Uni-pK_a: An Accurate and Physically Consistent pK_a Prediction through Protonation Ensemble Modeling

Weiliang Luo*
Peking University
DP Technology
luoweiliang7@pku.edu.cn

Gengmo Zhou*
Renmin University of China
DP Technology
zgm2015@ruc.edu.cn

Zhengdan Zhu
DP Technology
zhuzd@dp.tech

Guolin Ke
DP Technology
kegl@dp.tech

Zhewei Wei
Renmin University of China
zhewei@ruc.edu.cn

Zhifeng Gao†
DP Technology
gaozf@dp.tech

Hang Zheng†
DP Technology
zhengh@dp.tech

Abstract

Predicting pK_a values of small molecules has key applications in drug discovery and molecular simulation. However, current methods face challenges in rigorously interpreting experimental data and ensuring thermodynamic consistency between successive pK_a values. This study puts forward a protonation ensemble framework to address these limitations by modeling the full space of possible protonation microstates. Within this framework, we derive rigorous definitions connecting experimental macro-pK_as to underlying micro-pK_a equilibria. Under this new framework, we develop Uni-pK_a, an accurate and reliable pK_a predictor. Uni-pK_a first pretrains on over 1 million predicted pK_as from ChemBL to learn expressive molecular representations. It is then finetuned on experimental datasets that enforce consistency with the protonation ensemble definitions. The high-quality experimental pK_a datasets are fitted to this framework by recovering underlying microstates from macro-pK_as. Modeling the complete ensemble enables rigorous interpretation of macro-pK_a data, and inherently preserves thermodynamic consistency, improving the prediction accuracy of Uni-pK_a. Experiments demonstrate that Uni-pK_a achieves state-of-the-art performance, outperforming previous methods. This novel protonation ensemble approach significantly advances machine learning for pK_a prediction and molecular property modeling. Uni-pK_a provides a good example of how to combine chemical knowledge and machine learning methods. Users can utilize Uni-pKa for predicting and ranking the protonation states of molecules under various pH conditions via https://app.bohrium.dp.tech/uni-pka.

Keywords: pK_a, Protonation, Molecular Pretraining

1 Introduction

In Brønsted-Lowry acid-base theory [1, 2], an acid is a molecule with ionizable hydrogen, which can transfer to another molecule. In particular, an acid HA dissociates in a solution with the following chemical equilibrium,

$$\mathrm{HA} \rightleftharpoons \mathrm{H}^+(\text{solvated}) + \mathrm{A}^-, \quad K_a = \frac{[\mathrm{H}^+][\mathrm{A}^-]}{[\mathrm{HA}]}$$

*Equal contribution.

†Corresponding authors.

where $[\cdot]$ is a chemical species's activity (or dimensionless concentration, approximately). Then pK_a, the negative logarithm (base 10) of the acid dissociation constant K_a, is the key physical chemistry parameter describing the acid/base property.

Many small organic drug molecules contain acid/base groups like carboxyl groups, amino groups, and *N*-heterocyclic rings. Their pK_a values directly determine protonation states in physiological environments, influencing key properties like solubility, membrane permeability, and biomolecular interactions. As such, pK_a prediction plays an important role in screening drug-like molecules with optimal pharmacokinetics, toxicity, and activity [3]. In molecular simulations assessing activity evaluation, such as free energy perturbation methods, accurate pK_a values also enable proper structure preparation and thermodynamic correction, improving accuracy [4, 5]. Therefore, fast and reliable pK_a prediction approaches are highly valuable in drug discovery and related applications.

Quantum chemistry provides *ab initio* solutions for calculating thermodynamic properties like pK_a. Equipped with comprehensive conformation research and well-designed correction methods, Schrödinger's Jaguar pK_a prediction software has reached experimental accuracy in a large chemical space [6, 7]. Jaguar predicts pK_a based on DFT-calculated free energy. The results are very sensitive to the conformation because a conformational energy of merely 1 kcal/mol is corresponding to more than 0.7 pK_a unit. Therefore, the best accuracy is usually reached under the conformational ensemble average after a comprehensive conformational search, with a proportional increasing amount of computation. In practice, the trade-off between speed requirement and accuracy expectation determines the conformation search strategy.

With the growth of experimental data and cheminformatics, machine learning has enhanced molecular property prediction across tasks [8, 9]. However, pK_a modeling faces unique challenges compared to predicting properties like ADME/T characteristics. A salient difficulty is the prevalence of multiple ionizable groups within drug molecules. Both molecular and group-level information must be encoded, and general pK_a models should handle polyprotonated and amphoteric cases. Framing pK_a prediction as a simple multi-label regression problem with individual site labels overlooks this complexity.

With this consideration, recent chemoinformatics works use different descriptions of the molecular structure and ionization sites to achieve promising accuracy and outstanding speed:

- Template-based methods utilize ionization site matching to empirical fragment values, along with correction of surrounding structural context by Hammett linear free energy relationships [10], as implemented in early versions of Epik [11].
- Local atomic descriptors represent ionization sites while global molecular descriptors cover full structures in traditional machine learning techniques, including OPERA [12], the work of Baltruschat et al. [13], and SPOC [14].
- Graph neural networks learn hierarchical embeddings of sites and structures at different levels of molecular graphs, as demonstrated by MolGpKa [15], pKasolver [16], Graph-pKa [17], MF-SuP-pKa [18], and Epik 7 [19].

While recent methods have made progress on representing molecules and ionization sites for pK_a prediction [20], fundamental limitations remain in interpreting experimental data and ensuring thermodynamic consistency.

On the data side, most public pK_a data relies on macroscopic spectrophotometric or electrochemical measurements, reflecting an apparent equilibrium between various protonation states [21, 22]. However, these macro-pK_as are often ascribed to a single dominant site when incorporated into prediction datasets and training, inducing bias [23]. As discussed for decades, rigorous interpretation requires accounting for coupled contributions from all sites [24]. Recent attempts like the MIL framework proposed by Xiong et al. [17, 18] accommodate multiple ionization sites but remain ignorant of complex protonation networks and uniform treatment for amphoteric cases.

On the model side, thermodynamic coupling emerges when it comes to the modeling of poly-protonation [25]. Successive pK_a values along different protonation orders are constrained by chemical equilibrium relations, which is violated in predictions with independent site modeling. These models not only compromise the rigor but also risk thermodynamic inconsistency in the calculation of the pH-dependent distribution of protonation states [19]. Under the strong demand for protonation state ranking of given molecules, genuinely self-consistent pK_a prediction remains an unmet need.

These intertwined limitations of current approaches, both in interpreting experimental data and ensuring thermodynamic consistency in predictions, underscore the need for a new modeling perspective. For example, a recent study on the SAMPL6 challenge highlights the advantages of using standard free energies rather than pK_as when representing complex protonation systems [25]. Inspired by such works, we put forward a protonation ensemble framework, Uni-pK_a, that rethinks representations, data integration, and modeling under a unifying view that captures collections of microstates.

On the data side, we design a general format of the pK_a dataset, which stores the determined molecular structure of protonation states and is compatible with both micro- and macro-pK_a data. We reconstruct several publicly available datasets in this format and release them to provide a rigorous, molecule-level interpretation. This is a new benchmark for high-accuracy pK_a models.

On the model side, we introduce a modified Uni-Mol model into a free-energy-based machine learning framework with novel pretraining strategies. It allows the model to learn from both macro- and micro-pK_as, naturally preserves thermodynamic consistency, and enables multiple scenarios including pK_a prediction and protonation state determination. After pretraining on large-scale predicted pK_as and finetuning on experimental pK_as, Uni-pK_a achieves state-of-the-art accuracy for pK_a prediction compared to other chemoinformatics models.

Bridging the gap between data and model, we develop a structure enumerator to generate a protonation ensemble from given molecules. It helps to build the dataset and propose a workflow for structure preparation in molecular simulation, combining speed and accuracy.

In conclusion, we advance pK_a modeling by integrating chemical knowledge with data-driven techniques. The released datasets and Uni-pK_a framework synergistically improve the interpretation of experimental data and thermodynamic consistency.

2 Theory

Interpreting experimental macro-pK_a measurements requires modeling the underlying protonation microstates and equilibria, which are obscured in bulk techniques. We emphasize a protonation ensemble perspective that captures the collection of possible microstates corresponding to different protonation site combinations for a given molecule (Figure 1).

For a core structure A with n ionizable sites, there are theoretically 2^n microstate structures spanning different net charge macrostates from fully protonated to fully deprotonated. While unreasonable microstates can be excluded based on chemical knowledge, this still defines a broad ensemble.

A key distinction arises between micro-pK_as describing the equilibrium between determined microstates, and macro-pK_as reflecting the apparent equilibrium between total activities of all microstates with adjacent protonation levels [26].

Micro-pK_a values arise from the standard free energy change $\Delta_\mathrm{f}G_\mathrm{m}^\ominus$ between a defined microstate pair k-$\mathrm{H}_m\mathrm{A}^{m+}$ and i-$\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+}$. Let $\Delta_\mathrm{f}G_\mathrm{m}^\ominus(\cdot)$ be the standard molar Gibbs free energy change of formation at temperature T,

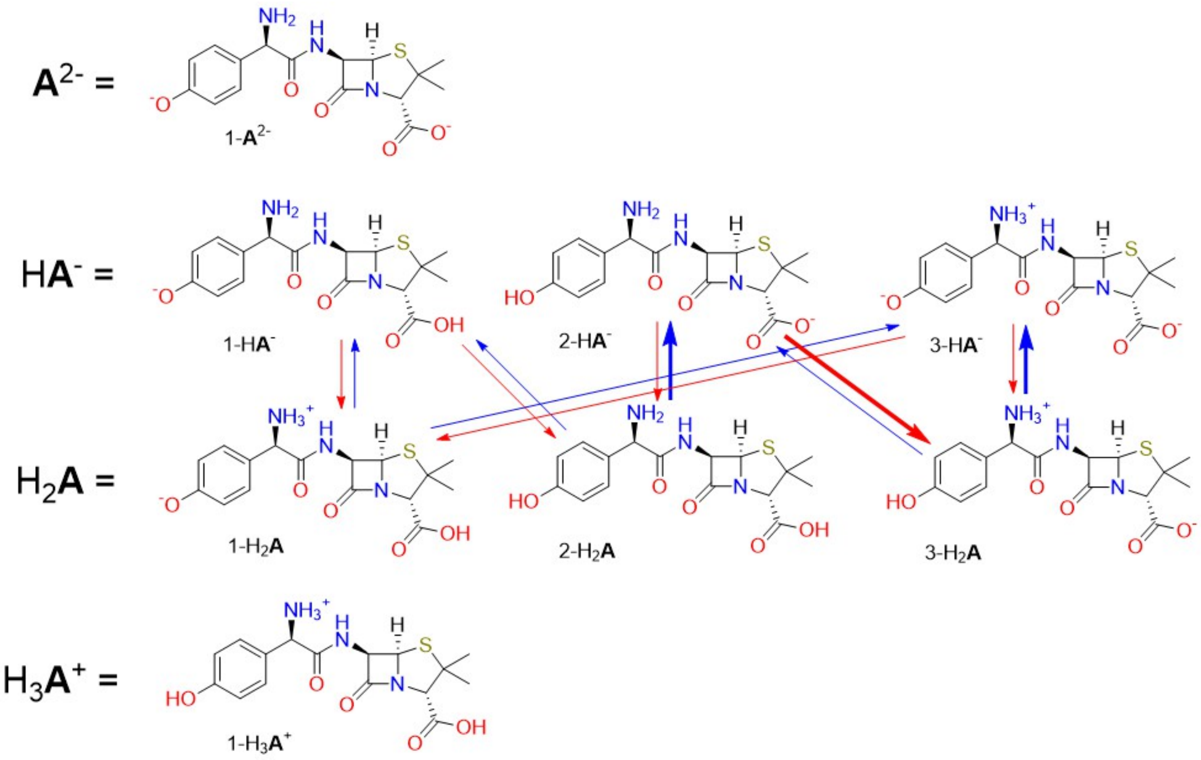


Figure 1: The protonation ensemble of Amoxicillin. Only the ionization of amino (basic), phenolic hydroxyl (acidic), and carboxyl (acidic) groups are taken into consideration. The red and blue arrows are all direct protonation and deprotonation reactions between macrostates HA^- and H_2A. The bolded arrows is one of the shortest path from 2-H_2A to 3-HA^-.

R be the molar gas constant, and $\beta = (RT)^{-1}$. The micro-pK_a is:

$$K^{k,i}_{\mathrm{a},m} := \frac{[k\text{-}\mathrm{H}_m\mathrm{A}^{m+}][\mathrm{H}^+]}{[i\text{-}\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+}]} = \exp\left\{-\beta\left[\Delta_\mathrm{f}G^\ominus_\mathrm{m}(k\text{-}\mathrm{H}_m\mathrm{A}^{m+}) - \Delta_\mathrm{f}G^\ominus_\mathrm{m}(i\text{-}\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+})\right]\right\} \tag{1}$$

Macro-pK_a values originate from the coarse-grained behaviors across all microstates, derived as:

$$K_{\mathrm{a},m} := \frac{[\mathrm{H}^+]\sum_i[i\text{-}\mathrm{H}_m\mathrm{A}^{m+}]}{\sum_i[i\text{-}\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+}]} = \frac{\sum_i \exp(-\beta\Delta_\mathrm{f}G^\ominus_\mathrm{m}(i\text{-}\mathrm{H}_m\mathrm{A}^{m+}))}{\sum_i \exp(-\beta\Delta_\mathrm{f}G^\ominus_\mathrm{m}(i\text{-}\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+}))} \tag{2}$$

This connection enables a proper interpretation of measured macro quantities in terms of specific microstate equilibria. The macro-pK_a-free-energy formula 2 degrades to the micro-pK_a-free-energy formula 1 when the microstate index i, k in both macrostates $H_mA^{m+}, H_{m+1}A^{(m+1)+}$ is unique. As a result, micro-pK_a is a special case of macro-pK_a, and both micro- and macro-pK_a are described by $\Delta_\mathrm{f}G^\ominus_\mathrm{m}$.

pH-dependence can also be introduced into the microstate free energies, as rigorously derived in Appendix A:

$$\Delta_\mathrm{f}G_\mathrm{m}(\mathrm{H}_m\mathrm{A}^{m+};\mathrm{pH}) := \Delta_\mathrm{f}G^\ominus_\mathrm{m}(\mathrm{H}_m\mathrm{A}^{m+}(\mathrm{aq})) + \frac{m\ln 10}{\beta}\mathrm{pH}, \tag{3}$$

This gives the fraction of each microstate across the ensemble under particular pH conditions:

$$w(k\text{-}\mathrm{H}_m\mathrm{A}^{m+};\mathrm{pH}) := \frac{[k\text{-}\mathrm{H}_m\mathrm{A}^{m+}]}{\sum_n\sum_i[i\text{-}\mathrm{H}_n\mathrm{A}^{n+}]} = \frac{\exp(-\beta\Delta_\mathrm{f}G_\mathrm{m}(k\text{-}\mathrm{H}_m\mathrm{A}^{m+};\mathrm{pH}))}{\sum_n\sum_i\exp(-\beta\Delta_\mathrm{f}G_\mathrm{m}(i\text{-}\mathrm{H}_n\mathrm{A}^{n+};\mathrm{pH}))}. \tag{4}$$

Unifying the micro-pK_a-free-energy formula 1, macro-pK_a-free-energy formula 2, and distribution-fraction-free-energy formula 4, we can see that free energies of all the microstates in the protonation ensemble contain the complete information of the acid/base equilibrium. By capturing coupled equilibria from an energy perspective, integrated equilibrium information can be extracted faithfully from both micro- and macro-pK_a data. Therefore, the protonation ensemble framework provides a foundation for accurate and rigorous pK_a prediction capabilities. It allows predicting both pK_a values and pH-dependent protonation states through protonation ensemble construction (Section 3) and microstate modeling of free energies (Section 4).

3 Toolkits and Datasets

3.1 Microstate enumerator

We implement a microstate enumerator for the systematic reconstruction of the protonation ensemble from a single structure. It processes the structure of a part of macrostate H_mA^{m+} to generate all microstates in H_mA^{m+} and a neighboring macrostate $H_{m+1}A^{(m+1)+}$ or $H_{m-1}A^{(m-1)+}$. The code is open source at https://github.com/dptech-corp/Uni-pKa.

The enumerator uses a template containing SMARTS patterns of ionizable sites (Table 1). It is modified, augmented, and annotated based on the template in MolGpKa [15] with chemical consideration. It contains 53 common acidic and basic groups with separate entries for deprotonation and protonation (Table 1) and covers all the ionization patterns demonstrated by the Dwar-iBond dataset introduced in the section 3.2.

Table 1: An example of double entries. Every group in the template has double-type entries that match the A to B ionization (deprotonation or basic ionization) and B to A ionization (protonation or acidic ionization). The A2B-type entry matches carboxyl acid, cyanic acid, and imidic acid and labels the oxygen as the atom to be deprotonated, and the B2A-type entry matches the deprotonated structure of those groups and labels the oxygen as the atom to be protonated.

type	SMARTS	atom index
A2B	[$([#6]=[#8,#7]),$(C#N):0]-[OX2:1]-[H:2]	1
B2A	[$([#6]=[#8,#7]),$(C#N):0]-[O-1:1]	1

When the enumeration starts, A and B Micro Pools are first built. They are dynamic sets containing microstates of higher and lower charged macrostates (**A**cids and **B**ases) respectively in two adjacent protonation levels(Figure 2). The algorithm then iteratively grows the pools:

- **A to B (A2B) round: deprotonation.** For each structure in A Micro Pool, substructure matching finds all possible deprotonation sites in the template and corresponding deprotonated structures go into B Micro Pool, like the blue line in Figure 2.
- **B to A (B2A) round: protonation.** For each structure in B Micro Pool, substructure matching finds all possible protonation sites in the template and corresponding protonated structures go into A Micro Pool, like the red line in Figure 2.

Therefore, beginning with some of H_mA^{m+}, if the macrostate $H_{m-1}A^{(m-1)+}$ is needed, the initial structures will be thrown into A Micro Pool with B Micro Pool empty, and an A2B round will go first (**A**cid mode). $H_{m+1}A^{(m+1)+}$ is also available when starting from a B2A round (**B**ase mode).

The two rounds alternate until the two pools are not growing anymore or the maximum number of iterations has been reached, then A and B Micro Pools are output as the final enumeration results. The max-iteration limit is customized to reduce memory consumption and increase efficiency when the huge enumeration results of very complex molecules are poured into the machine-learning model. In addition, another template filters out chemically unreasonable structures during enumeration (Structure Filter in Figure 2), like the coexistence of acidic ionization of amino group and basic ionization of amino group. These structures can be pruned because of their small contribution to the protonation ensemble.

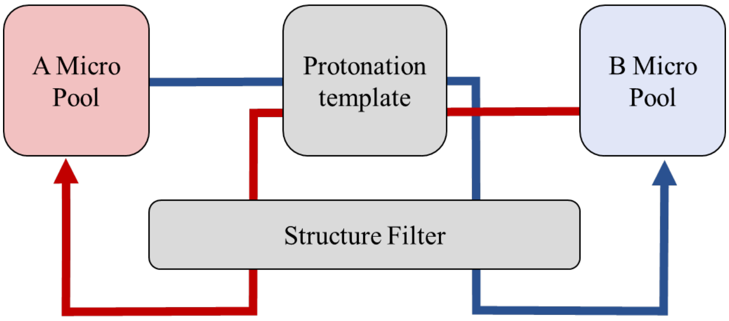


Figure 2: A structure chart of the microstate enumerator

The whole protonation ensemble is obtained by successively running the enumeration process above in the A and B modes. In the case of Figure 1, the whole macrostate of H_2A and HA^- can be enumerated from 1-H_2A in the A mode, H_3A^+ comes from H_2A in the B mode, and A^{2-} steps further from HA^- in the A mode.

The width (the number of microstates in macrostates) and depth (the number of macrostates) of the protonation ensemble enumeration are both determined by the coverage of the template. For example, if the template only contains the basic ionization of amino groups, acidic ionization of phenolic hydroxyl groups, and acidic ionization of carboxyl groups, the enumeration between H_2A and HA^- in Figure 1 will stop at the structures illustrated in the figure. However, if the acidic ionization of amide is recorded in the template, more structures with the proton on amide groups transferring to other sites will occur in H_2A and HA^-, increasing the width of enumeration results. Furthermore, when the amide group in 1-A^{2-} is deprotonated, the macrostates extend to $H_{-1}A^{3-}$, increasing the depth of enumeration results.

A more comprehensive template leading to wider and deeper enumeration certainly provides a more complete description of the protonation ensemble, but the number of enumerated structures grows exponentially with the number of matched ionization sites of the molecule. This complexity is suffered both in protonation ensemble reconstruction and the machine learning model. It is essential to prune negligible microstates with minimal influence on the accuracy of the distribution of the protonation ensemble. Template refinement and structural screening are respectively procedure-oriented and result-oriented solutions. However, reasonable pruning rules in the vast chemical space are case by case, and we finally choose the Dwar-iBond dataset as a representative of the ionization pattern to determine the standard coverage of the template. Manual screening also partially complements the structure filter to build a lightweight but effective training set for the machine learning model.

3.2 Reconstructed datasets

In existing public pK_a datasets, like DataWarrior pKaInWater [27] and i-Bond [28], each entry contains only one microstate structure with a designated ionization site and mode, empirically assumed to correspond to the reported macro-pK_a value. This risks a biased interpretation of experimental measurements reflecting coupled equilibria.

To enable rigorous modeling, we reconstructed several datasets by leveraging our microstate enumerator to recover the complete protonation ensembles underlying reported macro-pK_a values.

While the single provided structure is incomplete, properties like the core scaffold, initial charge, and reaction type contain sufficient information for the enumerator to regenerate the full macrostates involved in the macro-pK_a equilibrium through iterative templated protonation and deprotonation.

This process reformats the datasets into a unified table structure that stores the enumerated microstates mapped to each published macro-pK_a measurement (Table 2).

Table 2: An example of the dataset table

SMILES	pK_a	Reference
COc1cccc2c1-c1c3c(OC)cccc3[nH+]c3cccc(c13)N2C,COc1cccc2c1-c1c3c(OC)cccc3nc3cccc(c13)[NH+]2C»COc1cccc2c1-c1c3c(OC)cccc3nc3cccc(c13)N2C	8.9	[29]
CCN(CC)c1cccc2cccc([NH+](CC)CC)c12»CCN(CC)c1cccc2cccc(N(CC)CC)c12	2.7	[30]

Our release covers 7 experimental and predicted datasets relevant to drug-like chemical space (Table 3), including:

- Small molecule compilations like SAMPL - with exhaustive microstate enumeration
- Large predicted set from ChemBL - initial scaffold enumeration

Table 3: Basic information of the released datasets

Name	Type	Number of Datapoints	Number of Microstates
ChemBL[31, 18]	Predicted values	1116294	3139065
Dwar-iBond[27, 28]	Experimental values	8232	27138
Novartis Acid[32, 18]	Experimental values	112	345
Novartis Base[32, 18]	Experimental values	168	696
SAMPL6[33]	Experimental values	31	111
SAMPL7[34]	Experimental values	20	43
SAMPL8[35]	Experimental values	25	117

This work integrates robust chemical knowledge about protonation mechanisms with consistent experimental measurements into high-quality datasets tailored for developing accurate and physically consistent machine learning models. Full details of the source data and reconstruction process are provided in Appendix B. Relevant datasets can be obtained from https://www.aissquare.com/datasets/detail?pageType=datasets&name=Uni-pKa-Dataset.

4 Uni-pK_a Framework

4.1 Overview

The Uni-pK_a framework integrates the protonation ensemble, the microstate enumerator, and machine learning. The protonation ensemble framework establishes the theoretical foundation of microstate free-energy modeling for multiple pK_a-related tasks and formulates the data flow of the framework. The microstate enumerator implements the protonation ensemble generation and preprocesses the molecular data for the data flow of the framework.

The machine learning part serves as the core algorithm of microstate free-energy modeling. Receiving molecular structures from the microstate enumerator and organized by the protonation ensemble, it converts molecular inputs to free-energy outputs in the data flow. Following a pretraining-finetuning paradigm, it learns from data with different fidelity and grows into a highly accurate free-energy predictor in the model flow.

Figure 3 provides a schematic overview of the Uni-pK_a framework. Uni-pK_a employs a unified data preparation workflow across the stages of pretraining, finetuning, and inference. Instead of directly inputting a single ionization reaction into the model, we perform protonation enumeration on the data points to obtain microstates for the acid and base sides. Each microstate represents a molecule, and these molecules are then fed into the model.

The model backbone originates from Uni-Mol [36], an expressive and universal 3D molecular representation learning framework based on Transformer [37], which has demonstrated effectiveness across a range of molecular property prediction tasks. In Uni-pK_a, We make necessary modifications, including the incorporation of charge information and its FE2pK_a module under the protonation ensemble.

The pretraining phase involves four tasks: one weakly supervised task, pK_a prediction, and three self-supervised tasks, including 3D position recovery, masked atom prediction, and masked charge prediction. In the pK_a prediction task, unlike previous models that directly predict pK_a values, Uni-pK_a ensures the consistency of molecular protonation ensembles by taking individual microstate molecules as input, and the model's output is interpreted as predicted free energy. This is enabled by the theoretical analysis in Section 2 showing free energy models can represent protonation equilibria. With the free energy predicted for each molecule in the data point, we employ the free energy-pK_a formulas to predict the pK_a value for the entire data point and compute the loss with the ground truth. Details are in the following subsections.

After pretraining, we conduct finetuning with experimental pK_a labels, enabling our model to possess the capability of predicting high-precision pK_a, as depicted in the middle-right section of Figure 3.

Then, the resultant well-trained model is adept at serving three distinct tasks during the inference stage including macro-pK_a prediction, micro-pK_a prediction, and distribution fraction prediction.

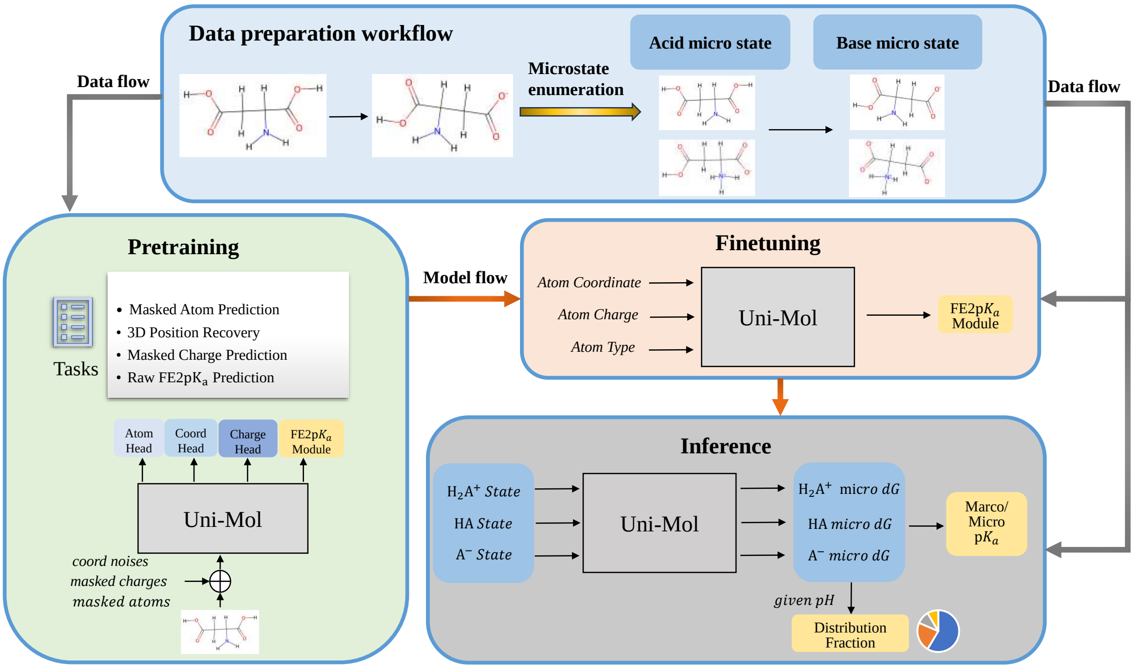


Figure 3: Schematic overview of Uni-pK_a framework. Top: Data preparation workflow. We implement a microstate enumerator to systematically build the protonation ensemble from a single structure. **Left:** pretraining workflow. Our pretraining strategy combines 1 weakly supervised task, pK_a-prediction, and 3 self-supervised pretraining tasks, masked atom prediction, masked charge prediction, and 3D position recovery, to make the most use of the chemical information in 3 million microstate structures. In the pK_a-prediction task, we introduce a free energy-to-pK_a(FE2pK_a) module to establish the relationship between the model-predicted free energy and pK_a. This module also enables us to predict pK_a from free energies. **Center right:** finetuning workflow. In this phase, we also employ the FE2pK_a module, training the model using experimental pK_a to enhance its capability for predicting pK_a with high accuracy. **Bottom right:** inference workflow. After pretraining and finetuning, the well-trained Uni-pK_a framework is equipped to handle three inference tasks, including macro pK_a prediction, micro pK_a prediction, and distribution fraction prediction.

4.2 Model input

Different from vanilla one, Uni-Mol in the Uni-pK_a framework has three inputs. Along with atom types and atom coordinates, we also consider the influence of atom charges, as it is closely related to molecular protonation. Atom representations are initialized through an embedding layer based on atom types. Atom charges are categorized into discrete and continuous charges. For discrete charges, we consider states with formal charge values of 0, 1, and -1, representing neutral, positively charged, and negatively charged atoms. Similar to atom representations, discrete charge representations are initialized through an embedding layer based on charge types. For continuous charges, we employ a Multi-Layer Perceptron (MLP) to obtain their initial representations. The acquisition of pair representations follows the same procedure as in vanilla Uni-Mol. It involves obtaining an atom distance matrix using atom coordinates and initializing pair representations through invariant positional encoding. Further details can be found in the Uni-Mol publication [36].

4.3 Pretraining

Inspired by the success of large language models in natural language processing [38, 39, 40, 41] and computer vision [42], the machine learning model in the Uni-pK_a framework follows the pretraining-finetuning paradigm. The objective of pretraining is to learn the underlying structures and features from the massive amount of data, enabling the model to capture high-level representations. And finetuning allows the model to optimize its performance on the specific prediction task through supervised learning.

In the scenario of pK_a prediction, previous work has proved the reasonability and effectivity of this paradigm, using predicted pK_a values in the ChemBL dataset as ”low fidelity data” for a weakly supervised pretraining of pK_a models [15, 16, 18]. Our strategy further combines one weakly supervised task and three self-supervised pretraining tasks to make the most use of the chemical information in these 3 million microstate structures.

4.3.1 Weakly supervised task: pK_a-prediction

Firstly, supervised pretraining is performed to predict the labels provided with the ChembL data, helping the model to learn mapping relationships from the large-scale labeled data. As mentioned previously, to ensure the consistency of molecular protonation ensembles, Uni-Mol in Uni-pK_a takes individual microstate molecules as input, and the output is interpreted as predicted free energy. Specifically, similar to the language model BERT [38], we introduce a special atom called [CLS]. Its coordinates represent the center of the molecule. We use this atom to represent the entire molecule.

Then, we introduce a FreeEnergy2pK_a (FE2pK_a) module. With a linear head, Uni-Mol utilizes the representation of [CLS] to obtain the raw vector output. This output will be interpreted as the predicted $\beta\Delta_\mathrm{f}G_\mathrm{m}^\ominus$s for given microstates, guaranteed by its relationship to the pK_a labels. In a data entry, if the free energy output of Uni-Mol is $g_1^\mathrm{A}, g_2^\mathrm{A}, \cdots$ for the microstates in A macrostate, and $g_1^\mathrm{B}, g_2^\mathrm{B}, \cdots$ for the microstates in B macrostate, then the final loss function of a single datapoint is a combination of Mean Square Error loss and the macro-pK_a-free-energy Equation 2:

$$\mathcal{L}_{\mathrm{p}K_\mathrm{a}}(\boldsymbol{g}^\mathrm{A}, \boldsymbol{g}^\mathrm{B}; \mathrm{p}K_\mathrm{a}) = \frac{1}{2}\left[\mathrm{p}K_\mathrm{a} + \log_{10}\frac{\sum_i \mathrm{e}^{-g_i^\mathrm{B}}}{\sum_i \mathrm{e}^{-g_i^\mathrm{A}}}\right]^2 \tag{5}$$

This loss function links the predicted free energy of Uni-Mol with the experimental macro-pK_a label to enforce consistency with the protonation ensemble view:

- For each data point, Uni-Mol in Uni-pK_a outputs free energy vectors $\boldsymbol{g}^\mathrm{A}$ and $\boldsymbol{g}^\mathrm{B}$ for the microstates in macrostates A and B.

- These are used to compute the total Boltzmann-weighted partition functions of A and B microstates based on Equation 2.
- The loss function (Equation 5) compares this logarithmic partition-function ratio to the reported macro-pK_a through a Mean Squared Error term.
- By back-propagating this ensemble-aware loss, Uni-Mol in Uni-pK_a learns consistent free energy predictions.

We also note that standard label preprocessing like scaling would break the physical meaning of the outputs. However, translation maintains interpretation as pH-dependent free energies, as proved in Appendix A.

4.3.2 Self-supervised tasks

Additionally, we introduce three self-supervised learning tasks in the pretraining phase. Apart from the existing masked atom prediction and 3D position recovery tasks in Uni-Mol, we add a new masked charge prediction task, as atom charges are closely related to pK_a prediction.

Specifically, similar to the approach used in masked language models, we randomly select 15% of the atoms in the molecule to mask and use $[\mathrm{MASK}]$ token prediction by replacing masked atom types with a $[\mathrm{MASK}]$ token and predicting their original ones during pretraining with a linear head. We utilize the cross-entropy loss function for this task and this loss constitutes a part of the original Uni-Mol loss. Here, we denote this loss as $\mathcal{L}_\mathrm{atom}$.

Then, in Uni-pK_a, we introduce a unique task known as masked charge prediction. Molecular electronegativity is closely related to acid-base properties, and in pK_a prediction, the transfer of protons in a molecule is often associated with the distribution of atom charges. By predicting atom charges, the model can learn about the electrostatic interactions between different atoms, thereby enhancing its understanding of proton transfer. Similar to the masked atom prediction, we also perform masking for discrete charges of these masked atoms. The masked charges are replaced with a $[\mathrm{MASK}]$ token and predict their original ones during pretraining. We also use the cross-entropy loss function in this task. The loss for this task is referred to as $\mathcal{L}_\mathrm{charge}$. We consider that masked charge prediction contributes to a deeper understanding of a molecule's chemical properties, leading to more accurate predictions of acid-base properties and pK_a values.

Furthermore, we aim for the model to learn the 3D structural information within molecules. Therefore, we retain the 3D positions recovery task from Uni-Mol. Since molecular coordinates are continuous values, we introduce noise to the masked atoms' coordinates instead of masking and train the model to recover the ground truth coordinates from corrupted ones. This allows the model to capture structural information during pretraining. We employ two additional heads to recover the true coordinates from corrupted ones. The first one is the pair-distance prediction head, where the model is tasked with recovering the original Euclidean distance matrix based on the pairwise distances computed from the corrupted coordinates. The second head is the SE(3)-equivariant coordinate prediction head, where the model aims to recover the true coordinates while preserving the equivariance to rotation and translation of the molecule. We use the smooth ℓ_1 loss for both of these tasks. They are denoted as $\mathcal{L}_\mathrm{coord}$ and $\mathcal{L}_\mathrm{dist}$, and they also constitute a part of the original Uni-Mol loss.

4.3.3 Training objective

Due to the combination of supervised and self-supervised pre-training, the training complexity increases, and we adjust the proportion of self-supervised task loss accordingly. The final composition of the loss function and corresponding formulas are as follows:

$$\mathcal{L}_\mathrm{all} = \mathcal{L}_{\mathrm{p}K_\mathrm{a}} + \mathcal{L}_\mathrm{charge} + \mathcal{L}_\mathrm{atom} + 2\mathcal{L}_\mathrm{coord} + \mathcal{L}_\mathrm{dist} \tag{6}$$

4.4 Finetuning

To ensure consistency with the pretraining phase, we maintain the same data preparation workflow during the finetuning process. During finetuning, we also follow the setup of the pK_a prediction task in the pretraining phase. The pretrained Uni-Mol model in Uni-pK_a is then finetuned on the Dwar-iBond dataset using the loss function 5.

For aiding model convergence, the pK_a target is translated by the average of the dataset in both the pretraining and finetuning stages. In addition, regarding molecules, leveraging the ability to swiftly generate multiple random conformations allows us to incorporate data augmentation techniques during finetuning. This approach enhances both performance and robustness.

In summary, pretraining and finetuning synergistically integrate the benefits of representation learning at scale from abundant inaccurate pK_as, with focused supervised tuning on limited accurate measurements.

4.5 Inference

The pretraining and finetuning together develop an accurate and robust machine learning model in Uni-pK_a, capable of effectively learning from macro-pK_a data while preserving thermodynamic consistency. Taking advantage of the physical interpretation of microstate free-energy modeling, Uni-pK_a supports multiple prediction tasks in a unified workflow (Figure 4):

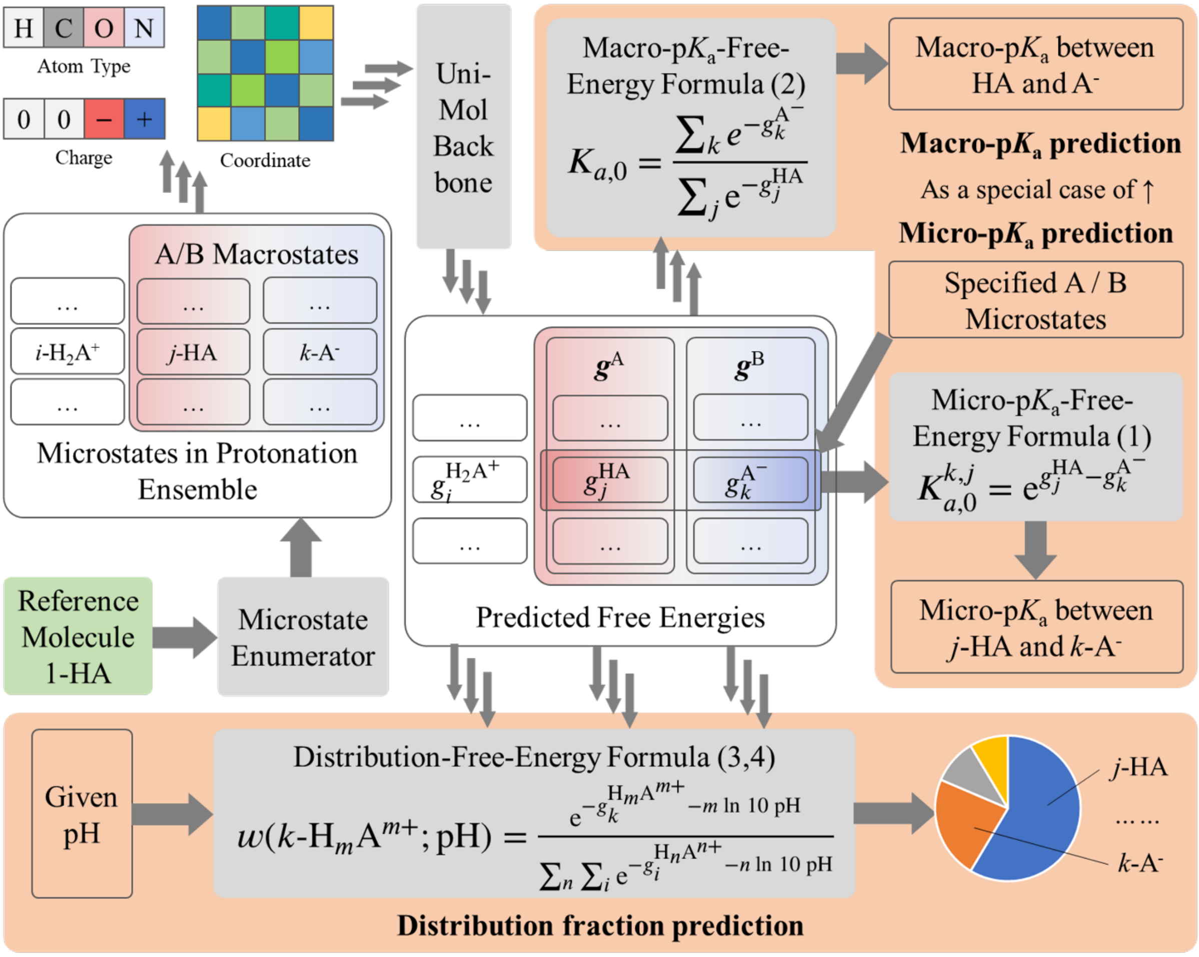

H C O N
Atom Type
0 0 − +
Charge
Coordinate
Uni-Mol Backbone
Macro-pK_a-Free-Energy Formula (2)
$K_{a,0} = \frac{\sum_k e^{-g_k^{A^-}}}{\sum_j e^{-g_j^{HA}}}$
Macro-pK_a between HA and A$^-$
Macro-pK_a prediction
As a special case of ↑
Micro-pK_a prediction
Specified A / B Microstates
A/B Macrostates
…
i-H_2A^+
j-HA
k-A$^-$
Microstates in Protonation Ensemble
g^A
g^B
$g_i^{H_2A^+}$
g_j^{HA}
$g_k^{A^-}$
Predicted Free Energies
Micro-pK_a-Free-Energy Formula (1)
$K_{a,0}^{k,j} = e^{g_j^{HA} - g_k^{A^-}}$
Micro-pK_a between j-HA and k-A$^-$
Reference Molecule 1-HA
Microstate Enumerator
Given pH
Distribution-Free-Energy Formula (3,4)
$w(k\text{-}H_mA^{m+}; \mathrm{pH}) = \frac{e^{-g_k^{H_mA^{m+}} - m \ln 10\ \mathrm{pH}}}{\sum_n \sum_i e^{-g_i^{H_nA^{n+}} - n \ln 10\ \mathrm{pH}}}$
j-HA
……
k-A$^-$
Distribution fraction prediction

Figure 4: The inference stage of Uni-pK_a

Macro-pK_a prediction Macro-pK_a prediction works analogously to the dataset reconstruction and finetuning process. The input is a molecule and its ionization mode (acidic or basic). The microstate enumerator first recovers the complete macrostates involved in the equilibrium as the model input. Uni-Mol in Uni-pK_a then predicts the pH-dependent free energies of the constituent microstates. Finally, the overall macro-pK_a is determined from the total Boltzmann-weighted partition functions of the macrostate microstates using the macro-pK_a free energy formula in Equation 2.

Micro-pK_a prediction Micro-pK_a prediction directly takes as input the specified reactant and product microstates. Their predicted free energy difference from Uni-Mol in Uni-pK_a is used to obtain the micro-pK_a value through the micro-pK_a free energy formula in Equation 1. This can be considered a special case of the macro-pK_a prediction workflow with a single, specified microstate in both macrostates.

Distribution fraction prediction The protonation state distribution fraction prediction starts from an initial molecule structure. The microstate enumerator works successfully across higher and lower net charge states, exploring the complete protonation ensemble until hitting the template limits. Uni-Mol in Uni-pK_a then predicts the free energies of all the generated microstates. Finally, by applying the distribution fraction formula in Equation 4 using the specified pH, the fraction of each microstate can be calculated. The dominant protonation state is typically chosen as the maximum fraction microstate for downstream usage.

By unifying the protonation ensemble, the microstate enumerator, and Uni-Mol, the Uni-pK_a framework enables reliable and consistent pK_a prediction workflow across diverse tasks.

5 Results and Discussion

5.1 Interpreting Macro-pKa Data

Accurately modeling macro-pK_a measurements requires accounting for the complete protonation ensemble, which refers to the collection of microstates with different protonation site combinations for a molecule. We analyzed our reconstructed datasets to quantify the additional information from the full enumeration.

As shown in Table 3, mapping the public pK_a datasets from individual structures to the underlying ensembles expands the data substantially. For instance, the Dwar-iBond dataset grows over 3-fold from 8,232 single data points to 27,138 enumerated microstates. This affirms the intrinsic complexity obscured by typical data representations.

We can visualize how ensemble modeling avoids biased assumptions about dominant sites. On the left of Figure 5, the acidity of the carboxyl group is known to be much stronger than the phenolic hydroxyl group, leading to the obvious assignment. While for molecules with chemical groups of similar acidity as shown on the right of Figure 5, ambiguities often exist in attributing macro-pK_as to specific sites, and any assignment is an oversimplification and introduces bias to data. Our protonation ensemble modeling reveals alternative chemically reasonable sites, including the dimethylamino group and the phenol group.

Recent works have adopted multi-instance learning (MIL) to decompose macro-pK_as into contributions from specific sites [17, 18]. However, MIL is a downgrade framework of the proton ensemble and risks misattributions for complex cases like Figure 1, shown in Appendix A. Our iterative, ensemble-aware enumerator explores the full space, avoiding assumptions. Thoroughly sampling the ensemble is imperative for rigorous macro-pK_a interpretation.

The accuracy of Uni-pK_a benefits from the dataset built under the protonation ensemble framework. In Table 11, ablation studies show that full microstates in the Dwar-iBond dataset in the finetuning stage improve the

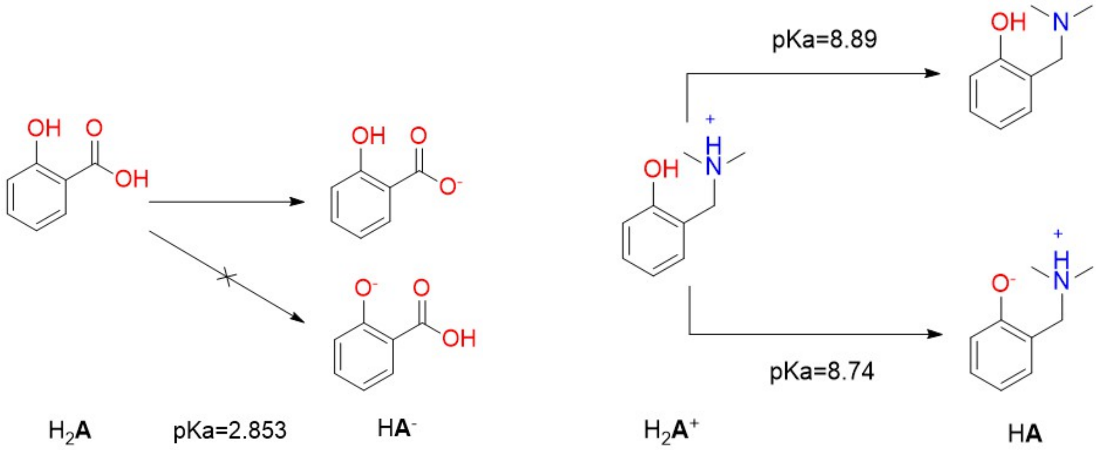


Figure 5: Two examples of the ionization site assignment: 2-hydroxybenzoic acid [43] (left) and 2-((dimethylamino)methyl)phenol [44] (right)

RMSE in the cross-validation and on most external datasets. As we have emphasized, a correct interpretation of data is key to the progression of the model. Our reconstructed datasets show chemical soundness as well as help the model to grasp the chemical properties.

In summary, modeling the complete protonation ensemble provides a stronger foundation for integrating experimental data, as quantified by the expansion of our reconstructed datasets. By preventing biased assumptions, ensemble-based modeling enables more accurate pK_a prediction.

5.2 Preserving Thermodynamic Consistency

Our protonation ensemble framework inherently preserves thermodynamic consistency between coupled pK_a values, while traditional independent site predictions risk inconsistencies. We demonstrated this through example protonation cycles and quantitative validation.

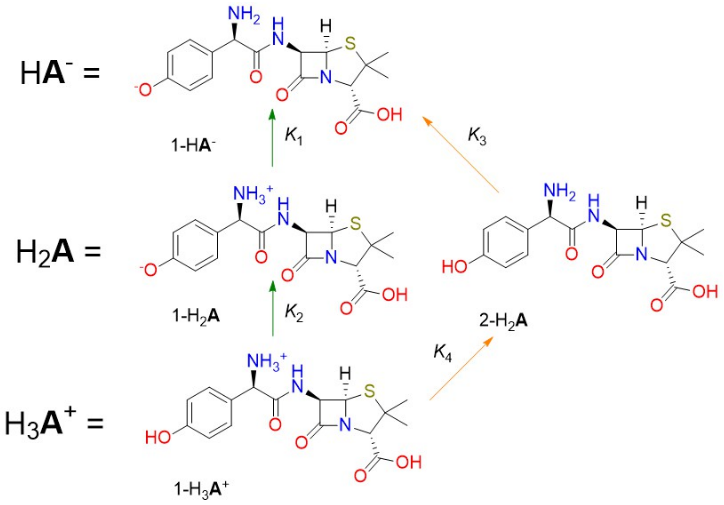


Figure 6: A thermodynamic cycle in the protonation of Amoxicillin. K_i is the dissociation equilibrium constant. The green and yellow arrows stand for different protonation routes.

As illustrated in Figure 6, the predicted micro-pK_as from Uni-pK_a automatically satisfy the thermodynamic relationship $K_1K_2 = K_3K_4$ for a sample molecule. In the microstate free energy modeling under the protonation ensemble framework, Uni-Mol in Uni-pK_a gives 4 free energies. We denote predicted equilibrium constant as $\tilde{K}_i$, predicted $\beta\Delta_{\mathrm{f}}G_{\mathrm{m}}^{\ominus}$ as g. As the workflow shown in Figure 4, the output of Uni-pK_{a} satisfies

$$\begin{aligned}\tilde{K}_1\tilde{K}_2 &= \exp\left\{-\left[g(1\text{-HA}^-) - g(1\text{-H}_2\text{A})\right]\right\}\cdot\exp\left\{-\left[g(1\text{-H}_2\text{A}) - g(1\text{-H}_3\text{A}^+)\right]\right\}\\ &= \exp\left\{-\left[g(1\text{-HA}^-) - g(1\text{-H}_3\text{A}^+)\right]\right\}\\ &= \exp\left\{-\left[g(1\text{-HA}^-) - g(2\text{-H}_2\text{A})\right]\right\}\cdot\exp\left\{-\left[g(2\text{-H}_2\text{A}) - g(1\text{-H}_3\text{A}^+)\right]\right\}\\ &= \tilde{K}_3\tilde{K}_4\end{aligned}$$

In contrast, baseline methods treating each site independently may violate this constraint, introducing unphysical results.

In conclusion, the protonation ensemble framework uniquely preserves thermodynamic constraints between interdependent pK_{a} values. By inherently encoding coupled equilibria, our approach provides a basis for rigorous pK_{a} prediction.

5.3 Model Accuracy and Generalizability

We evaluated Uni-pK_{a}'s performance on external datasets spanning diverse chemical spaces to assess generalizability. As summarized in Table 4, Uni-pK_{a} achieves state-of-the-art accuracy compared to recent chemoinformatics methods on the Novartis, SAMPL6, and SAMPL7 benchmarks.

Table 4: Performance on External Datasets

Method	Novartis				SAMPL6		SAMPL7	
	Acid		Base					
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Schrödinger Epik Classical [11][a,b]	0.99	1.531	0.876	1.175	0.784	0.962	1.121	1.648
ChemAxon Marvin[a,b]	0.808	1.144	0.835	1.145	1.007	1.248	0.559	0.708
ACD/Labs[b]	——	——	——	——	0.55	0.783	——	——
SPOC + XGBoost [14][c]	——	——	——	——	0.767	1.011	1.476	1.622
SPOC + NN [14][c]	——	——	——	——	0.832	1.141	0.932	1.156
OPERA [12][d]	——	——	——	——	0.97	1.283	2.135	2.515
MolGpKa [15][a,e]	0.849	1.287	0.789	1.064	**0.522**	0.773	0.797	0.98
GraphpKa [17][e]	——	——	——	——	0.594	0.726	0.758	0.934
MF-SuP-pKa [18][e,f]	0.85	1.09	0.61	0.79	0.687	0.751	0.656	0.816
Schrödinger Epik v7 [19][g]	——	——	——	——	——	0.92	——	——
Uni-pK_{a} (FC[h])	**0.810**	**1.061**	**0.493**	**0.653**	0.554	**0.716**	0.570	0.735
Uni-pK_{a} (GC[i])	0.846	1.109	0.550	0.697	0.719	0.959	0.487	0.609
Uni-pK_{a} (xC[j])	0.787	1.078	0.551	0.722	0.781	1.068	**0.433**	**0.603**

[a] Novartis Acid / Base results were reported in [15].
[b] SAMPL6 and/or SAMPL7 results were reported in the SAMPL challenge summaries [33, 34].
[c] SAMPL6 results were reported in original literature [14], and SAMPL7 results in [17].
[d] SAMPL6 and SAMPL7 results were reported in [17].
[e] SAMPL6 and SAMPL7 results were reported in [18].
[f] Novartis Acid / Base results were reported in original literature [18].
[g] SAMPL6 result was reported in original literature [19].
[h] using **F**ormal **C**harge in the pretraining, finetuning and inference stages, the default setting.
[i] using **G**asteiger **C**harge [45] in the pretraining, finetuning and inference stages.
[j] using **x**tb-GFN2 **C**harge [46] in the pretraining, finetuning and inference stages.

We also benchmarked against blind challenge submissions on the SAMPL8 dataset in Table 5. The submitted macro-pK_{a} results were not assigned to specific macrostate pairs in the challenge. We choose the macro-pK_{a}

prediction closest to the experimental value as the final answer of the submission. For Uni-pK_a, we manually assign experimental values to macrostate pairs based on chemical experience. Uni-pK_a significantly outperforms all entries, this provides an unbiased assessment of the predictive advantages.

Table 5: SAMPL8 Challenge

Submission Name	MAE	RMSE
Chemaxon	1.300	1.511
RobertRaddi_DeepGP	2.365	3.407
3DS	1.291	1.448
SabatinoRodriguezPaluch_uESE_extra	2.666	3.468
ZhiyiWu	3.212	4.642
ECRISM	1.545	2.420
Uni-pK_a (FC)	0.631	**0.878**
Uni-pK_a (GC)	0.642	0.949
Uni-pK_a (xC)	**0.619**	0.927

As detailed in Section 3.1 and Appendix B, the macrostates in the training set of Uni-pK_a are abridged to reduce the training consumption, while the microstates in the external test sets come directly from the full enumeration without any hand-pick. The biggest risk of the tandem of the enumerator and the neural network is that the unusual structures generated by radical enumeration are unfamiliar to the neural network trained on the pruned dataset. Therefore, the results also reveal the effectiveness of the lightweight training set, the reliability of the enumerator, and the extrapolation ability of the model, contributing to the performance of the whole prediction workflow.

In summary, experiments on standardized benchmarks demonstrate that the enumerator and the neural network in Uni-pK_a cooperate to achieve state-of-the-art accuracy compared to prior chemoinformatics techniques. The consistent improvements across heterogeneous evaluation sets validate the effectiveness of our protonation ensemble approach.

5.4 Comparison to Quantum Mechanics Methods

To evaluate Uni-pK_a's performance against rigorous first-principles calculations, we conducted comparative case studies with the state-of-the-art *ab initio* pK_a prediction software Jaguar.

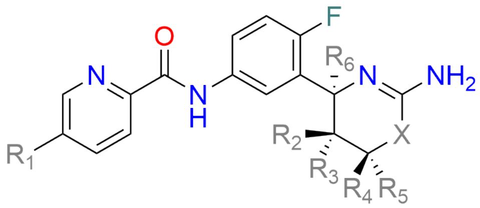
As shown in Tables 6, 7 and 8, Uni-pK_a demonstrates promising accuracy relative to Jaguar, given practical computational constraints. For example, without conformational sampling, Uni-pK_a matches or exceeds Jaguar's accuracy on a family of drug-like molecules in Table 7. This highlights the benefits of data-driven training on large datasets.

However, accuracy challenges remain for certain complex systems in Table 6, where Jaguar's accuracy improves significantly with exhaustive conformational modeling. While Uni-pK_a cannot match this, it provides a much faster alternative within reasonable tolerances for many applications.

While Jaguar's DFT calculations provide rigorous pK_a estimates, systematic errors remain. To compensate, Jaguar employs an empirical ”shell model” that assigns molecules to classes with parameterized corrections. However, this classification contains some arbitrariness, as the original authors note when evaluating guanidine derivatives in Table 8.

By default, these molecules fall under the guanidine shell. Yet the “partially substituted amidine” shell yields superior accuracy, with a mean absolute error of just 0.38 pK_a units without conformational sampling. The authors

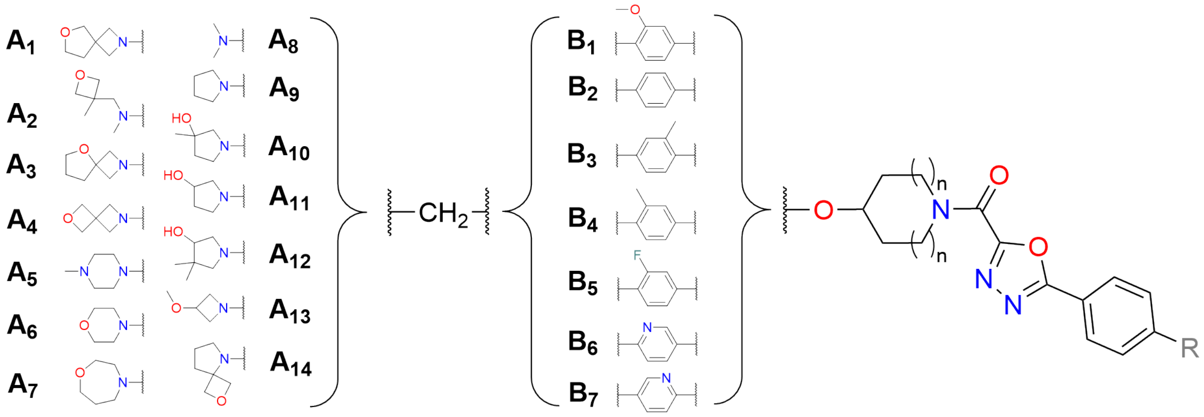
Table 6: Comparison between Uni-pK_a and Jaguar results: amidines in rings



R_1	R_2-R_5[a]	R_6	X	Exp. value	Uni-pK_a	Jaguar[b]	Jaguar+[c]	Jaguar++[d]
CN	——	CH_3	O	9.7	7.9	**9.1**	9.0	9.7
CN	$R_3=F$	CH_3	O	8.1	**6.7**	6.6	7.7	8.2
CN	$R_2=F$	CH_3	O	7.4	**6.8**	6.0	6.0	7.4
CN	——	CHF_2	O	7.3	5.9	**7.3**	7.3	7.7
CN	$R_4=CF_3$	CH_3	O	7.3	**6.7**	6.6	5.4	7.0
CN	$R_5=CF_3$	CH_3	O	7.0	6.6	**6.8**	6.8	7.3
CN	$R_3=F$	CH_2F	O	6.7	**6.0**	4.4	5.0	7.3
CN	$R_5=CF_3$	CH_2F	O	6.3	**5.8**	4.7	6.1	6.3
CN	$R_5=CF_3, R_3=F$	CH_3	O	5.9	**5.7**	5.5	5.8	6.0
CN	$R_4=CF_3, R_3=F$	CH_3	O	5.8	**5.7**	3.8	5.2	6.0
CN	$R_2=F, R_3=F$	CH_3	O	5.8	**5.5**	5.1	5.8	6.3
CN	$R_2=F, R_3=F$	CH_2F	O	5.1	**5.0**	3.2	5.5	4.7
Cl	——	CH_3	O	9.8	8.2	**9.2**	9.4	9.8
Cl	$R_2=F, R_3=F$	CH_3	O	6.3	**5.9**	5.1	6.2	6.4
Cl	——	CH_3	S	9.0	**8.0**	7.8	9.5	9.0
Cl	$R_2=F, R_3=F$	CH_3	S	6.1	**5.9**	5.3	6.1	6.3
MAE					**0.70**	1.07	0.56	0.20
Outlier number[e]					**5**	8	3	0

[a] H by default
[b] w/o conformational search
[c] w/ simple conformational search
[d] w/ comprehensive conformational search, weighting 10 conformers
[e] A prediction with the error larger than 1 pK_a unit is regarded as an outlier.

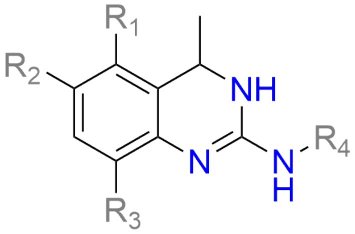
Table 7: Comparison between Uni-pK_a and Jaguar results: tertiary amines



A	B	*n*	R	Exp. value	Uni-pK_a	Jaguar[a]	Jaguar+	Jaguar++
A_9	B_2	0	OCH_3	9.9	9.13	**9.43**	8.13	9.11
A_8	B_7	0	OCH_3	8.5	**8.56**	7.94	8.23	7.90
A_9	B_7	1	OCH_3	9.2	**8.76**	8.23	8.68	8.69
A_9	B_6	1	OCH_3	9.1	8.86	9.02	9.30	**9.13**
A_8	B_5	0	OCH_3	8.6	**8.50**	7.81	9.04	8.23
A_8	B_4	0	OCH_3	9.3	8.89	8.49	8.34	**9.43**
A_8	B_3	0	H	9.3	**8.97**	8.68	8.62	8.66
A_5	B_2	1	OCH_3	8.5	**8.21**	6.60	7.96	8.18
A_5	B_5	0	OCH_3	8.4	**8.11**	7.34	7.03	7.79
A_5	B_5	0	H	8.4	**8.14**	7.80	7.22	7.78
A_6	B_2	1	OCH_3	6.5	**6.72**	7.96	7.91	8.25
A_6	B_1	0	OCH_3	7.5	6.88	8.50	7.04	**7.89**
A_{11}	B_2	1	H	8.9	8.26	8.33	8.12	**8.61**
A_{11}	B_2	1	OCH_3	8.9	8.27	**8.76**	9.05	8.29
A_{10}	B_2	0	H	9.0	8.52	**8.72**	8.33	8.62
A_{10}	B_2	0	OCH_3	9.0	8.56	8.61	**9.17**	8.32
A_{12}	B_2	1	OCH_3	9.0	7.85	8.24	9.41	**9.27**
A_{13}	B_2	0	OCH_3	8.5	7.29	8.20	**8.42**	8.00
A_2	B_2	0	OCH_3	8.0	8.36	**8.15**	7.69	7.11
A_2	B_2	0	H	8.5	**8.32**	6.54	8.22	7.71
A_7	B_2	0	OCH_3	8.1	7.93	**8.21**	7.56	6.47
A_{14}	B_2	0	OCH_3	6.0	7.68	7.78	7.07	**6.93**
A_1	B_2	0	OCH_3	8.6	7.98	**8.51**	8.32	8.53
A_1	B_2	0	H	8.6	7.91	8.36	**8.57**	8.38
A_3	B_2	0	OCH_3	8.4	7.25	7.98	7.88	**8.35**
A_4	B_2	0	OCH_3	8.2	7.76	7.83	**8.00**	7.96
A_4	B_2	0	H	8.0	7.73	7.63	7.80	**7.90**
MAE					0.52	0.69	0.56	**0.51**
Outlier number					4	6	5	**2**
Best number[b]					**9**	6	4	8

[a] The symbol is the same as 6, where the conformational search takes place in the vacuum.
[b] Frequency to be the best among 4 methods

Table 8: Comparison between Uni-pK_a and Jaguar results: guanidines in rings



R_1	R_2	R_3	R_4	Exp. value	Uni-pK_a	Jaguar[a]	Jaguar+	Jaguar++
Cl	H	H	H	9.9	**9.32**	8.62	8.70	8.70
Cl	H	H	CH_2CHF_2	8.9	**8.32**	7.02	6.93	7.43
H	H	H	CH_3	10.6	10.64	10.47	**10.58**	10.49
H	H	H	CH_2CHF_2	9.7	8.93	8.42	7.98	**8.61**
H	H	H	CH_2CF_3	9.2	**8.15**	7.36	6.65	7.65
Cl	Cl	H	CH_2CHF_2	8.5	**8.10**	6.17	6.23	6.80
H	H	OCH_3	CH_2CHF_2	10.2	8.72	**9.53**	9.07	9.23
OCH_3	Cl	H	CH_2CHF_2	8.9	**8.53**	6.68	7.19	6.65
Cl	H	Cl	CH_2CHF_2	7.8	**7.33**	5.00	5.26	6.11
MAE					**0.64**	1.60	1.68	1.34
Outlier number					**2**	7	8	7

[a] The symbol is the same as 6, where the conformational search takes place in the vacuum, and the default Guanidine shell is used for calibration.

suggest structural differences between the guanidine training set and these targets contribute to the discrepancy.

In contrast, Uni-pK_a adapts more flexibly across chemical spaces. Rather than human-crafted classes, it relies on automated pretraining over diverse data to incorporate chemical knowledge. While Uni-pK_a does not match the amidine shell's accuracy here, it still outperforms Jaguar's default corrections.

These benchmarks reveal a complementary synergy between the computational expense of QM methods and the data efficiency of machine learning techniques like Uni-pK_a. Integrating the two approaches to balance speed and accuracy is an exciting direction for future hybrid modeling.

In conclusion, comparisons to rigorous QM calculations substantiate Uni-pK_a's viability as an efficient surrogate for pK_a prediction, within limitations. Targeted integration of first-principles training data could help address areas for improvement revealed by QM benchmarks. This further motivates the development of unified ensemble modeling frameworks.

6 Conclusions

This work puts forward the protonation ensemble framework to enable machine learning models like our Uni-pK_a to represent acid-base equilibria with greater rigor and thermodynamic consistency. Uni-pK_a leverages pretraining on abundant inaccurate data and finetuning on curated experimental measurements to learn highly expressive molecular representations. By modeling microstate collections, it circumvents limitations of conventional independent site assumptions and improves the interpretation of macro-pK_a measurements.

We develop high-quality reconstructed datasets mapping macro-pK_a values to complete underlying microstate information. These rigorous benchmarks integrate chemical knowledge and experimental data to enable accurate machine learning.

Our microstate enumerator toolkit and prediction workflow unify speed and efficiency for tasks ranging from pK_a prediction to determining pH-dependent protonation fractions. In addition to improved performance over previous methods on small molecule datasets, Uni-pK_a demonstrates promising competitiveness against costly quantum chemistry calculations like Jaguar, given practical computational constraints.

However, accuracy challenges persist for certain complexes with subtle stereoelectronic effects like the proton sponge [47, 48] and Meldrum's acid [49, 50, 51, 52], not well represented in training data. Tautomerism also remains difficult to model comprehensively due to the scarcity of experimental data. Integrating data-driven techniques like our framework with first-principles training is an exciting path forward. Our protonation ensemble approach establishes a strong foundation for future synergistic hybrid modeling to address these interconnected equilibria.

By rethinking thermodynamics, data, and modeling under a unifying perspective, this work makes important progress in integrating chemical knowledge with machine learning. Our concepts, datasets, and techniques pave the way for continued advances in this productive fusion.

References

[1] Johann Nicolaus Brönsted. Einige bemerkungen über den begriff der säuren und basen. *Recueil des Travaux Chimiques des Pays-Bas*, 42(8):718–728, 1923.

[2] TM Lowry. The uniqueness of hydrogen. *Journal of the Society of Chemical Industry*, 42(3):43–47, 1923.

[3] David T Manallack, Richard J Prankerd, Elizabeth Yuriev, Tudor I Oprea, and David K Chalmers. The significance of acid/base properties in drug discovery. *Chemical Society Reviews*, 42(2):485–496, 2013.

[4] Wei Chen, Yuqing Deng, Ellery Russell, Yujie Wu, Robert Abel, and Lingle Wang. Accurate calculation of relative binding free energies between ligands with different net charges. *Journal of chemical theory and computation*, 14(12):6346–6358, 2018.

[5] César De Oliveira, Haoyu S Yu, Wei Chen, Robert Abel, and Lingle Wang. Rigorous free energy perturbation approach to estimating relative binding affinities between ligands with multiple protonation and tautomeric states. *Journal of chemical theory and computation*, 15(1):424–435, 2018.

[6] Art D Bochevarov, Mark A Watson, Jeremy R Greenwood, and Dean M Philipp. Multiconformation, density functional theory-based p k a prediction in application to large, flexible organic molecules with diverse functional groups. *Journal of chemical theory and computation*, 12(12):6001–6019, 2016.

[7] Haoyu S Yu, Mark A Watson, and Art D Bochevarov. Weighted averaging scheme and local atomic descriptor for p k a prediction based on density functional theory. *Journal of Chemical Information and Modeling*, 58(2):271–286, 2018.

[8] Oliver Wieder, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, and Thierry Langer. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37:1–12, 2020.

[9] Zhen Li, Mingjian Jiang, Shuang Wang, and Shugang Zhang. Deep learning methods for molecular representation and property prediction. *Drug Discovery Today*, page 103373, 2022.

[10] Louis P Hammett. Linear free energy relationships in rate and equilibrium phenomena. *Transactions of the Faraday Society*, 34:156–165, 1938.

[11] John C Shelley, Anuradha Cholleti, Leah L Frye, Jeremy R Greenwood, Mathew R Timlin, and Makoto Uchimaya. Epik: a software program for pk a prediction and protonation state generation for drug-like molecules. *Journal of computer-aided molecular design*, 21: 681–691, 2007.

[12] Kamel Mansouri, Neal F Cariello, Alexandru Korotcov, Valery Tkachenko, Chris M Grulke, Catherine S Sprankle, David Allen, Warren M Casey, Nicole C Kleinstreuer, and Antony J Williams. Open-source qsar models for pka prediction using multiple machine learning approaches. *Journal of Cheminformatics*, 11(1):1–20, 2019.

[13] Marcel Baltruschat and Paul Czodrowski. Machine learning meets pk a. *F1000Research*, 9, 2020.

[14] Qi Yang, Yao Li, Jin-Dong Yang, Yidi Liu, Long Zhang, Sanzhong Luo, and Jin-Pei Cheng. Holistic prediction of the pka in diverse solvents based on a machine-learning approach. *Angewandte Chemie*, 132(43):19444–19453, 2020.

[15] Xiaolin Pan, Hao Wang, Cuiyu Li, John ZH Zhang, and Changge Ji. Molgpka: A web server for small molecule p k a prediction using a graph-convolutional neural network. *Journal of Chemical Information and Modeling*, 61(7):3159–3165, 2021.

[16] Fritz Mayr, Marcus Wieder, Oliver Wieder, and Thierry Langer. Improving small molecule pk a prediction using transfer learning with graph neural networks. *Frontiers in Chemistry*, 10, 2022.

[17] Jiacheng Xiong, Zhaojun Li, Guangchao Wang, Zunyun Fu, Feisheng Zhong, Tingyang Xu, Xiaomeng Liu, Ziming Huang, Xiaohong Liu, Kaixian Chen, et al. Multi-instance learning of graph neural networks for aqueous p k a prediction. *Bioinformatics*, 38(3):792–798, 2022.

[18] Jialu Wu, Yue Wan, Zhenxing Wu, Shengyu Zhang, Dongsheng Cao, Chang-Yu Hsieh, and Tingjun Hou. Mf-sup-pka: multi-fidelity modeling with subgraph pooling mechanism for pka prediction. *Acta Pharmaceutica Sinica B*, 2022.

[19] Ryne C. Johnston, Kun Yao, Zachary Kaplan, Monica Chelliah, Karl Leswing, Sean Seekins, Shawn Watts, David Calkins, Jackson Chief Elk, Steven V. Jerome, Matthew P. Repasky, and John C. Shelley. Epik: pka and protonation state prediction through machine learning. *Journal of Chemical Theory and Computation*, 19(8):2380–2388, 2023.

[20] Jialu Wu, Yu Kang, Peichen Pan, and Tingjun Hou. Machine learning methods for pka prediction of small molecules: Advances and challenges. *Drug Discovery Today*, page 103372, 2022.

[21] Mehtap Işık, Dorothy Levorse, Ariën S Rustenburg, Ikenna E Ndukwe, Heather Wang, Xiao Wang, Mikhail Reibarkh, Gary E Martin, Alexey A Makarov, David L Mobley, et al. p k a measurements for the sampl6 prediction challenge for a set of kinase inhibitor-like fragments. *Journal of computer-aided molecular design*, 32:1117–1138, 2018.

[22] Adam C Lee and Gordon M Crippen. Predicting p k a. *Journal of chemical information and modeling*, 49(9):2013–2033, 2009.

[23] Matthias Rupp, Robert Korner, and Igor V Tetko. Predicting the pka of small molecules. *Combinatorial chemistry & high throughput screening*, 14(5):307–327, 2011.

[24] Lewis J Leeson, James E Krueger, and Robert A Nash. Concerning the structural assignment of the second and third acidity constants of the tetracycline antibiotics. *Tetrahedron Letters*, 4(18):1155–1160, 1963.

[25] Marilyn R Gunner, Taichi Murakami, Ariën S Rustenburg, Mehtap Işık, and John D Chodera. Standard state free energies, not pk as, are ideal for describing small molecule protonation and tautomeric states. *Journal of computer-aided molecular design*, 34:561–573, 2020.

[26] John Tileston Edsall and Jeffries Wyman. *Biophysical chemistry. Vol. 1, Themrodynamics, electrostatics, and the biological significance of the properties of matter*. Academic Press, 1958.

[27] Thomas Sander, Joel Freyss, Modest von Korff, and Christian Rufener. Datawarrior: an open-source program for chemistry aware data visualization and analysis. *Journal of chemical information and modeling*, 55(2):460–473, 2015.

[28] Internet bond-energy databank (pka and bde)–ibond home page., 2017. http://ibond.nankai.edu.cn.

[29] Antoine Wallabregue, Petr Sherin, Joyram Guin, Celine Besnard, Eric Vauthey, and Jérôme Lacour. Modular synthesis of ph-sensitive fluorescent diaza [4] helicenes. *European Journal of Organic Chemistry*, 2014(29):6431–6438, 2014.

[30] Heinz A Staab and Thomas Saupe. "proton sponges" and the geometry of hydrogen bonds: aromatic nitrogen bases with exceptional basicities. *Angewandte Chemie International Edition in English*, 27(7):865–879, 1988.

[31] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1): D1100–D1107, 2012.

[32] Chenzhong Liao and Marc C Nicklaus. Comparison of nine programs predicting p k a values of pharmaceutical substances. *Journal of chemical information and modeling*, 49(12):2801–2812, 2009.

[33] Mehtap Işık, Ariën S Rustenburg, Andrea Rizzi, Marilyn R Gunner, David L Mobley, and John D Chodera. Overview of the sampl6 p k a challenge: evaluating small molecule microscopic and macroscopic p k a predictions. *Journal of computer-aided molecular design*, 35 (2):131–166, 2021.

[34] Teresa Danielle Bergazin, Nicolas Tielker, Yingying Zhang, Junjun Mao, Marilyn R Gunner, Karol Francisco, Carlo Ballatore, Stefan M Kast, and David L Mobley. Evaluation of log p, p k a, and log d predictions from the sampl7 blind challenge. *Journal of computer-aided molecular design*, 35(7):771–802, 2021.

[35] Aakankschit Nandkeolyar, Matthew N Bahr, and David L Mobley. Insights from the sampl8 physical properties blind prediction challenge. *Biophysical Journal*, 122(3):423a, 2023.

[36] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[39] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask

learners. *OpenAI blog*, 1(8):9, 2019.

[41] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

[42] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[43] Ali Farajtabar and Farrokh Gharib. Solvent effect on protonation constants of salicylic acid in mixed aqueous organic solutions of dmso. *Monatshefte für Chemie-Chemical Monthly*, 141:381–386, 2010.

[44] AB Teitel'baum, KA Derstuganova, NA Shishkina, LA Kudryavtseva, VE Bel'skii, and BE Ivanov. Tautomerism in the ortho-aminomethylphenols. *Bulletin of the Academy of Sciences of the USSR, Division of chemical science*, 29:558–562, 1980.

[45] Johann Gasteiger and Mario Marsili. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, 36(22):3219–3228, 1980.

[46] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation*, 15(3):1652–1671, 2019.

[47] Lee Belding, Peter Stoyanov, and Travis Dudding. Synthesis, theoretical analysis, and experimental p k a determination of a fluorescent, nonsymmetric, in–out proton sponge. *The Journal of organic chemistry*, 81(1):6–13, 2016.

[48] Davor Margetić, Tsutomu Ishikawa, and Takuya Kumamoto. Exceptional superbasicity of bis (guanidine) proton sponges imposed by the bis (secododecahedrane) molecular scaffold: A computational study. *European Journal of Organic Chemistry*, 34(2010):6563–6572, 2010.

[49] Edward M Arnett and John A Harrelson Jr. Ion pairing and reactivity of enolate anions. 7. a spectacular example of the importance of rotational barriers: the ionization of meldrum's acid. *Journal of the American Chemical Society*, 109(3):809–812, 1987.

[50] Xuebao Wang and KN Houk. Theoretical elucidation of the origin of the anomalously high acidity of meldrum's acid. *Journal of the American Chemical Society*, 110(6):1870–1872, 1988.

[51] Kyoungrim Byun, Yirong Mo, and Jiali Gao. New insight on the origin of the unusual acidity of meldrum's acid from ab initio and combined qm/mm simulation study. *Journal of the American Chemical Society*, 123(17):3974–3979, 2001.

[52] Satoshi Nakamura, Hajime Hirao, and Tomohiko Ohwada. Rationale for the acidity of meldrum's acid. consistent relation of c- h acidities to the properties of localized reactive orbital. *The Journal of Organic Chemistry*, 69(13):4309–4316, 2004.

[53] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):1–14, 2011.

[54] Rdkit: Open-source cheminformatics. URL https://www.rdkit.org.

[55] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

A pH-dependent Free Energy

A.1 The definition of pH-dependent free energy

We start from the fundamental acid/base equilibrium theory. Considering the simplest acidic ionization reaction in water,

$$\mathrm{HA^+(aq)} \rightleftharpoons \mathrm{H^+(aq)} + \mathrm{A\,(aq)}, \quad K_\mathrm{a}^\ominus = \frac{[\mathrm{H^+}][\mathrm{A}]}{[\mathrm{HA^+}]},$$

its standard molar Gibbs free energy change of the reaction is

$$\Delta_\mathrm{r}G_\mathrm{m}^\ominus = \Delta_\mathrm{f}G_\mathrm{m}^\ominus(\mathrm{A\,(aq)}) - \Delta_\mathrm{f}G_\mathrm{m}^\ominus(\mathrm{HA^+(aq)}) = -\frac{1}{\beta}\ln K_\mathrm{a}^\ominus, \quad \beta = \frac{1}{RT},$$

to which $\mathrm{p}K_\mathrm{a}$ is related as

$$\mathrm{p}K_\mathrm{a} = \frac{\beta}{\ln 10}\Delta_\mathrm{r}G_\mathrm{m}^\ominus.$$

The $\Delta_\mathrm{r}G_\mathrm{m}^\ominus$ above can be regarded as the free energy difference between the thermodynamic standard state of A (aq) and $\mathrm{HA^+}$(aq) in the presence of the thermodynamic standard state of $\mathrm{H^+}$(aq) (pH = 0). Therefore, the pH-related free energy difference is

$$\Delta_\mathrm{r}G_\mathrm{m}(\mathrm{pH}) = \Delta_\mathrm{r}G_\mathrm{m}^\ominus + RT\ln[\mathrm{H^+(aq)}] = \Delta_\mathrm{r}G_\mathrm{m}^\ominus - \frac{\ln 10}{\beta}\mathrm{pH}.$$

If we define the pH-dependent free energy of $\mathrm{AH^+}$(aq) as

$$\Delta_\mathrm{f}G_\mathrm{m}(\mathrm{HA^+};\mathrm{pH}) = \Delta_\mathrm{f}G_\mathrm{m}^\ominus(\mathrm{A\,(aq)}) - \Delta_\mathrm{r}G_\mathrm{m}(\mathrm{pH}) = \Delta_\mathrm{f}G_\mathrm{m}^\ominus(\mathrm{AH^+(aq)}) + \frac{\ln 10}{\beta}\mathrm{pH},$$

then it can be generalized to the case of multi-protonated acid, that is,

$$\mathrm{H}_m\mathrm{A}^{m+}\mathrm{(aq)} \rightleftharpoons \mathrm{mH^+(aq)} + \mathrm{A\,(aq)}, \quad \Delta_\mathrm{r}G_\mathrm{m}(\mathrm{pH}) = \Delta_\mathrm{r}G_\mathrm{m}^\ominus + \frac{m}{\beta}\ln[\mathrm{H^+(aq)}]$$

and we have

$$\Delta_\mathrm{f}G_\mathrm{m}(\mathrm{H}_m\mathrm{A}^{m+};\mathrm{pH}) = \Delta_\mathrm{f}G_\mathrm{m}^\ominus(\mathrm{A\,(aq)}) - \Delta_\mathrm{r}G_\mathrm{m}(\mathrm{pH}) = \Delta_\mathrm{f}G_\mathrm{m}^\ominus(\mathrm{H}_m\mathrm{A}^{m+}\mathrm{(aq)}) + \frac{m\ln 10}{\beta}\mathrm{pH}, \tag{7}$$

in which there is a linear relationship between the free energy and pH, with a slope proportional to the net charge.

A.2 The relationship between MIL and the proton ensemble learning framework

MIL framework implements the macro-micro $\mathrm{p}K_\mathrm{a}$ formula in the model training [17, 18]. For acidic dissociation, like 2-$\mathrm{H_2A}$ to 1-$\mathrm{HA^-}$ (phenolic hydroxyl group) and 2-$\mathrm{HA^-}$ (carboxylic group) in Figure 1, the formula is

$$\mathrm{p}K_{\mathrm{a},m} = -\log_{10}\left(\sum_k 10^{-\mathrm{p}K_{\mathrm{a},m}^k}\right) \tag{8}$$

For basic dissociation, like 2-$\mathrm{HA^-}$ to 2-$\mathrm{H_2A}$ (carboxylic group) and 3-$\mathrm{H_2A}$ (amino group) in Figure 1, the formula is

$$\mathrm{p}K_{\mathrm{a},m} = \log_{10}\left(\sum_i 10^{\mathrm{p}K_{\mathrm{a},m}^i}\right) \tag{9}$$

This framework successfully handles the example on the right of Figure 5 with the formula 8. However, it still cannot deal with more complex molecules. Like the case of Amoxicillin $\mathrm{H_2A}$ to $\mathrm{HA^-}$ in Figure 1, starting from any microstate, MIL always overlooks three in the six microstates, whichever the initial microstate is chosen. Furthermore, due to the difference between formula 8 and 9, two different models are responsible for acidic and basic cases, which increases the training cost and loses chemical information of bidirected ionization.

Formula 8 and 9 are accommodated in our proton ensemble framework as special cases where there is only one microstate in a macrostate. If i is unique in the summation of the denominator (A macrostates) in the formula 2 and is thus omitted, we have

$$\begin{aligned}
\mathrm{p}K_{\mathrm{a},m} &= -\log_{10}\frac{[\mathrm{H}^+]\sum_k[k\text{-}\mathrm{H}_m\mathrm{A}^{m+}]}{[\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+}]} = -\log_{10}\left(\sum_k\frac{[\mathrm{H}^+][k\text{-}\mathrm{H}_m\mathrm{A}^{m+}]}{[\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+}]}\right)\\
&= -\log_{10}\left(\sum_k K^k_{\mathrm{a},m}\right) = \log_{10}\left(\sum_k 10^{-\mathrm{p}K^k_{\mathrm{a},m}}\right)
\end{aligned}$$

deriving Equation 8. If i is unique in the summation of the numerator (B macrostates) in the formula 2 and is thus omitted, we have

$$\begin{aligned}
\mathrm{p}K_{\mathrm{a},m} &= -\log_{10}\frac{[\mathrm{H}^+][\mathrm{H}_m\mathrm{A}^{m+}]}{\sum_i[i\text{-}\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+}]} = \log_{10}\frac{\sum_i[i\text{-}\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+}]}{[\mathrm{H}^+][\mathrm{H}_m\mathrm{A}^{m+}]} = \log_{10}\left(\sum_i\frac{[i\text{-}\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+}]}{[\mathrm{H}^+][\mathrm{H}_m\mathrm{A}^{m+}]}\right)\\
&= \log_{10}\left(\sum_i\frac{1}{K^i_{\mathrm{a},m}}\right) = \log_{10}\left(\sum_i 10^{\mathrm{p}K^i_{\mathrm{a},m}}\right)
\end{aligned}$$

deriving Equation 9.

A.3 Label translation in Uni-pK_a framework

It is obvious that $\Delta_\mathrm{f}G^\ominus_\mathrm{m} = \Delta_\mathrm{f}G_\mathrm{m}(\mathrm{pH}=0)$. Hence, the macro p$K_\mathrm{a}$-free energy formula can be generalized as

$$\begin{aligned}
&-\log_{10}\frac{\sum_i\exp\left(-\beta\Delta_\mathrm{f}G_\mathrm{m}(i\text{-}\mathrm{H}_m\mathrm{A}^{m+};\mathrm{pH}=t)\right)}{\sum_i\exp\left(-\beta\Delta_\mathrm{f}G_\mathrm{m}(i\text{-}\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+});\mathrm{pH}=t\right)}\\
&=-\log_{10}\frac{\sum_i\exp\left(-\beta\Delta_\mathrm{f}G^\ominus_\mathrm{m}(i\text{-}\mathrm{H}_m\mathrm{A}^{m+})-m\ln10\cdot t\right)}{\sum_i\exp\left(-\beta\Delta_\mathrm{f}G^\ominus_\mathrm{m}(i\text{-}\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+})-(m+1)\ln10\cdot t\right)}\\
&=-\log_{10}\frac{10^{-mt}\sum_i\exp\left(-\beta\Delta_\mathrm{f}G^\ominus_\mathrm{m}(i\text{-}\mathrm{H}_m\mathrm{A}^{m+})\right)}{10^{-(m+1)t}\sum_i\exp\left(-\beta\Delta_\mathrm{f}G^\ominus_\mathrm{m}(i\text{-}\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+})\right)}\\
&=-\log_{10}\frac{10^{-t}\sum_i\exp\left(-\beta\Delta_\mathrm{f}G^\ominus_\mathrm{m}(i\text{-}\mathrm{H}_m\mathrm{A}^{m+})\right)}{\sum_i\exp\left(-\beta\Delta_\mathrm{f}G^\ominus_\mathrm{m}(i\text{-}\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+})\right)}\\
&=-\log_{10}\frac{\sum_i\exp\left(-\beta\Delta_\mathrm{f}G^\ominus_\mathrm{m}(i\text{-}\mathrm{H}_m\mathrm{A}^{m+})\right)}{\sum_i\exp\left(-\beta\Delta_\mathrm{f}G^\ominus_\mathrm{m}(i\text{-}\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+})\right)}+t\\
&=\mathrm{p}K_{\mathrm{a},m}+t
\end{aligned}$$

and that is why the physical meaning of Uni-Mol's raw output is preserved when the pH label is translated with a constant t.

B Datasets

- **ChemBL dataset**: There are a bulk of small organic molecules in ChemBL [31] with predicted pK_a values by ChemAxon software. Our version is based on the edited version and the protonation templates in MF-SuP-pKa [18]. The core structure and initial net charge are directly determined by the initial one, and the ionization mode is determined by the acid/base classification in MF-SuP-pKa. A simple enumerator, which only iterates once from the initial structure, is chosen to adapt to the data size and quality.
- **Dwar-iBond dataset**: pKaInWater.dwar is a DataWarrior [27] sample dataset containing initial structures, color-labeled ionization sites and modes, and measuring methods. It has been used for model training in several works [12, 16, 17]. i-Bond [28] is a bond energy database with pK_a subdatabase, containing initial structures, color-labeled ionization sites, measuring methods, and references. We combine almost all entries from pKaInWater and selected entries from i-Bond. The core structure is directly determined by

the initial one, while the initial net charge and the ionization mode are judged and curated comprehensively with chemical sense by the initial structure, pK_a value, color label, and original reference. Our microstate enumerator is fully used. The enumeration results are further cleaned to exclude chemically unreasonable and non-contributive structures.

- **Novartis Acid / Base dataset**: As a dataset mainly containing drug-like molecules originally released by Liao et al. [32], it was commonly used for external tests in previous studies [12, 13, 15, 16, 18, 19]. We follow the separation of acid and base in MF-SuP-pKa and determine the ionization mode. The core structure is directly determined by the initial one, and the initial net charge is judged and curated comprehensively with chemical sense by the initial structure and pK_a value. Our microstate enumerator is fully used.
- **SAMPL6, SAMPL7, and SAMPL8 dataset**: Macro-pK_a prediction is one of the tracks in SAMPL challenges and has attracted a variety of approaches [33, 34, 35], and is also used for external tests in previous studies [14, 17, 18, 19]. The core structure is directly determined by the initial one after washing off counterions, while the initial net charge and the ionization mode are judged and curated comprehensively with chemical sense by the initial structure and pK_a value. Our microstate enumerator is fully used, instead of existing microstates given in the challenge repositories.

In the released dataset, a new standard table format is to store the microstates after enumeration. If the structure is represented by SMILES [55], every entry has a SMILES field A1,A2,...>>B1,B2..., where SMILES like A1 and B1 on both sides of the arrow >> respectively correspond to the structures in A and B Micro Pool after enumeration and necessary curation. As shown in Table 2, this format naturally stores all microstates enumerated for macro-pK_a data (like in the first line), and is compatible with specified microstates for micro-pK_a data, as same as the case of only one microstate in both macrostates (like in the second line).

C Molecular preprocessing

If starting with SMILES like in Table 2, 2D and 3D conformers are generated from the SMILES by RDkit [54] or OpenBabel [53] at first. The atom-type list, atomic charge list, and distance matrix are calculated and packaged as the input of the Uni-Mol model. The workflow supported 3 sources of atomic charge:

- **Formal Charge**: discrete values of 0 or ±1 directly read in covalent structures. It is our default setup.
- **Gasteiger Charge**: continuous values calculated by empirical rules [45]. It is not supported in RDkit for molecules containing uncommon elements like selenium and arsenic, and in these cases, the formal charge is used instead.
- **xtb-GFN2 Charge**: continuous values obtained after quantum chemistry geometry optimization and single point energy calculations. xtb-GFN2 [46], as our first choice of quantum chemistry in the workflow, is a semi-empirical tight-binding DFT supporting a wide range of molecular systems with satisfying speed, stability, and usability. Partial charges from other quantum chemistry program packages and levels of theory are also supported if needed and available.

For the special atom-type word like [CLS] dressed at the tail of the list, the corresponding continuous charge is set to zero.

Table 9: Uni-pK_a hyperparameters setup during pretraining

Hyperparameter	Pretraining
Layers	15
Peak learning rate	1e-4
Batch size	128
Max training epoches	100
Warmup ratio	0.06
Attention heads	64
FFN dropout	0.1
Attention dropout	0.1
Embedding dropout	0.1
Weight decay	1e-4
Embedding dim	512
FFN hidden dim	2048
Gaussian kernel channels	128
Corrupt ratio	0.15
Activation function	GELU
Learning rate decay	Linear
Adams ϵ	1e-6
Adams (β_1, β_2)	(0.9, 0.99)
Gradient clip norm	1.0
pK_a loss function and its weight	MSE, 1.0
Atom loss function and its weight	Cross entropy, 1.0
Charge loss function and its weight	Cross entropy, 1.0
Coordinate loss function and its weight	Smooth L1, 2.0
Distance loss function and its weight	Smooth L1, 1.0
Vocabulary size (atom types)	30
Vocabulary size (charge types)	7

Table 10: Search space for finetuning

Hyperparameter	Finetuning
Learning rate	[3e-4, 1e-4, 5e-5]
Batch size	[16, 32, 64, 128]
Epochs	[20 40 60 80]
Pooler dropout	0.1
Warmup ratio	0.06

D Experiments details & reproduce

D.1 Pretraining setup

We report the detailed hyperparameters setup of Uni-pK_a during pretraining in Table D. Uni-pK_a pretraining loss is summed up by five components, pK_a loss, atom (token) loss, charge loss, coordinate loss, and pair-distance loss. For pK_a, Uni-pK_a predict it through the molecular [CLS] token. Atoms and charges are masked, and noise is added to coordinate as described in Sec. 4.3. Predicting pK_a of the whole data points is not an easy task in the presence of masking and adding noise within the data points, where molecular information is incomplete. Compared the molecular pretraining in Uni-Mol, we reduce the multiples of coordinate loss and distance loss. At the same time, in order to make them have a similar influence, we appropriately enlarge the coordinate loss. For the sake of training stability, we also retain the norm loss in Uni-Mol with a very small weight of 0.01, which will not affect the model results. The pretraining runs on 8 V100 GPUs (32GB memory, the same below), and the training time is about 1 day and 22 hours.

D.2 Finetuning setup

Data split In our experiments, referring to previous work [15, 17, 18], we use use a 5-fold cross-validation splitting to divide the dataset into training and validation. In all experiments, we select the checkpoint with the best validation metric for each fold separately and report the average metric across 5 folds.

Hyperparameter search space Referring to previous works, we use a grid search to find the best combination of hyperparameters in finetuning. The specific search space is shown in Table 10. And we run finetuning on a single V100 GPU.

D.3 Inference setup

As mentioned in Sec. 4.4, we generate multiple conformations for a molecule to enhances both performance and robustness of the model. During the inference stage, in the pK_a prediction task, the free energy predicted for different conformations of the same molecule under the same model is averaged. Then, the FE2pK_a module is used to predict the pK_a. We average the pK_a results predicted by the finetuned 5-fold models as the final prediction result.

In the case of distribution fraction prediction, since only the free energy predicted by the model is needed, we directly average the free energy predicted by the 5-fold models as the final free energy result.

E Ablation Experiments

Table 11: RMSE on experimental datasets with different settings

Settings				Datasets					
Charge	Microstate	Pretraining	Label	Dwar-iBond[a]	Novartis		SAMPL		
					Acid	Base	6	7	8
Formal[b]	Full[c]	ChemBL[d]	Translated[e]	0.883	1.061	0.653	0.716	0.735	0.878
None[f]	Full	ChemBL	Translated	0.941	1.115	0.696	0.897	0.651	0.849
Formal	Single[g]	ChemBL	Translated	0.925	1.234	0.749	0.845	0.678	0.927
Formal	Full	Uni-Mol[h]	Translated	0.910	1.373	0.845	0.845	0.818	0.927
Formal	Full	None[i]	Translated	1.151	1.702	1.431	1.157	1.867	1.159
Formal	Full	ChemBL	Original[j]	0.980	1.022	0.670	0.911	0.557	0.860

[a] 5-fold cross validation

[b] The same as FC in Table 4 and 5

[c] Using all microstates provided by our Dwar-iBond dataset in the finetuning stage.

[d] Following the pretraining strategy in the main text.

[e] The pK_a label is preprocessed by a translation of average of the dataset in the pretraining and finetuning stage.

[f] Without Charge Repr. in Figure 3 in pretraining, finetuning and inference stages.

[g] Using only one microstate picked with chemical knowledge in the finetuning stage.

[h] Finetuned upon the pretrained model released by Uni-Mol [36].

[i] Without the pretraining stage.

[j] The pK_a label is not preprocessed in pretraining and finetuning stage.