate than the ensemble<sup>26</sup> and precludes the calculation of ensemble-based uncertainties.

**Microstates Populations Estimation.** Obtaining the microstates' populations first starts with calculating the pH-dependent populations in a self-consistent manner using thermodynamic

cycles. Pairs of microstates within the same ionization state (i.e., tautomers) in the network are connected by  $pK_T$  values, which can be calculated from the  $pK_a$  values to common ionized states (Scheme 1).

**Scheme 2.** Thermodynamically consistent cycles and consistency relations used to calculate  $pK_T$  values.



Experimentally, the  $pK_T$  value between a tautomer pair is invariant to the thermodynamic cycle used. In practice,  $pK_a$ s are usually predicted in isolation from the entire thermodynamic landscape, and therefore are not self-consistent.<sup>27</sup> To correct for the  $pK_a$  inconsistencies, we calculate  $pK_a$  weights using two types of four-membered thermodynamic cycles (Scheme 2), the absolute sum of which,  $c$ , is a measure of consistency, where  $c = 0$  is perfectly consistent. The  $pK_{a,j}$  weight,  $w_j$ , is determined by the number of consistency relations involving  $pK_{a,j}$  that are poorly met ( $c \geq 1$ ) versus those that are well met ( $c < 1$ ).

An  $N \times N$  microstates coefficients matrix  $C$  is compiled from the  $pK_a$  weights, and an  $N \times 1$  matrix  $V$  is compiled statewise from the  $\Delta pK_{a,j}$  (i.e.,  $pK_{a,j} - pH_{ref}$ ) weighted by  $w_j$ . For convenience, we work with  $u_i$  values from the pH-dependent state populations  $p_i$ :

$$u_i = -\log_{10}(p_i). \quad (3)$$

These  $u_i$  values are solved collectively in the  $N \times 1$  matrix  $U$  collectively via

$$V = C \cdot U. \quad (4)$$

The  $u_i$  values are finally converted back to  $p_i$  populations, which are then normalized and converted to state penalties to be reported as molecular properties.

Once the pH-dependent populations of all the microstates have been estimated, the microspecies across all ionization states can be ranked according to their populations and returned. The default population threshold ( $p_{\text{thresh}}$ ) to return microstates is  $p_{\text{thresh}} = 0.1$ , but this can be changed by the user. The cutoff can also be expressed in terms of a pH tolerance,  $\Delta\text{pH}$ , as

$$p_{\text{thresh}} = 10^{-\Delta\text{pH}}. \quad (5)$$

This allows the user to define a pH window ( $\text{pH}_{\text{ref}} \pm \Delta\text{pH}$ ) in which to search. This pH tolerance setting is by default 1.0 as per eqn. 5 with the default  $p_{\text{thresh}} = 0.1$ , but it can similarly be changed by the user.

The pH-independent populations are obtained as the ratio of a single state's pH-dependent population over all the pH-dependent populations for the charge state  $q$  the state belongs to, also known as the charge population factor ( $f_q$ ).

$$p_i^{\text{pH-indep.}} = \frac{p_i^{\text{pH-dep.}}}{\sum_{j=1}^{n_q} p_j^{\text{pH-dep.}}} = \frac{p_i^{\text{pH-dep.}}}{f_q} \quad (6)$$

Finally, macroscopic  $\text{p}K_a$ s are calculated between pairs of successive charge states, e.g.,  $q = +1$  &  $q = 0$  using the  $[\text{H}^+]$  at the reference pH (via  $[\text{H}^+] = 10^{-\text{pH}_{\text{ref}}}$ ) in the following equation:

$$\text{p}K_a^{\text{macro}, q+1-q} = -\log_{10} \frac{f_q([\text{H}^+])}{f_{q+1}} \quad (7)$$

The collection of all macroscopic  $\text{p}K_a$  values spanning the queried charge levels are added as a molecular property on all returned structures.

We have also adapted from Macro- $\text{p}K_a$ <sup>28</sup> a feature that produces a single ligand  $\text{p}K_a$  report as an HTML page (e.g., Figure S6). The report includes the major contributing species spanning



the queried charge levels, their pH-dependent populations at the queried pH, their pH-independent populations, the macro- $pK_a$ s, and a speciation diagram.

**Datasets.** The  $pK_a$  data (42,890 total values) contain both mono- and polyprotic compounds and were collected from five sources of which roughly half are experimental and the other half are calculated from quantum mechanics using Jaguar  $pK_a$ : 1) 4,326 experimental values from the Epik Classic training set<sup>7</sup>, itself compiled from various experimental sources; 2) a curated selection of 13,698 experimental values from the pKaData database compiled by IUPAC<sup>29</sup>; 3) 9,920 Jaguar  $pK_a$  calculated values from the Epik Classic training and IUPAC sets for “off”-sites, i.e., titratable sites that are not the primary contributor to the experimentally observed  $pK_a$  and whose protonated forms are minor tautomers; 4) Jaguar  $pK_a$  calculated values for an enumeration of 11,432 heterobicycles of 11 heavy atoms or fewer from the GDB-13 dataset<sup>30</sup> (see Supporting Information for methods); and 5) 3,514 additional values for refinement either calculated with Jaguar or compiled from the literature.

The structures in the experimental datasets were adjusted to what we determined to be the single most populated tautomeric form that contributes to the observed  $pK_a$ . By including the “off”-sites we intended to train the model to be able to predict the populations of minor tautomers with reasonable accuracy. Inclusion of the GDB13 heterocycles expands the chemical diversity of the model to novel scaffolds, to a total of 7,759 Bemis-Murcko (BM) scaffolds. Overall, Jaguar  $pK_a$  calculations complement the experimental values to extend model coverage (Fig. S2) and improve the accuracy for predicting experimental values (Fig. S3).

From the Epik Classic training and the additional, non-IUPAC experimental datasets, a random sample of 1,031  $pK_a$  values was withheld for validation, and the remainder was used for training.

It is important to note that an external set for validation was never held out from Epik Classic, which was trained on the molecules in this validation set and thus will show bias.

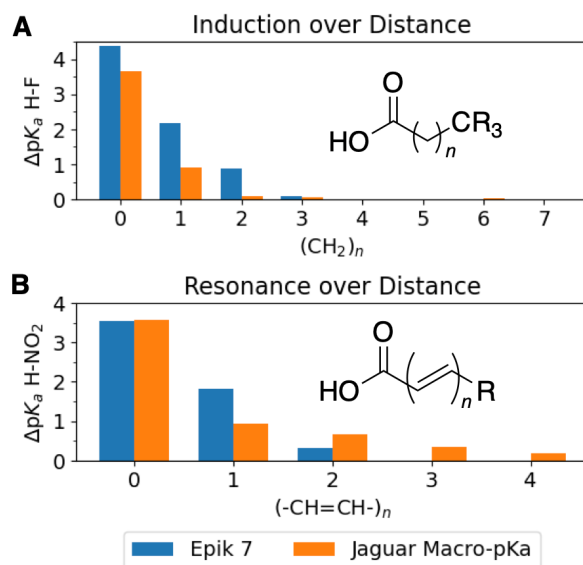
The datasets show a bimodal distribution of  $pK_a$  values with two means at  $\sim 3$  and  $\sim 9$ , owing to a predominance of carboxylic acids and basic amines, respectively. While the training set has nearly twice as many entries of  $pK_a \sim 3$  than  $pK_a \sim 9$ , the validation set is more balanced (Fig. S1). Whereas the training set had 7,559 BM scaffolds, there are 874 BM scaffolds across all of the test sets, of which 515 are unique to the test sets. That over half of the scaffolds in the test sets are unique provides a rough indication of how well Epik v 7 is expected to handle chemistry it has not been trained on.

**Model Training.** The Epik v 7 model is built on the DeepAutoQSAR framework<sup>31,32</sup>, a machine learning platform based upon the DeepChem<sup>33</sup> package. The approach underlying Epik v 7 is the atomic GCNN, where the model behaves as a fingerprint of the local subgraph of the molecular graph centered about an ionizing atom, conceptually similar to Morgan fingerprints.<sup>34</sup>

The automated training mechanism in DeepAutoQSAR enables a mostly hands-off approach, requiring only the maximum model search time to be specified by the user before launching a training session. DeepAutoQSAR first rapidly featurizes the molecular graph with minimal, mostly one-hot atomic features, e.g., element, hybridization, formal charge, etc. as calculated by RDKit. DeepAutoQSAR then automatically samples hyperparameter space, including number of training epochs, normalization scheme, model algorithm, and layer depth.

We found optimal accuracy on our training set for atomic GCNNs with a layer depth ( $D$ ) of three, corresponding to a topological neighborhood of out to  $2D$  (six) bonds away from the ionizing site (Figure 1). This regime models  $pK_a$  as being a function of the local chemical environment

only, similar to Roszak's approach but with graph pooling between layers. Such a model has a limitation in that remote effects are not taken into account.



**Figure 1.** Inductive substituent effects (A) diminish faster than resonance (B). In Epik v 7, substituent effects cut off at > 6 bonds between the acidic oxygen and the R atoms. The bars in the induction plot represent moving out from the carboxylic acid by one methylene carbon, but those in the resonance plot move out by two vinylic carbons. Directly compare the plots by doubling the resonance plot's x-axis values.

However, for many cases, substitution effects attenuate rapidly or gradually with distance ( $r$ ) from the ionizing site depending on the type of effect. Whereas induction can diminish more rapidly than either  $1/r$  or  $1/r^2$  for charged or dipolar substituents<sup>35</sup>, respectively, resonance effects drop off much more slowly. For example, we compared the  $pK_a$  values of trifluoromethyl- ( $R = CF_3$ ) and methyl-terminated ( $R = H$ ) carboxylic acids of increasing lengths to measure induction (Figure 1, A). After  $n = 4$  methylene units, Epik v 7 reports the exact same  $pK_a$  values regardless of the terminal substituent because it is too far away from the ionizing site to be registered, but at this distance the perturbing effect is already negligible as can be seen in the Macro- $pK_a$  predictions,

which do take into account all the atoms in the molecule. We measured resonance substituent effects similarly with nitrovinyl- ( $R = \text{NO}_2$ ) and vinyl-terminated ( $R = \text{H}$ ) carboxylic acids (Figure 1, B). The Jaguar predictions show a longer tail, where substituents can perturb out to at least ten bonds away through resonance. The six-bond cutoff in Epik v 7 results in at most a 0.33  $\Delta\text{p}K_a$  unit error for a group right outside this range, but decreases with distance. The cutoff was chosen to strike a balance between accuracy, training time and training set size, the latter two of which increase at larger  $D$  (Table S1).

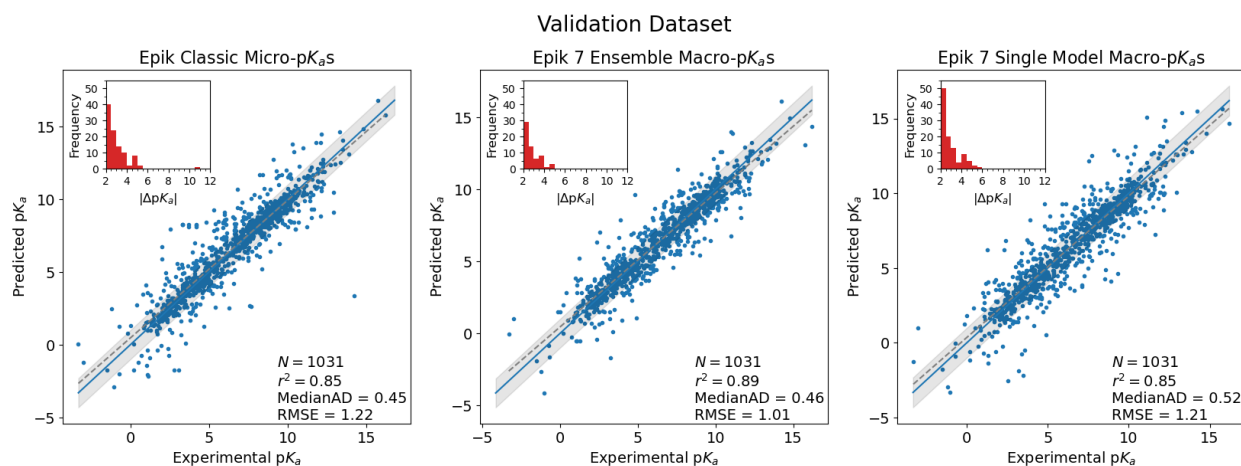
We attempted to introduce long-range effects by incorporating an additional “master” atom<sup>36</sup> that is connected to all atoms, but this resulted in a non-predictive model (Figure S5). The regression is presumably due to the fact that this approach puts all atoms on equal footing, and with  $D = 1$ , leading to an exaggeration of remote effects. Computing and including Gasteiger charges to account for remote effects as Roszak does<sup>21</sup> was not fast enough for our application.

The Epik v 7 training set is simply a collection of structures protonated at the site of ionization, along with the atom index of the site of ionization, and the reference  $\text{p}K_a$  value. To improve Epik v 7 accuracy for a specific chemotype, one simply needs to add additional representative structures to the training set and retrain.

## RESULTS AND DISCUSSION

**Accuracy.** We first evaluated Epik v 7 on the validation set for both the ensemble and the single-model modes and compared them to Epik Classic (Figure 2). Because Epik Classic has already been trained on most of the validation set, it is expected to and indeed does perform quite well, with a median absolute error of 0.45 and an RMSE of 1.23  $\text{p}K_a$  units. By comparison, Epik v 7 ensemble predictions find a median absolute error of 0.46 and RMSE of 1.01  $\text{p}K_a$  units, similar to the accuracy of Epik Classic though with fewer outliers ( $>2 \text{p}K_a$  units) leading to a smaller RMSE.

The single-model mode fares slightly worse but is still acceptable for some applications, with a median absolute error of 0.52 and RMSE of 1.21  $pK_a$  units.



**Figure 2.** Comparing Epik Classic, Epik v 7 ensemble mode, and Epik v 7 single-model mode on the validation set. Frequency of absolute errors  $\geq 2$   $pK_a$  units for each method are inset. Note that Epik Classic was trained on this data, and so this does not represent a true test set for that version of Epik.

We additionally tested a number of publicly-available datasets, (Table 1). Some of the publicly-available sets contain molecules that were already in the Epik v 7 training sets, and so these molecules were excluded from evaluation. However, as with the validation set, many of these data were already included in Epik Classic’s training set and so can’t be considered true prospective tests for that program. The Epik v 7 ensemble model on average had a median absolute error of 0.42 and RMSE of 0.72  $pK_a$  units, outperforming Epik Classic in most cases. Furthermore, the standard deviations of the MAD and RMSE are much lower in Epik v 7 ensemble (0.13 and 0.16, respectively) than in Epik Classic (0.74 and 0.76, respectively) indicating improved consistency.

**Table 1.** Comparison of Epik Classic and different modes of Epik v 7 on various publicly available test sets.

Dataset	$N$	Epik Classic			Epik v 7 Ensemble			Epik v 7 Single Model		
		$r^2$	MAD	RMSE	$r^2$	MAD	RMSE	$r^2$	MAD	RMSE
Validation	1,031	0.85	0.45	1.23	0.89	0.46	1.01	0.85	0.53	1.21
AstraZeneca <sup>37</sup>	243	0.85	0.37	0.83	0.87	0.44	0.79	0.83	0.48	0.92
Manchester <sup>38</sup>	142	0.96	0.29	0.48	0.86	0.61	0.80	0.87	0.37	0.80
Morgenthaler <sup>39</sup>	43	0.53	2.34	2.73	0.94	0.21	0.62	0.94	0.24	0.72
Novartis <sup>40</sup>	152	0.96	0.23	0.55	0.97	0.31	0.54	0.93	0.34	0.74
SAMPL6 <sup>41</sup>	31	0.91	0.52	0.84	0.95	0.39	0.61	0.89	0.56	0.92
Vertex <sup>42</sup>	51	0.74	0.59	1.17	0.92	0.49	0.68	0.89	0.52	0.86
<i>Total/Average</i>	<i>1,693</i>	<i>0.85</i>	<i>0.68</i>	<i>1.08</i>	<i>0.91</i>	<i>0.42</i>	<i>0.72</i>	<i>0.89</i>	<i>0.43</i>	<i>0.88</i>

**Comparison to Other Methods.** The SAMPL6 set of 31  $pK_a$  values were excluded from the training and validation sets. Had the Epik v 7 ensemble model been submitted to the final results<sup>36</sup>, it would have ranked first by both MeanAE and RMSE (Table 2) ahead of Grimme’s rigorous QM with linear fitting submission<sup>43</sup>, S+ $pK_a$ , ACD/ $pK_a$ , MoKa, and Epik Classic.

**Table 2.** Top ranked submissions<sup>44</sup> for the SAMPL6 set.

Rank	Method	ID	RMSE	MeanAE
1	Epik v 7 ensemble	N/A	0.61	0.48
2	Grimme	xvxzd	0.68	0.58
3	S+pK <sub>a</sub>	gyuhx	0.73	0.59
4	ACD/pK <sub>a</sub>	xmyhm	0.79	0.56
5	MoKa	nb017	0.94	0.77
6	Epik Classic	nb007	0.95	0.78

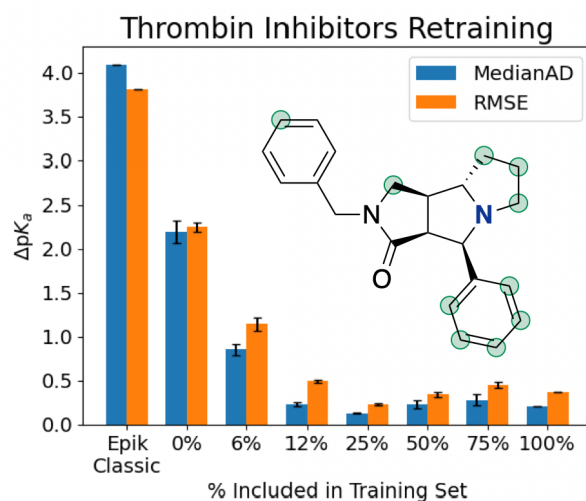
From AstraZeneca’s dataset of pK<sub>a</sub> values on publicly disclosed compounds we took a subset of 243 of the strongest basic values which has previously been evaluated with ChemAxon version Fermium.<sup>44</sup> ChemAxon had a median absolute error of 0.54 and RMSE of 1.11 pK<sub>a</sub> units. By contrast, Epik v 7 ensemble was more accurate, with a MAD of 0.44 and RMSE of 0.79 pK<sub>a</sub> units.

We also evaluated the 142 pK<sub>a</sub> values spanning 85 molecules in the supplementary information of Manchester’s original work<sup>38</sup> with Epik v 7. Epik Classic outperforms Epik v 7 ensemble with mean absolute errors of 0.37 and 0.56 pK<sub>a</sub> units, respectively. Although Epik Classic was on par with S+pK<sub>a</sub> v7, which achieved a mean absolute error of 0.41 on the same set<sup>19</sup>, Epik Classic was explicitly and heavily trained on this dataset.<sup>45</sup> While Epik v 7 was trained on molecular fragments similar to those in the Manchester set, none of the Manchester molecules in full are members of the training set. Unfortunately, the Epik v 7 model performs worse than both Epik Classic and S+pK<sub>a</sub> on this dataset.

**Model Customization.** Because many medicinal chemistry programs focus on the diversity within a narrow chemical space, it would be advantageous to have an Epik v 7 model trained to

the specific program chemistry. In these cases, we can train a new model by adding to the full training set either a portion of the existing data or synthetic data from Jaguar  $pK_a$  calculations.

*Retraining on existing data.* The Morgenthaler set contains several examples of the tricyclic thrombin inhibitors in Figure 3. The amine  $pK_a$  value is very sensitive to intramolecular forces and can be modulated by different substituents around the molecule to vary from between  $< 2$  to 7 (Fig. S16). Since there exist across the training and test sets<sup>39,42</sup> a total of 84 of these thrombin inhibitors, we considered this series to be representative of a small drug design program and to use them to demonstrate the ability to train a custom model to a particular chemotype.



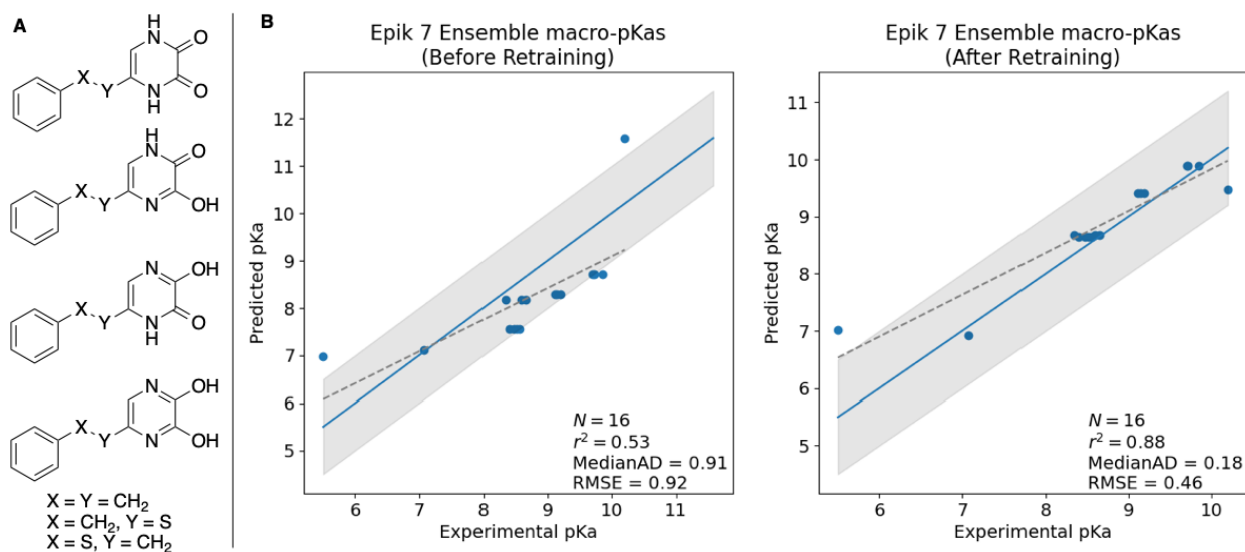
**Figure 3.** The  $pK_a$  values of the amine in 84 tricyclic thrombin inhibitors are highly tunable depending on the substituents at various points (green circles) around the molecule. Errors from the ensemble model with different amounts of examples included in the training set.

To get a baseline performance, we first removed all examples of this tricyclic core from the training set and then retrained the ensemble model. Next, we held out 20% of the removed structures for testing and then retrained a series of models by adding increasing amounts of the remaining thrombin inhibitors to reduce the error to a reasonable level. Tranches were selected such that their  $pK_a$  value distributions mirrored that of the entire set.



The absence of this chemotype from the baseline model results in an erroneous prediction on the holdout set, with a median absolute error of 2.25 and RMSE of 2.28  $pK_a$  units. Starting with adding just 6% of the training examples, we already see a halving of the error. This halving continues as we increase the number of examples all the way to 25%, where we see a sizable reduction in median absolute error and RMSE (Figure 3) down to 0.13 and 0.23  $pK_a$  units, respectively. This does appear to be a slight minimum, however, because further increasing the number of examples erodes both MAD and RMSE slightly. A second series of models was trained with a different random selection with a  $pK_a$  distribution matching that of the full dataset, and again we found a slight minimum at 25% inclusion. We suspect that this is due to the fact that the extrema have already been well described by this point, and additional data introduces noise to the model. This result suggests that, at least for sets covering narrow chemical space, a training set should be representative, but not overly so.

*Retraining on synthetic data.* As an illustrative example of training on calculated data, we took an internal program containing a challenging tautomerizable heterocycle core<sup>46</sup> where we thought we might be able to improve accuracy with a custom model. We first performed Jaguar  $pK_a$  calculations on the three unfunctionalized variants of the core to obtain the  $\mu pK_a$  values for each of their four tautomers (Figure 4, A). Because these compounds are not members of the test set, we did not need to take a holdout. To the full training set we added these 12 structures and retrained the model for 48 h. This exercise is relatively straightforward and requires little expertise. The longest parts of the process are the retraining and the Jaguar calculations. The improvement over the default model is clear (Figure 4, B). Whereas the default Epik v 7 ensemble model has a median absolute error of 0.91  $pK_a$  units, the custom model has an error of 0.18  $pK_a$  units while the RMSE improved by a factor of 2.



**Figure 4.** A custom Epik v 7 model was trained to the specific chemistry of an internal drug discovery program. A) The four tautomers added to the training set; B) Comparing the Epik v 7 ensemble models before and after custom retraining.

**Performance.** To test how Epik v 7 will perform in real-world screening we performed a state prediction calculation at  $\text{pH } 7.4 \pm 2.0$  with a charge range of  $\pm 2$  for a batch of 10,000 randomly chosen drug-like molecules from the Enamine REAL database<sup>47</sup>. On a single 3.50 GHz Intel<sup>®</sup> Xeon<sup>®</sup> processor and an NVIDIA<sup>®</sup> GeForce<sup>®</sup> GTX 1080 GPGPU, Epik v 7 takes 0.047 s to predict the  $\text{pK}_a$  values and state populations in the single-model mode, and in ensemble mode the speed is 0.072 s per molecule. By comparison, Epik Classic has an average speed of 0.106 s per molecule on the same hardware (although Epik Classic does not use the GPGPU). Additionally, Epik v 7 and Epik Classic agree on the most populated state in 95% of the cases. And in over half of the remaining cases, Epik v 7 recovers Epik Classic's top state with a population  $\geq 10\%$ .

**Limitations.** In addition to the aforementioned neglect of remote effects of atoms separated by  $> 6$  bonds, Epik v 7, like Epik Classic before it, does not explicitly take into account non-covalent and conformational effects, e.g., conformation dependent hydrogen bonding, shielding, etc.

However, in some cases, these effects may be implicitly included in a conformationally averaged way resulting from experiments or calculations<sup>15</sup>. Likewise, the Epik v 7 model also does not consider stereochemistry in its  $pK_a$  calculation, and thus different stereoisomers will all return the same  $pK_a$  value. Such stereochemistry dependent  $pK_a$  effects are rare and are usually the result of underlying conformational or hydrogen-bonding effects (e.g., Chart S2A).

By default, Epik v 7 enumerates the ionization states  $\pm 2$  net charge levels with respect to the input charge, which for some molecules would not be sufficient to locate the most populated state at the queried pH, e.g., Chart S2B. The range of ionization states considered in enumeration can be modified by the user, but at the expense of processing time.

## CONCLUSIONS

In this work, we disclosed Epik v 7, a new ML-based software program for predicting  $pK_a$  values and protonation state distributions of small, druglike molecules. Our atomic GCNN approach trained on over 42,000  $pK_a$  values produces a robust ML model that accurately captures the local chemistry surrounding an ionizing site out to six bonds away. Protonation state populations were calculated from the predicted  $pK_a$ s using a self-consistent method.

We demonstrated across almost 1,700 publicly available compounds contained within seven different datasets that Epik v 7 ensemble performs competitively to Epik Classic and other commercial programs at predicting  $pK_a$  values. We also showed two approaches to customize a highly accurate local model through additional training on a minimal set of druglike ligands, either by using known experimental values or by computing values through physics-based approaches, like with Jaguar  $pK_a$  or Macro- $pK_a$ .

Beyond standalone usage, which has been the focus of this report, Epik v 7 is a critical supporting technology in Schrödinger's ligand preparation workflows, such as LigPrep<sup>48</sup> and the Protein Prep

Workflow<sup>49</sup>. We have also integrated it as a computational model in our informatics platform LiveDesign<sup>50</sup>. Overall, Epik v 7 is not only a replacement for, but an improvement over, Epik Classic in most cases.

## **ASSOCIATED CONTENT**

### **Supporting Information.**

The following files are available free of charge.

Composition of the training and validation sets; details on the methods, including the Macro-pKa approach, the effects of varying model layer depth, and the use of a “master” atom; an example speciation report; details on the approach used to obtain Epik Classic and Epik v 7 values for the test sets; additional results on results on the test sets; additional details on the similarity between the training set versus the test sets (PDF)

Archive of the Epik v 7 results in CSV form separated by test set (ZIP)

## **AUTHOR INFORMATION**

### **Corresponding Author**

\*Email: [ryne.johnston@schrodinger.com](mailto:ryne.johnston@schrodinger.com)

### **ORCID**

Ryne C. Johnston: 0000-0002-6606-9401

Kun Yao: 0000-0003-2032-7441

Steven V. Jerome: 0000-0001-7510-9682

Matthew P. Repasky: 0000-0002-0259-7053

John C. Shelley: 0000-0001-6223-4804

## **Author Contributions**

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

## **Notes**

The authors declare the following competing financial interest(s): all authors are Schrödinger employees and hold financial interests in the company. Epik Classic & Epik v 7 are products sold by Schrödinger, LLC.

## **ACKNOWLEDGMENT**

RCJ wishes to thank Art D. Bochevarov and Mark A. Watson for their insights into macro- $pK_a$  and tautomerism as well as for their development of Macro- $pK_a$ . RCJ also appreciates Mats A. Svensson for helpful discussions on medicinally relevant chemistry.

## **ABBREVIATIONS**

BM, Bemis-Murcko; GCNN, graph convolutional neural network; GPGPU, general-purpose graphical processing unit; HT, Hammett-Taft; ML, machine learning; QM, quantum mechanics; QSPR, quantitative structure-property relationship; MAE, median absolute error; RMSE, root mean square error; SDF, structure-data file; SMARTS, SMILES arbitrary target specification where SMILES is Simplified molecular-input line-entry system

## **REFERENCES**

(1) de Oliveira, C.; Yu, H. S.; Chen, W.; Abel, R.; Wang, L. Rigorous Free Energy Perturbation Approach to Estimating Relative Binding Affinities between Ligands with Multiple Protonation and Tautomeric States. *J. Chem. Theory Comput.* **2019**, *15* (1), 424–435. <https://doi.org/10.1021/acs.jctc.8b00826>.

(2) Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C. Towards the Comprehensive, Rapid, and Accurate Prediction of the Favorable Tautomeric States of Drug-like Molecules in Aqueous Solution. *J. Comput. Aided Mol. Des.* **2010**, *24* (6), 591–604. <https://doi.org/10.1007/s10822-010-9349-1>.

(3) Katritzky, A. R.; Hall, C. D.; El-Gendy, B. E.-D. M.; Draghici, B. Tautomerism in Drug Discovery. *J. Comput. Aided Mol. Des.* **2010**, *24* (6), 475–484. <https://doi.org/10.1007/s10822-010-9359-z>.

(4) Pospisil, P.; Ballmer, P.; Scapozza, L.; Folkers, G. Tautomerism in Computer-Aided Drug Design. *J. Recept. Signal Transduct.* **2003**, *23* (4), 361–371. <https://doi.org/10.1081/RRS-120026975>.

(5) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. *pK<sub>a</sub> Prediction for Organic Acids and Bases*; Chapman and Hall: London, 1981.

(6) ACD/pK<sub>a</sub>. <https://www.acdlabs.com/products/percepta-platform/physchem-suite/pka/> (accessed 2021-12-17).

(7) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A Software Program for pK<sub>a</sub> prediction and Protonation State Generation for Drug-like

Molecules. *J. Comput. Aided Mol. Des.* **2007**, *21* (12), 681–691. <https://doi.org/10.1007/s10822-007-9133-z>.

- (8) Schrodinger Release 2022-4: Epik, 2022.
- (9) pK<sub>a</sub> Plugin. <https://chemaxon.com/> (accessed 2021-12-17).
- (10) Szegezdi, J.; Csizmadia, F. Prediction of Dissociation Constant Using Microconstants. In *227th American Chemical Society National Spring Meeting*; Anaheim, California, 2004.
- (11) Szegezdi, J.; Csizmadia, F. A Method for Calculating the pK<sub>a</sub> Values of Small and Large Molecules. In *233rd American Chemical Society National Spring Meeting*; Chicago, Illinois, 2007.
- (12) Schrodinger Release 2022-4: Jaguar, 2022.
- (13) Klicić, J. J.; Friesner, R. A.; Liu, S.-Y.; Guida, W. C. Accurate Prediction of Acidity Constants in Aqueous Solution via Density Functional Theory and Self-Consistent Reaction Field Methods. *J. Phys. Chem. A* **2002**, *106* (7), 1327–1335. <https://doi.org/10.1021/jp012533f>.
- (14) Bochevarov, A. D.; Watson, M. A.; Greenwood, J. R.; Philipp, D. M. Multiconformation, Density Functional Theory-Based pK<sub>a</sub> Prediction in Application to Large, Flexible Organic Molecules with Diverse Functional Groups. *J. Chem. Theory Comput.* **2016**, *12* (12), 6001–6019. <https://doi.org/10.1021/acs.jctc.6b00805>.
- (15) Yu, H. S.; Watson, M. A.; Bochevarov, A. D. Weighted Averaging Scheme and Local Atomic Descriptor for pK<sub>a</sub> Prediction Based on Density Functional Theory. *J. Chem. Inf. Model.* **2018**, *58* (2), 271–286. <https://doi.org/10.1021/acs.jcim.7b00537>.

(16) Milletti, F.; Storchi, L.; Sforza, G.; Cruciani, G. New and Original  $pK_a$  Prediction Method Using Grid Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, *47* (6), 2172–2181. <https://doi.org/10.1021/ci700018y>.

(17) Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M. E. 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* **2003**, *107* (44), 9380–9386. <https://doi.org/10.1021/jp034688o>.

(18) Eckert, F.; Diedenhofen, M.; Klamt, A. Towards a First Principles Prediction of  $pK_a$ : COSMO-RS and the Cluster-Continuum Approach. *Mol. Phys.* **2010**, *108* (3–4), 229–241. <https://doi.org/10.1080/00268970903313667>.

(19) Fraczek, R.; Lobell, M.; Göller, A. H.; Krenz, U.; Schoenle, R.; Clark, R. D.; Hillisch, A. Best of Both Worlds: Combining Pharma Data and State of the Art Modeling Technology To Improve in Silico  $pK_a$  Prediction. *J. Chem. Inf. Model.* **2015**, *55* (2), 389–397. <https://doi.org/10.1021/ci500585w>.

(20) Li, M.; Zhang, H.; Chen, B.; Wu, Y.; Guan, L. Prediction of  $pK_a$  Values for Neutral and Basic Drugs Based on Hybrid Artificial Intelligence Methods. *Sci. Rep.* **2018**, *8* (1), 3991. <https://doi.org/10.1038/s41598-018-22332-7>.

(21) Roszak, R.; Beker, W.; Molga, K.; Grzybowski, B. A. Rapid and Accurate Prediction of  $pK_a$  Values of C–H Acids Using Graph Convolutional Neural Networks. *J. Am. Chem. Soc.* **2019**, *141* (43), 17142–17149. <https://doi.org/10.1021/jacs.9b05895>.



(22) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *ArXiv150909292 Cs Stat* **2015**.

(23) Baltruschat, M.; Czodrowski, P. Machine Learning Meets  $pK_a$ . F1000Research April 27, 2020. <https://doi.org/10.12688/f1000research.22090.2>.

(24) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Alga, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566* (7743), 224–229. <https://doi.org/10.1038/s41586-019-0917-9>.

(25) Warr, W. A.; Nicklaus, M. C.; Nicolaou, C. A.; Rarey, M. Exploration of Ultralarge Compound Collections for Drug Discovery. *J. Chem. Inf. Model.* **2022**, *62* (9), 2021–2034. <https://doi.org/10.1021/acs.jcim.2c00224>.

(26) Dietterich, T. G. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*; Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, 2000; pp 1–15. [https://doi.org/10.1007/3-540-45014-9\\_1](https://doi.org/10.1007/3-540-45014-9_1).

(27) Gunner, M. R.; Murakami, T.; Rustenburg, A. S.; Işık, M.; Chodera, J. D. Standard State Free Energies, Not  $pK_a$ s, Are Ideal for Describing Small Molecule Protonation and Tautomeric States. *J. Comput. Aided Mol. Des.* **2020**, *34* (5), 561–573. <https://doi.org/10.1007/s10822-020-00280-7>.

(28) See the Supporting Information for a Brief Description of the Macro- $pK_a$  Protocol.

- (29) Slater, A. M. The IUPAC Aqueous and Non-Aqueous Experimental  $pK_a$  Data Repositories of Organic Acids and Bases. *J. Comput. Aided Mol. Des.* **2014**, 28 (10), 1031–1034. <https://doi.org/10.1007/s10822-014-9764-9>.
- (30) Arús-Pous, J.; Blaschke, T.; Ulander, S.; Reymond, J.-L.; Chen, H.; Engkvist, O. Exploring the GDB-13 Chemical Space Using Deep Generative Models. *J. Cheminformatics* **2019**, 11 (1), 20. <https://doi.org/10.1186/s13321-019-0341-z>.
- (31) Yang, Y.; Yao, K.; Repasky, M. P.; Leswing, K.; Abel, R.; Shoichet, B. K.; Jerome, S. V. Efficient Exploration of Chemical Space with Docking and Deep Learning. *J. Chem. Theory Comput.* **2021**, 17 (11), 7106–7119. <https://doi.org/10.1021/acs.jctc.1c00810>.
- (32) Schrodinger Release 2022-4: Active Learning Glide, 2022.
- (33) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*, 1st edition.; O'Reilly Media: Sebastopol, CA, 2019.
- (34) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures- A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, 5 (2), 107–113. <https://doi.org/10.1021/c160017a018>.
- (35) Kirkwood, J. G.; Westheimer, F. H. The Electrostatic Influence of Substituents on the Dissociation Constants of Organic Acids. I. *J. Chem. Phys.* **1938**, 6 (9), 506–512. <https://doi.org/10.1063/1.1750302>.
- (36) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *ArXiv170401212 Cs* **2017**.

(37) Wenlock, M.; Tomkinson, N. *Experimental in vitro DMPK and physicochemical data on a set of publicly disclosed compounds*. Document Report Card. [https://www.ebi.ac.uk/chembl/document\\_report\\_card/CHEMBL3301361/](https://www.ebi.ac.uk/chembl/document_report_card/CHEMBL3301361/) (accessed 2021-12-30).

(38) Manchester, J.; Walkup, G.; Rivin, O.; You, Z. Evaluation of  $pK_a$  Estimation Methods on 211 Druglike Compounds. *J. Chem. Inf. Model.* **2010**, *50* (4), 565–571. <https://doi.org/10.1021/ci100019p>.

(39) Morgenthaler, M.; Schweizer, E.; Hoffmann-Röder, A.; Benini, F.; Martin, R. E.; Jaeschke, G.; Wagner, B.; Fischer, H.; Bendels, S.; Zimmerli, D.; Schneider, J.; Diederich, F.; Kansy, M.; Müller, K. Predicting and Tuning Physicochemical Properties in Lead Optimization: Amine Basicities. *ChemMedChem* **2007**, *2* (8), 1100–1115. <https://doi.org/10.1002/cmdc.200700059>.

(40) Liao, C.; Nicklaus, M. C. Comparison of Nine Programs Predicting  $pK_a$  Values of Pharmaceutical Substances. *J. Chem. Inf. Model.* **2009**, *49* (12), 2801–2812. <https://doi.org/10.1021/ci900289x>.

(41) 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.  $pK_a$  Measurements for the SAMPL6 Prediction Challenge for a Set of Kinase Inhibitor-like Fragments. *J. Comput. Aided Mol. Des.* **2018**, *32* (10), 1117–1138. <https://doi.org/10.1007/s10822-018-0168-0>.

(42) Settimo, L.; Bellman, K.; Knegtel, R. M. A. Comparison of the Accuracy of Experimental and Predicted  $pK_a$  Values of Basic and Acidic Compounds. *Pharm. Res.* **2014**, *31* (4), 1082–1095. <https://doi.org/10.1007/s11095-013-1232-z>.

(43) Pracht, P.; Wilcken, R.; Udvarhelyi, A.; Rodde, S.; Grimme, S. High Accuracy Quantum-Chemistry-Based Calculation and Blind Prediction of Macroscopic  $pK_a$  Values in the Context of the SAMPL6 Challenge. *J. Comput. Aided Mol. Des.* **2018**, *32* (10), 1139–1149. <https://doi.org/10.1007/s10822-018-0145-7>.

(44) Işık, M.; Rustenburg, A. S.; Rizzi, A.; Gunner, M. R.; Mobley, D. L.; Chodera, J. D. Overview of the SAMPL6  $pK_a$  Challenge: Evaluating Small Molecule Microscopic and Macroscopic  $pK_a$  Predictions. *J. Comput. Aided Mol. Des.* **2021**, *35* (2), 131–166. <https://doi.org/10.1007/s10822-020-00362-6>.

(45) Shelley, J. C.; Calkins, D.; Sullivan, A. P. Comments on the Article “Evaluation of  $pK_a$  Estimation Methods on 211 Druglike Compounds.” *J. Chem. Inf. Model.* **2011**, *51* (1), 102–104. <https://doi.org/10.1021/ci100332m>.

(46) Tang, H.; Jensen, K.; Houang, E.; McRobb, F. M.; Bhat, S.; Svensson, M.; Bochevarov, A.; Day, T.; Dahlgren, M. K.; Bell, J. A.; Frye, L.; Skene, R. J.; Lewis, J. H.; Osborne, J. D.; Tierney, J. P.; Gordon, J. A.; Palomero, M. A.; Gallati, C.; Chapman, R. S. L.; Jones, D. R.; Hirst, K. L.; Sephton, M.; Chauhan, A.; Sharpe, A.; Tardia, P.; Dechaux, E. A.; Taylor, A.; Waddell, R. D.; Valentine, A.; Janssens, H. B.; Aziz, O.; Bloomfield, D. E.; Ladha, S.; Fraser, I. J.; Ellard, J. M. Discovery of a Novel Class of D-Amino Acid Oxidase Inhibitors Using the Schrödinger Computational Platform. *J. Med. Chem.* **2022**. <https://doi.org/10.1021/acs.jmedchem.2c00118>.

(47) *REAL Database - Enamine*. <https://enamine.net/compound-collections/real-compounds/real-database> (accessed 2021-12-17).

(48) Schrodinger Release 2022-4: LigPrep, 2022.

(49) Schrodinger Release 2022-4: Protein Prep Workflow, 2022.

(50) Schrodinger Release 2022-4: LiveDesign, 2022.

## TABLE OF CONTENTS GRAPHIC

