JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

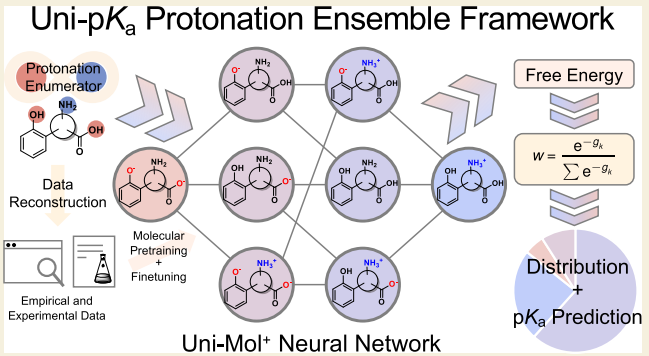
pubs.acs.org/jacsau
Article

Bridging Machine Learning and Thermodynamics for Accurate pK_a Prediction

Weiliang Luo,[‖] Gengmo Zhou,[‖] Zhengdan Zhu, Yannan Yuan, Guolin Ke, Zhewei Wei, Zhifeng Gao,* and Hang Zheng*

Cite This: *JACS Au* 2024, 4, 3451−3465
Read Online
ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Integrating scientific principles into machine learning models to enhance their predictive performance and generalizability is a central challenge in the development of AI for Science. Herein, we introduce Uni-pK_a, a novel framework that successfully incorporates thermodynamic principles into machine learning modeling, achieving high-precision predictions of acid dissociation constants (pK_a), a crucial task in the rational design of drugs and catalysts, as well as a modeling challenge in computational physical chemistry for small organic molecules. Uni-pK_a utilizes a comprehensive free energy model to represent molecular protonation equilibria accurately. It features a structure enumerator that reconstructs molecular configurations from pK_a data, coupled with a neural network that functions as a free energy predictor, ensuring high-throughput, data-driven prediction while preserving thermodynamic consistency. Employing a pretraining-finetuning strategy with both predicted and experimental pK_a data, Uni-pK_a not only achieves state-of-the-art accuracy in chemoinformatics but also shows comparable precision to quantum mechanics-based methods.



KEYWORDS: *pK_a, machine learning, protonation ensemble, pretraining-finetuning strategy, free energy modeling, chemical thermodynamics*

1. INTRODUCTION

Machine learning's integration into scientific research, known as AI for Science, has greatly improved our problem-solving capabilities.[1,2] However, challenges of accuracy and generalizability from issues with data quantity and quality persist, requiring innovative solutions.[3−5] The fusion of established scientific principles with advanced machine learning is essential, enhancing model performance, robustness, and versatility in scientific applications.[6,7]

The acid dissociation constant (pK_a) is a complex modeling and calculation challenge due to the intricate chemical equilibria among various protonated forms of a molecule. While machine learning excels in accuracy and speed for individual molecular properties, pK_a prediction is complicated by these equilibria, posing difficulties for experimental measurements, quantum chemical calculations, and machine learning methods alike. To construct robust models, it is imperative to integrate scientific principles into machine learning at multiple stages of the modeling process.

The prediction of pK_a is one of the foundations of accurate chemical modeling in computer-aided molecular discovery. In particular, ubiquitous acid−base equilibrium adds complexity to molecular structures in water, which is of great concern in chemical, material, health, and environmental sciences. Functional molecules universally contain acidic/basic chemical groups such as carboxyl groups, amino groups, and *N*-heterocyclic rings, where pK_a is the key physical chemistry parameter describing their acid−base equilibria. pK_a serves as an informative descriptor for designing catalysis systems[8−10] and environmental impact assessment.[11] From a drug discovery perspective, it directly determines structures in physiological environments, influencing key properties such as solubility, membrane permeability, and biomolecular interactions. As such, pK_a prediction also plays an important role in screening drug-like molecules with optimal pharmacokinetics, toxicity, and activity.[12] As seen in free energy perturbation calculations, a molecular simulation method assessing activity, accurate pK_a values also enable proper structure preparation and thermodynamic correction and improve the accuracy.[13,14] Therefore, fast and reliable pK_a prediction approaches are

Received: March 26, 2024
Revised: July 7, 2024
Accepted: July 10, 2024
Published: July 17, 2024

highly valuable in various applications of molecular discovery and production.

Due to the prevalence of multiple ionizable groups within functional molecules, framing pK_a prediction as a simple multilabel regression problem with individual site labels overlooks its complexity, where both global and local structures must be encoded, and polyprotonated and amphoteric cases should be handled. With this consideration, recent chemoinformatics works use different descriptions of the molecular structure and ionization:[15] (1) Template-based methods utilize ionization site matching to empirical fragment values, along with correction of surrounding structural context by Hammett linear free energy relationships,[16] as implemented in early versions of Epik.[17] (2) Local atomic descriptors represent ionization sites, while global molecular descriptors cover full structures in traditional machine learning techniques, including OPERA,[18] the work of Baltruschat and Czodrowski,[19] and SPOC.[20] (3) Graph neural networks learn hierarchical embeddings of sites and structures at different levels of molecular graphs, as demonstrated by MolGpKa,[21] pKasolver,[22] Graph-pK_a,[23] MF-SuP-pK_a,[24] and Epik 7.[25]

However, fundamental limitations remain in interpreting experimental data and ensuring thermodynamic consistency. On the data side, most pK_a measurements reflect coupled equilibria[26,27] but are often ascribed to one dominant equilibrium in data sets and algorithms, inducing bias.[28] As has been discussed for decades, rigorous interpretation requires contributions from all equilibria.[29] Recent attempts like the multi-instance learning (MIL) framework proposed by Xiong et al.[23,24] accommodate multiple ionization sites but still ignore complex protonation networks. On the model side, thermodynamic coupling emerges when it comes to the modeling of polyprotonation.[30] Independent site modeling loses the awareness of the coupling, compromises the rigor, and risks self-contradiction when predicting the distribution of protonation states.[25] Under the strong demand for the protonation state ranking of given molecules, genuinely self-consistent pK_a prediction remains an unmet need. These intertwined limitations of current approaches underscore the need for a new modeling perspective. A recent study on the SAMPL6 challenge highlights the necessity of using standard free energies rather than pK_as when representing complex protonation systems.[30]

This preliminary effort has validated the feasibility of applying scientific principles to the prediction of pK_a values. Inspired by such works, we introduce Uni-pK_a, a protonation-ensemble-based framework bridging thermodynamics and machine learning. We have reorganized pK_a data of varying types and accuracies and meticulously crafted both the modeling and training methodologies. Ultimately, we encapsulate these elements within a unified representational framework designed to predict a diverse array of pK_a-related tasks.

On the data side, we design a general format of the pK_a data set, which stores the determined molecular structure of protonation states and is compatible with all kinds of pK_a measurements. We reconstruct several publicly available data sets in this format and release them as a new, fine-grained 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 pK_a from different measurements, naturally preserves thermodynamic consistency, and enables multiple scenarios, including pK_a prediction and protonation state scoring. 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 comprehensively generate protonation states from given molecules. It helps to build the data set and propose a workflow for structure preparation in molecular simulation, combining speed and accuracy.

In conclusion, by bridging scientific principles with advanced machine learning techniques from representation to modeling, we have established Uni-pK_a, a robust framework that not only accommodates diverse pK_a measurements but also upholds the fundamental laws of thermodynamics, thereby providing a versatile tool for both pK_a prediction and protonation state evaluation.

2. RESULTS

2.1. Overview of the Uni-pK_a Framework

In Brønsted-Lowry acid−base theory,[31,32] the acidity and basicity of a molecule in an aqueous solution are defined by the mutual protonation between itself and the water molecule. pK_a quantitatively describes the extent and direction of protonation equilibria. Interpreting experimental pK_a measurements requires modeling the underlying chemical structures and equilibria, which are differently detected in bulk experimental techniques.[26,27,33]

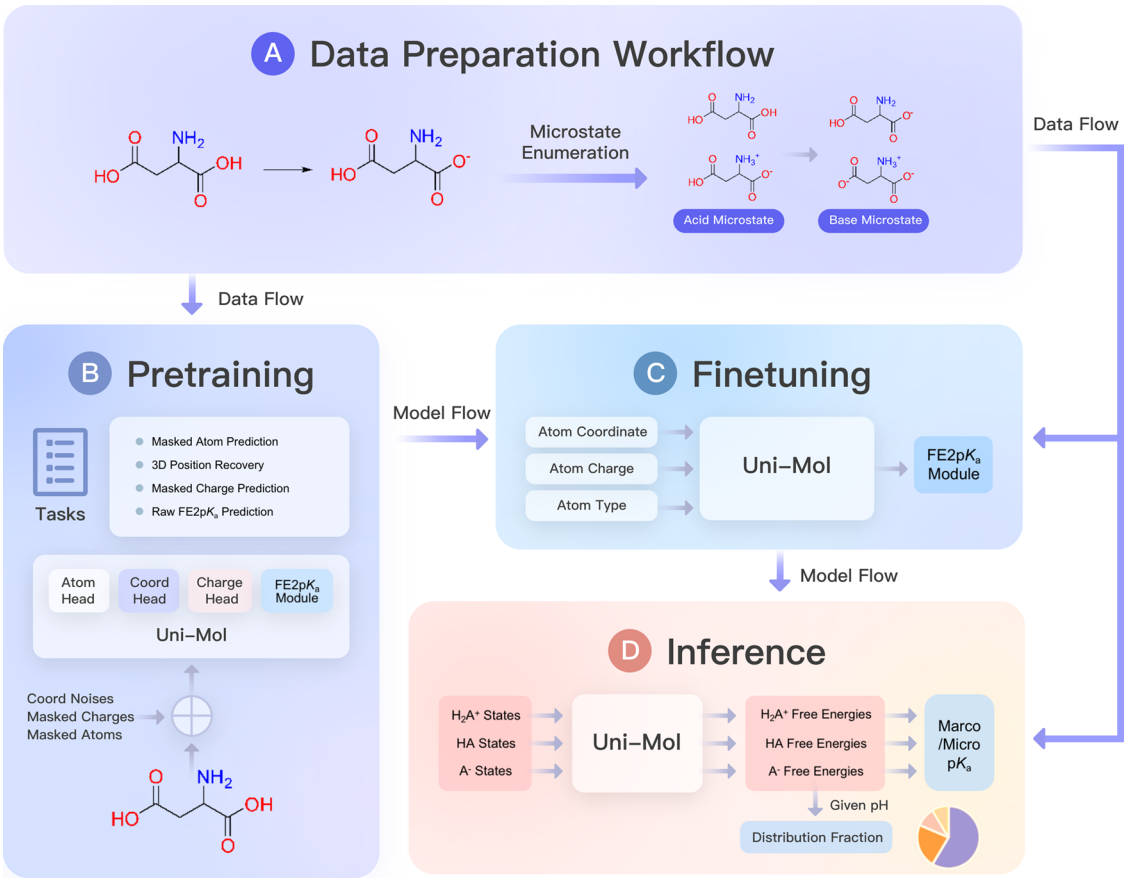
1. A **microstate** of a molecule is a well-defined chemical structure of its protonation state. A **micro**-pK_a describes the hidden, detailed equilibrium between two microstates, which NMR and kinetics methods can observe.
2. A **macrostate** of a molecule is a collection of its microstates with a particular net charge. A **macro**-pK_a describes the apparent, coarse-grained equilibrium between two macrostates, which electrochemical and spectroscopic methods mostly measure.

The microstate-macrostate hierarchy describes a molecule's **protonation ensemble**, which is our core concept to be revealed. It captures possible protonation states corresponding to different ionization site combinations for a given molecule (example of amoxicillin in Figure S1). We emphasize that **the free energies** (pH-dependent dimensionless Gibbs free energy change of formation at a certain pH, $\beta\Delta_f G_m(\cdot\ ;\ \mathrm{pH})$, see formulas eqs 1−4 in The free energy description of the protonation ensemble) **of all the microstates contain the complete information on coupled acid/base equilibria in the protonation ensemble** (see Section 5.1). This establishes the theoretical foundation of extracting the integrated equilibrium information faithfully from both micro- and macro-pK_a measurements by microstate free energy modeling, as well as predicting both pK_a values and pH-dependent protonation states by microstate free energy prediction.

Our Uni-pK_a framework integrates the theory of the protonation ensemble, a microstate enumerator, and molecular machine learning. The microstate enumerator implements the construction of the protonation ensemble of the target molecules, as explained in the Sections 5.2 and 5.3. The machine learning part serves as the core algorithm for the microstate free energy. Receiving molecular structures from the microstate enumerator and organized by the protonation ensemble, it converts molecular inputs to free energy outputs

https://doi.org/10.1021/jacsau.4c00271
JACS Au 2024, 4, 3451−3465

Scheme 1. Schematic Overview of Uni-pK_a Framework[a]



[a](A) Data preparation workflow. We implement a microstate enumerator to systematically build the protonation ensemble from a single structure. (B) 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. (C) 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. (D) 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.

along the data flow. Following a pretraining-finetuning paradigm, it learns from data with different fidelity and grows into a highly accurate free energy predictor along the model flow.

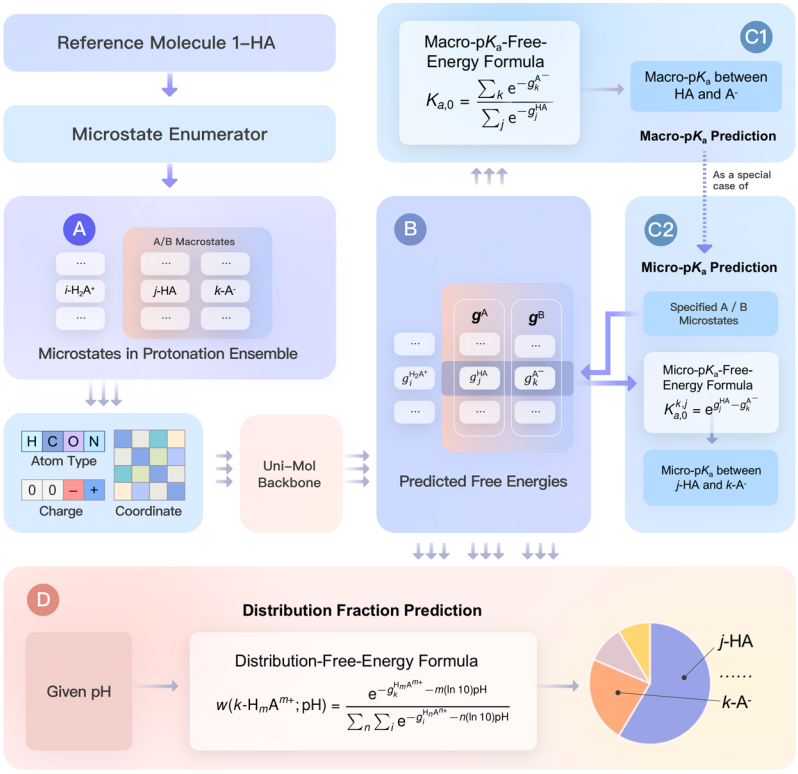
Scheme 1 provides a schematic overview of the Uni-pK_a framework. Uni-pK_a employs a unified data preparation procedure across the stages of pretraining, finetuning, and inference. Instead of directly inputting a single ionization reaction into the model, we recover the protonation ensemble on the data points to obtain microstates for the acid and base sides and feed them into the model. The model backbone originates from Uni-Mol,[34] an expressive and universal 3D molecular representation learning framework based on Transformer,[35] 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 free-energy-to-pK_a (FE2pK_a) module under the protonation ensemble theory.

The pretraining phase leverages about 1 million molecules with empirical pK_a values in the ChEMBL database,[36] which contains more than 3 million protonation states after microstate enumeration (Table S3). Four tasks are designed to exploit its abundant chemical information: 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 free energy consistency within the whole protonation ensemble by taking individual microstates as input and directly predicting microstate-free energies as the output. We employ the free-energy pK_a formulas (2) to predict the pK_a value for the entire data point and compute the loss with the ground truth, as explained in the Section 5.5.1.

After pretraining, we conduct finetuning with experimental pK_a labels, which endows our model to predict high-precision pK_a, as depicted in Scheme 1 C. The public pK_a data set in DataWarrior[37] and the selected entries from the i-BonD database[38] are sampled from a broad chemical space from the simple carboxylic acid to the complex alkaloids and porphyrins, covering the whole pH range in the aqueous solution (Figure S3) and spanning the 106 ionization patterns in our template. After our microstate enumeration, it is filled with well-constructed macrostates with at most 18 microstates (Figure S3).

https://doi.org/10.1021/jacsau.4c00271
JACS Au 2024, 4, 3451−3465

Scheme 2. Inference Stage of Uni-pK_a[a]



[a](A) Structures of microstates in the protonation ensemble of one reference molecule are reconstructed by the microstate generator. (B) The atom types, atomic charges, and geometry information of the microstates are fed into the Uni-Mol backbone, and the free energies are predicted for each microstate. (C) If the acid and base macrostates are specified by the user input, the macro-pK_a-free-energy formula is used to transform the free energy prediction to macro-pK_a prediction. If the microstates are further specified, the micro-pK_a-free-energy formula is used as a special case of the macro-pK_a prediction where there is only one microstate in both macrostates. (D) If pH is given by the user input, the distribution-free-energy formula is used to calculate the fraction of all the microstates in the protonation ensemble.

Table 1. Performance on External Data Sets

	Novartis											
	acid		base		total		SAMPL6		SAMPL7		SAMPL8	
method	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Schrödinger Epik Classical[17]	0.99	1.531	0.876	1.175	0.83	1.16	0.784	0.962	1.121	1.648		
ChemAxon Marvin[39]	**0.808**	1.144	0.835	1.145	0.86	1.17	1.007	1.248	0.559	0.708	1.300	1.511
ACD/Labs[40]							0.55	0.783				
SPOC + XGBoost[20]							0.767	1.011	1.476	1.622	*1.108*[a]	*1.547*
SPOC + NN[20]							0.832	1.141	0.932	1.156		
OPERA[18]							0.97	1.283	2.135	2.515		
MolGpKa[21]	0.849	1.287	0.789	1.064	0.87	1.27	**0.522**	0.773	0.797	0.98	*0.835*	*1.150*
GraphpKa[23]							0.594	0.726	0.758	0.934	*0.916*	*1.230*
pKasolver[22]					0.71	0.93					*1.244*	*1.590*
MF-SuP-pK_a[24]	0.85	1.09	0.61	0.79	0.71	0.92	0.687	0.751	0.656	0.816		
Schrödinger Epik v7[25]								0.92				
Uni-pK_a[b]	0.810	**1.061**	**0.493**	**0.653**	**0.620**	**0.840**	0.554	**0.716**	**0.570**	**0.735**	**0.631**	**0.878**

[a]The bold figures are the highest accuracy on each dataset. The italic figures present the performance of other methods tested by us on public platforms. Other external results are from the literature. The data and program sources are explained in Table S1. [b]Uni-pK_a with the formal charge descriptor is the default version. The performance of other versions is listed in Table S1.

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 (Scheme 2).

2.2. Model Accuracy and Generalizability

We evaluated Uni-pK_a's performance on external data sets spanning diverse chemical spaces to assess generalizability. Novartis and SAMPL6 collect a variety of possible protonation

https://doi.org/10.1021/jacsau.4c00271
JACS Au 2024, 4, 3451−3465

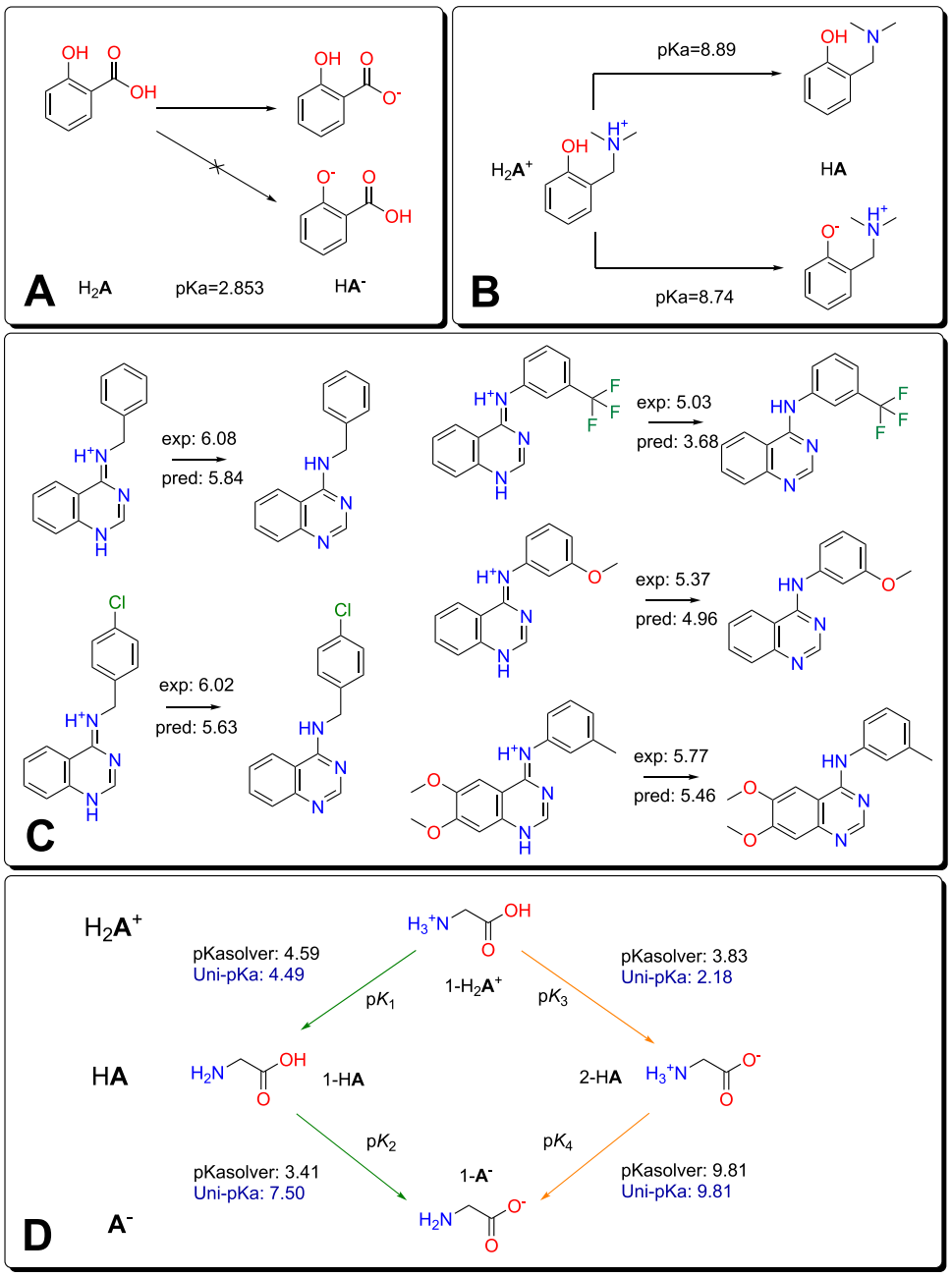


Figure 1. Uni-pK_a's concern for detailed acid–base equilibria. (A) Example of 2-hydroxybenzoic acid,[41] where one of the dissociation is dominant. (B) Example of 2-((dimethylamino)methyl)phenol,[42] where both reactions are dominant. (C) Uni-pK_a results on SAMPL6 micro-pK_a data sets involving tautomerism. (D) Thermodynamic cycle of the glycine. pK_i is the dissociation equilibrium constant. The green and orange arrows indicate different protonation routes.

patterns in complex drug-like molecules, which faithfully reflect the realities of a pK_a prediction model's application. SAMPL7 includes sulfonamides in different scaffolds and chemical environments, which tests the model's capacity to resolve subtle substituent effects. As summarized in Table 1, Uni-pK_a outperforms recent chemoinformatics methods on the Novartis, SAMPL6, and SAMPL7 data sets.

SAMPL8 is the most recent SAMPL challenge series that involves a pK_a prediction contest; therefore, it refreshes the benchmarks for previous relevant works. Uni-pK_a significantly exceeds the real submission entries in the SAMPL8 challenge (Table S2). As shown in Table 1, we further compared Uni-pK_a with several methods with their web platform and their deployed model, which reflects the best available pK_a prediction service to the public before. Uni-pK_a decreases the MAE by more than 0.2 pK_a units (0.631 for Uni-pK_a compared to 0.835 for MolGpKa) on this data set containing drug-like molecules with multiple ionization sites and successive ionization.

The macrostates in the training set of Uni-pK_a are abridged to reduce computational cost (Supporting Information: Data Sets), while the microstates in the external test sets come directly from the full enumeration without any handpick. The biggest risk of the direct tandem of the enumerator and the neural network is that the unusual structures generated by radical enumeration are unfamiliar to the neural network

https://doi.org/10.1021/jacsau.4c00271
JACS Au 2024, 4, 3451−3465

trained on the pruned data set. Therefore, the results above 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 with prior chemoinformatics techniques. The consistent improvements across heterogeneous evaluation sets validate the effectiveness of our protonation ensemble approach on a broad range of the chemical space.

2.3. Interpreting Macro-pK_a 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 data sets to quantify the additional information from the full enumeration. As shown in Table S3, mapping the public pK_a data sets from individual structures to the underlying ensembles expands the data substantially. For instance, the Dwar-iBonD data set grows over 3-fold from 8232 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. As shown in Figure 1A, the acidity of the carboxyl group is known to be much stronger than that of the phenolic hydroxyl group, leading to the obvious assignment. While for molecules with chemical groups of similar acidity as shown in Figure 1B, ambiguities often exist in attributing macro-pK_as to specific sites, and any assignment is an oversimplification and introduces bias to the data. Our protonation ensemble modeling reveals alternative chemically reasonable sites, including the dimethylamino group and the phenol group.

Recent works have adopted MIL to decompose macro-pK_as into contributions from specific sites.[23,24] However, MIL is a special case of the proton ensemble and risks misattributions for complex cases like Figure S1 (also eqs S1 and S2 in Supporting Information: Theoretical Details). Our iterative, ensemble-aware enumerator explores the full space, avoiding assumptions. Thoroughly sampling the ensemble is imperative in pursuit of rigorous macro-pK_a interpretation.

The accuracy of Uni-pK_a benefits from the data set built under the protonation ensemble framework. In Table S5, ablation studies show that full microstates in the Dwar-iBonD data set in the finetuning stage improve the RMSE in the cross-validation and on most external data sets. As we have emphasized, a correct interpretation of data is key to the progression of the model. Our reconstructed data sets show chemical soundness as well as help the model to grasp the chemical properties.

We further test Uni-pK_a on the SAMPL6 micro-pK_a data set, including 10 micro-pK_as of SAMPL6 challenge molecules measured by NMR titration. Uni-pK_a ends up with an MAE of 0.592 and an RMSE of 0.719, which is comparable to the results on the SAMPL6 macro-pK_a prediction in Table 1. It reveals that Uni-pK_a learns accurate knowledge of microlevel free energies from macrolevel equilibria, giving the right answer for the right reason. It is also worth noting that 5 of the 10 micro-pK_as involve tautomerism between enamines and imines (Figure 1C). Given that our model is trained on data sets that solely consider acid−base equilibrium, the test suggests that it has a generalizable perception of the underlying relationship between molecular structures and thermodynamics.

In conclusion, modeling the complete protonation ensemble provides a stronger foundation for leveraging experimental data, as fulfilled by our data set reconstruction. By preventing biased assumptions, it enables a more accurate pK_a prediction and the potential of extension into a general scope of various chemical equilibria.

2.4. Preserving Thermodynamic Consistency

pK_a itself is a thermodynamic property between the acid and base, constrained by fundamental thermodynamic cyclic relations. We illustrate this principle quantitatively in Figure 1D through an amino acid, glycine. The dominance of the zwitterion form of a neutral amino acid is a classical topic in biochemistry,[43] which can be derived from the micro-pK_as of the carboxyl group and the amino group. c(2-HA)/c(1-HA) = K_3/K_1 is based on the acidic ionization of 1-H_2A^+, while c(2-HA)/c(1-HA) = K_2/K_4 is based on the basic ionization of 1-A^-. The consistency $K_3/K_1 = K_2/K_4$ between the two results reported in both ways relies on the cyclic relation pK_1 + pK_2 = pK_3 + pK_4, which is deducted from the definition of the equilibrium constant. However, the relation is not guaranteed in arbitrary individual micro-pK_a predictions. Violation results in contradiction when evaluating the relative importance of different protonation configurations in different ways. Therefore, thermodynamic consistency not only promotes coherent micro-pK_a prediction but also avoids conflicts in the distribution coefficient prediction task derived from micro-pK_a prediction.

We exemplify the risk of methods without thermodynamic concern on the pKasolver, a graph-neural-network-based model that provides one of the most recent micro-pK_a prediction interfaces in its Colab notebook.[44] It shares the input format with Uni-pK_a that takes in both the acid and base when predicting micro-pK_a but with no thermodynamic constraints between individual predictions. We observe that pK_1 + pK_2 = 8.00 but pK_3 + pK_4 = 13.64 in its predicted values, which causes K_3/K_1 = 5.8, $K_2/K_4 = 2.5 \times 10^7$, more than 5 orders of magnitude's difference in the same quantity. As a result, the model fails to clarify this biochemical problem quantitatively.

By contrast, our free energy modeling of the protonation ensemble framework inherently preserves thermodynamic consistency between coupled pK_a values, under which the thermodynamic cycle is automatically satisfied (Supporting Information: The Verification of Thermodynamic Consistency). Hence, Uni-pK_a reports pK_1 + pK_2 = pK_3 + pK_4 = 11.99 and derives c(2-HA)/c(1-HA) = $K_3/K_1 = K_2/K_4 = 2.0 \times 10^2$ without any reference-dependent behavior.

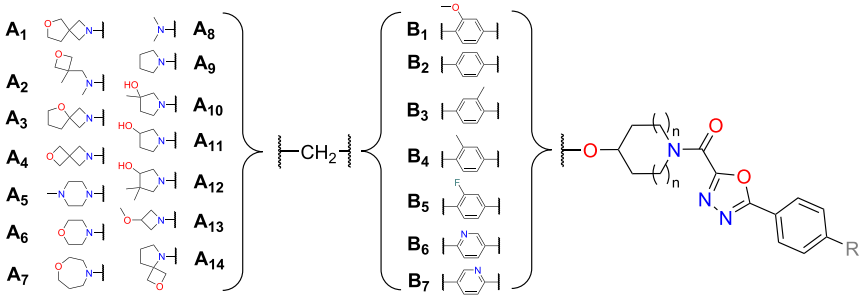
In general, with inherently encoded coupled acid−base equilibria, Uni-pK_a's thermodynamic awareness improves current micro-pK_a prediction schemes toward a self-consistent distribution coefficient prediction. This advance enables the holistic analysis of the protonation network with complex acid−base reactions.

2.5. Comparison to Quantum Chemistry Methods

The most accurate solutions for calculating thermodynamic properties, such as pK_a, are provided by quantum chemistry. Schrödinger's Jaguar, one of the state-of-the-art quantum-

https://doi.org/10.1021/jacsau.4c00271
JACS Au 2024, 4, 3451−3465

Table 2. Comparison between Uni-pK_a and Jaguar Results: Tertiary Amines



A	B	*n*	R	exp. value	Uni-pK_a	Jaguar	Jaguar+[a]	Jaguar++[b]
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 count[c]					4	6	5	**2**
best count[d]					**9**	6	4	8

[a]Jaguar with ordinary conformational search. [b]Jaguar with comprehensive conformational search, weighting 10 conformers. [c]A prediction with the error larger than 1 pK_a unit is regarded as an outlier. [d]The best among 4 methods each molecule is marked by bold figures.

chemistry-based pK_a prediction software, has reached experimental accuracy in a large chemical space.[45,46] This DFT-based prediction is very sensitive to the conformation energy because 1 kcal/mol corresponds to more than 0.7 pK_a unit. Thus, it heavily relies on a comprehensive conformational search and weighted average, with a proportionally increasing amount of computation. In practice, the trade-off between speed requirement and accuracy expectation determines the conformation search strategy.

Typically, compared to hours for conformational search and geometry optimization in an implicit solvent model of a typical-size molecule in Jaguar, Uni-pK_a's inference speed is 28 macro-pK_a per second, and the average prediction time for a macro-pK_a is approximately 0.036 s for the Novartis Acid data set. With this significant speed up against precise quantum chemistry calculations, we conducted comparative case studies with Jaguar to evaluate Uni-pK_a's accuracy loss.

Given practical computational constraints, Uni-pK_a demonstrates promising accuracy relative to that of Jaguar. For example, without conformational sampling, Uni-pK_a matches or exceeds Jaguar's accuracy on a family of drug-like molecules in Table 2. This highlights the benefits of data-driven training on large data sets. However, accuracy challenges remain for certain complex systems, where Jaguar's accuracy improves significantly with exhaustive conformational modeling (from MAE = 1.07 down to 0.20 in Table S6). While Uni-pK_a cannot match this (MAE = 0.70), it provides a much faster alternative within reasonable tolerances for many applications.

While Jaguar's DFT calculations provide *ab initio* pK_a estimates, systematic errors remain. To compensate, Jaguar employs an empirical "shell model" that assigns molecules to classes with parametrized corrections. However, this classification contains some arbitrariness, as the original authors note when evaluating guanidine derivatives (Table S7). By

https://doi.org/10.1021/jacsau.4c00271
JACS Au 2024, 4, 3451−3465

default, these molecules fall under the guanidine shell, giving an MAE of 1.34, even with exhaustive conformational sampling. Yet the "partially substituted amidine" shell yields superior accuracy, with an MAE of just 0.38 pK_a units without conformational sampling. The authors 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, giving an MAE of 0.64.

In conclusion, these benchmarks reveal a complementary synergy between the computational expense of quantum chemistry methods and the data efficiency of machine learning techniques like Uni-pK_a. Comparisons to in-depth quantum chemistry calculations substantiate Uni-pK_a's viability as an efficient surrogate for pK_a prediction, within limitations. Integrating the two approaches to balance speed and accuracy is an exciting direction for future hybrid modeling. Targeted integration of first-principles training data could help address areas for improvement revealed by quantum chemistry benchmarks. This further motivates the development of unified ensemble modeling frameworks.

3. DISCUSSION

Uni-pK_a comprises two key components under protonation ensemble theory: the microstate enumerator and the neural network predictor. The microstate enumerator systematically generates the protonation states of molecules, representing the complete collection of molecular protonation equilibria. The neural network predictor, trained on well-designed tasks and well-constructed data sets, ensures precise and reliable predictions.

The microstate enumerator plays a key role in the applicability of Uni-pK_a in a broad chemical space. A more thorough ionization pattern template undoubtedly leads to wider and deeper enumeration and 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 neural network computation. Nonetheless, the system size is controlled in the realm of small organic molecules. Because the batch-wise inference holds dozens of structures in parallel, even extremely complex molecules like porphyrins and small peptides' macrostates can be evaluated concurrently at an averaged time. In a manually built, ethylenediaminetetraacetic acid (EDTA)-like complex molecule example with at most 6 ionization sites, the macro-pK_a prediction between central macrostates regarding the most microstates takes a nearly constant time across different sizes of macrostates (Figure S5). Moreover, it is reasonable 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. Although reasonable pruning rules in the vast chemical space are case by case, our SMARTS-based template is well-documented with interpretable chemical names for specific demands in different chemical domains. We choose the Dwar-iBonD data set 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.

Among the models being compared against, the three technical routes summarized in the Introduction aiming to fast, general, data-driven pK_a prediction methods for small organic molecules in the aqueous solution are all covered. Epik Classical,[17] ChemAxon,[39] and ACD/Labs[40] represent popular choices of commercial software for molecular property prediction, which share the template matching and empirical correction methods. From OPERA[18] to SPOC,[20] molecular descriptors are refined and customized for pK_a prediction with traditional machine learning. The evolution of deep learning methods since MolGpKa,[21] reflects the deployment of more powerful graph neural networks that capture the long-range effects,[23,24] the benefits from large pretraining data set[22] and quantum chemistry data set,[25] and the progressive consideration of complex acid−base equilibria.[23−25,47] Our Uni-pK_a belongs to the deep learning class as well. It adopts the transformer architecture on 3D molecular structures instead of GNNs on molecular graphs, which leverages the advantage of long-range modeling of attentive modules and embraces cutting-edge molecular representation learning techniques. It follows the successful practice of multifidelity learning with a pretraining-finetuning strategy but completes the data with our microstate enumerator and augments the tasks with self-supervised learning inspired by the original Uni-Mol work.[34] It gets out of the box of the molecule-site representation of protonation reactions and complements the partial concerns of acid−base equilibria in a unified, complete framework, namely, the free energy modeling of the protonation ensemble. It is those advances that enable Uni-pK_a to explore the boundary of data-driven pK_a prediction under the existing data, toward the pK_a measurement's uncertainty limit of 0.5 pK_a unit.[45]

The use of an atomic charge is one of the adaptive modifications of the general-purpose Uni-Mol model in the pK_a prediction task. The atomic charge describes the local electrostatic environment around the ionization sites and reflects the local electronic structure polarization. Thus, it is a common choice for pK_a-related tasks in the feature sets of traditional statistical learning methods,[18−20,48−52] a part of molecular embeddings in deep learning methods,[22−25] and the semiempirical correction of quantum chemistry in solvent models.[46,53,54] In addition to the formal charge, the performance of Gasteiger's empirical charge scheme[55] and GFN2-xTB partial charge[56] is also explored during the whole training and exploration process. We choose the formal charge as default because it is directly read from the SMILES input without any further time-consuming calculation and achieves the best accuracy on the most diverse Novartis, SAMPL6, and SAMPL8 data sets (Table S5) among the three charge schemes. However, other schemes are still able to outperform other methods with the Uni-pK_a framework (Table S1). Continuous charge schemes inform more than the discrete and sparse formal charge when the chemical environment varies subtly. In SAMPL7, a more homogeneous data set filled with sulfonamides, the GFN2-xTB partial charge on the dominant ionization site has a strong correlation with the pK_a value (Figure S6), which explains why Uni-pK_a behaves best with the GFN2-xTB partial charge on this data set (Table S5). Therefore, we believe that refined descriptors, not limited to atomic charge schemes, should help Uni-pK_a gain more precision in specific domains of the chemical space. Uni-pK_a

https://doi.org/10.1021/jacsau.4c00271
JACS Au 2024, 4, 3451−3465

as a general framework always accommodates different descriptors in its neural network at the molecular and atomic embedding level and exploits their potential by the comprehensive free energy modeling of the protonation ensemble.

However, accuracy challenges persist for certain complexes with subtle stereoelectronic effects like the proton sponge[57,58] and Meldrum's acid,[59−62] not well represented in training data. Tautomerism has not been explicitly addressed in our framework. Although Figure 1C gives a positive signal of our model's robustness on tautomerism, limited accuracy is obtained in additional tests on specialized tautomer data sets, like Tautobase.[63] An MAE of 3.435 for log K of the transition between the tautomer pairs in the aqueous solution was near room temperature. Observation in Figure S4 shows the model tends to attach similar free energies to different tautomers and fails to distinguish the tautomers with drastic energy differences. The main difficulty in modeling various tautomerism exhaustively is due to the scarcity of experimental data. Thus, future work will focus on integrating more first-principles training data and optimizing the enumeration process to address these challenges.

4. CONCLUSIONS

In this work, we present Uni-pK_a, a novel framework that integrates thermodynamic principles with advanced machine learning techniques to predict the acid dissociation constants (pK_a) of organic molecules. The significance of this work lies in its introduction of physical knowledge into machine learning modeling, which establishes a reliable framework for the pK_a prediction.

A highlight of Uni-pK_a is its innovative approach to unifying micro- and macro-pK_a values using thermodynamic free energy relationships. This method ensures thermodynamic consistency across different protonation states, addressing a common limitation in traditional models. By modeling the free energies of all microstates through well-defined thermodynamic equations, Uni-pK_a provides a comprehensive and coherent depiction of acid−base equilibria. This approach not only maintains consistency but also extends to the distribution coefficient prediction across various pH values.

Uni-pK_a leverages Uni-Mol as the molecular encoder, a transformer-based architecture designed for molecular representation. Its invariant spatial positional encoding allows the model to distinguish the 3D spatial positions of atoms, while the pair representation captures the spatial relationships between atom pairs. Uni-pK_a follows a pretraining-finetuning paradigm. During pretraining, the model learns high-level chemistry from large-scale, low-fidelity, computationally predicted pK_a data. This process includes one weakly supervised task of predicting pK_a values from ChEMBL data and three self-supervised tasks: masked atom prediction, masked charge prediction, and 3D position recovery. Subsequently, Uni-pK_a is fine-tuned on high-fidelity experimental pK_a data for enhanced accuracy.

Our extensive evaluation demonstrates the state-of-the-art accuracy of Uni-pK_a. On the Novartis, SAMPL6, and SAMPL7 benchmarks, Uni-pK_a achieves mean absolute errors (MAEs) of 0.810, 0.493, and 0.554, respectively. Moreover, in the blind SAMPL8 challenge, Uni-pK_a significantly outperforms other competitors with an MAE of 0.619. These results qualify Uni-pK_a as a competitive tool in pK_a prediction. Case studies further consolidate the model's accuracy and robustness. For instance, in predicting the pK_a of tertiary amines, Uni-pK_a exhibits an MAE of 0.52, outperforming traditional quantum chemistry methods such as Jaguar, which has an MAE of 0.69. Additionally, Uni-pK_a demonstrates its capability in handling complex cases involving multiple ionizable groups and tautomerism, as evidenced by its performance on the SAMPL6 micro-pK_a data set.

Integrating data-driven techniques such as our framework with first-principles training is an exciting path forward. With free energy as a pivot, our protonation ensemble approach establishes a strong foundation to accommodate all kinds of interconnected equilibria for future synergistic hybrid modeling.

5. METHOD

5.1. Free Energy Description of the Protonation Ensemble

The simplest acidic ionization reaction in water is

$$HA^+(aq) \rightleftharpoons H^+(aq) + A(aq),\ K_a^{\ominus} = \frac{[H^+][A]}{[HA^+]}$$

where $[\cdot]$ is a chemical species's activity (or dimensionless concentration, approximately). The equilibrium produces various chemical structures of the same initial molecule, namely, its protonation states in the solution. pK_a is the negative logarithm (base 10) of the acid dissociation equilibrium constant K_a.

The micro-pK_a-macro-pK_a hierarchy is further clarified from a free-energy perspective. If we denote the k-th microstate in the microstate of a neutral structural core A with m protons as k-H_mA^{m+}. Micro-pK_a values arise from the standard free energy change $\Delta_f G_m^{\ominus}$ between a defined microstate pair k-H_mA^{m+} and i-$H_{m+1}A^{(m+1)+}$. Let R be the molar gas constant, and $\beta = (RT)^{-1}$. The micro-pK_a is

$$\begin{aligned} K_{a,m}^{k,i} &:= \frac{[k - H_mA^{m+}][H^+]}{\left[i - H_{m+1}A^{(m+1)+}\right]} \\ &= \exp\left\{-\beta\left[\Delta_f G_m^{\ominus}\left(k - H_mA^{m+}\right) - \Delta_f G_m^{\ominus}\left(i - H_{m+1}A^{(m+1)+}\right)\right]\right\} \end{aligned} \quad (1)$$

Macro-pK_a values originate from the collective contribution of all microstates in adjacent macrostates, derived as

$$\begin{aligned} K_{a,m} &:= \frac{[H^+]\sum_i [i - H_mA^{m+}]}{\sum_i \left[i - H_{m+1}A^{(m+1)+}\right]} \\ &= \frac{\sum_i \exp(-\beta\Delta_f G_m^{\ominus}(i - H_mA^{m+}))}{\sum_i \exp\left(-\beta\Delta_f G_m^{\ominus}\left(i - H_{m+1}A^{(m+1)+}\right)\right)} \end{aligned} \quad (2)$$

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 $\Delta_f G_m^{\ominus}$.

We can define a pH-dependent free energy for the general case of a multiprotonated acid,

$$\Delta_f G_m(H_mA^{m+}; \mathrm{pH}) = \Delta_f G_m^{\ominus}(H_mA^{m+}(aq)) + \frac{m\ln 10}{\beta}\mathrm{pH} \quad (3)$$

https://doi.org/10.1021/jacsau.4c00271
JACS Au 2024, 4, 3451−3465

where $\Delta_f G_m^{\ominus}(\cdot)$ is the standard molar Gibbs free energy change of formation at the temperature T we study. The motivation of this definition is shown in Supporting Information: Theoretical Details. pH-dependent free energies give the fraction of each microstate across the ensemble under particular pH conditions:

$$w(k - \mathrm{H}_m\mathrm{A}^{m+}; \mathrm{pH}) := \frac{[k - \mathrm{H}_m\mathrm{A}^{m+}]}{\sum_n \sum_i [i - \mathrm{H}_n\mathrm{A}^{n+}]} = \frac{\exp(-\beta \Delta_f G_m(k - \mathrm{H}_m\mathrm{A}^{m+}; \mathrm{pH}))}{\sum_n \sum_i \exp(-\beta \Delta_f G_m(i - \mathrm{H}_n\mathrm{A}^{n+}; \mathrm{pH}))} \quad (4)$$

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 on the acid/base equilibrium.

5.2. 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 the macrostate $\mathrm{H}_m\mathrm{A}^{m+}$ to generate all microstates in $\mathrm{H}_m\mathrm{A}^{m+}$ and a neighboring macrostate $\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+}$ or $\mathrm{H}_{m-1}\mathrm{A}^{(m-1)+}$.

The enumerator uses a template containing SMARTS patterns of ionizable sites. It is modified, augmented, and annotated based on the template in MolGpKa[21] with chemical consideration. It contains 53 common acidic and basic groups with separate entries for deprotonation and protonation (examples in Table S8) and covers all the ionization patterns demonstrated by the Dwar-iBonD data set introduced in Reconstructed data sets.

When the enumeration starts, A and B Micropools 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 S2). The algorithm then iteratively grows the pools:

1. **A to B (A2B) round: deprotonation.** For each structure in A Micropool, substructure matching finds all possible deprotonation sites in the template, and corresponding deprotonated structures go into B Micropool.
2. **B to A (B2A) round: protonation.** For each structure in B Micropool, substructure matching finds all possible protonation sites in the template, and corresponding protonated structures go into A Micropool.

Therefore, beginning with some $\mathrm{H}_m\mathrm{A}^{m+}$, if the macrostate $\mathrm{H}_{m-1}\mathrm{A}^{(m-1)+}$ is needed, the initial structures will be thrown into the A Micropool with the B Micropool empty, and an A2B round will go first (**A**cid mode). $\mathrm{H}_{m+1}\mathrm{A}^{(m+1)+}$ 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, and then A and B Micropools are output as the final enumeration results. The maximum 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 S2), like the coexistence of acidic ionization of the amino group and basic ionization of the amino group. These structures can be pruned because of their small contribution to the protonation ensemble.

The whole protonation ensemble is obtained by successively running the enumeration process above in A and B modes. In the case of Figure S1, the whole macrostate of $\mathrm{H_2A}$ and can be enumerated from 1-$\mathrm{H_2A}$ in the A mode, $\mathrm{H_3A^+}$ comes from $\mathrm{H_2A}$ in the B mode, and $\mathrm{A^{2-}}$ steps further from $\mathrm{HA^-}$ 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 $\mathrm{H_2A}$ and $\mathrm{HA^-}$ in Figure S1 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 $\mathrm{H_2A}$ and $\mathrm{HA^-}$, increasing the width of the enumeration results. Furthermore, when the amide group in 1-$\mathrm{A^{2-}}$ is deprotonated, the macrostates extend to $\mathrm{H_{-1}A^{3-}}$, increasing the depth of the enumeration results.

5.3. Reconstructed Data Sets

In existing public pK_a data sets, like DataWarrior pKaInWater[37] and iBonD,[38] 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. Benefiting from the advancement of the protonation ensemble framework, we reconstructed several data sets 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 such as 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 data sets into a unified SMILES-like structure that stores the enumerated microstates mapped to each published macro-pK_a measurement.

Our release covers 7 experimental and predicted data sets relevant to drug-like chemical space from ChEMBL,[24,36] DataWarrior,[37] iBonD,[38] Novartis,[24,64] SAMPL6,[65] SAMPL7,[66] and SAMPL8[67] (Table S3), including:

1. Small molecule compilations like SAMPL - with exhaustive microstate enumeration
2. Large predicted set from ChEMBL - with one-iteration enumeration

This work integrates robust chemical knowledge about protonation mechanisms with consistent experimental measurements into high-quality data sets tailored for developing accurate and physically consistent machine learning models. Full details of the source data and reconstruction process are provided in Supporting Information: Data Sets.

5.4. Backbone and Model Input

We chose Uni-Mol as the encoder for individual molecules. Uni-Mol is a standard transformer[35] based on Pre-Layer Normalization[68] with several modifications.

Uni-Mol incorporates invariant spatial positional encoding. Vanilla Transformer cannot distinguish the positions of inputs

https://doi.org/10.1021/jacsau.4c00271
JACS Au 2024, 4, 3451−3465

without positional encoding because the model architecture is permutation invariant. Uni-Mol uses Euclidean distances of all-atom pairs, plus with the edge type aware Gaussian kernels,[69] as the spatial positional encoding. This encoding is invariant with global rotation and translation. Formally, the D channel positional encoding of atom pair ij is denoted as

$$\begin{aligned} \boldsymbol{p}_{ij} &= \{\mathcal{G}(\mathcal{A}(d_{ij}, t_{ij}; \boldsymbol{a}, \boldsymbol{b}), \mu^k, \sigma^k) | k \in [1, D]\} \\ &, \ \mathcal{A}(d, r; \boldsymbol{a}, \boldsymbol{b}) \\ &= a_r d + b_r \end{aligned} \tag{5}$$

where $\mathcal{G}(\cdot, \mu^k, \sigma^k)$ is a Gaussian density function with parameters μ^k and σ^k, d_{ij} is the Euclidean distance of atom pair ij, and t_{ij} is the edge type of atom pair ij. The edge here is not the chemical bond, and the edge type is determined by the atom types of pair ij. $\mathcal{A}(\cdot, \cdot; \boldsymbol{a}, \boldsymbol{b})$ is the affine transformation with parameters $\boldsymbol{a}$ and $\boldsymbol{b}$, it affines d_{ij} corresponding to its edge type.

Transformer typically maintains the token(atom) level representation. In Uni-Mol, spatial positions are encoded at the pair level; therefore, pair-level representations are maintained as well. Spatial information is propagated through atom-to-pair and pair-to-atom communications to better understand spatial representations. Specifically, the pair representation is initialized by invariant spatial positional encoding and updated using atom-to-pair communication through the multihead Query-Key resulting in self-attention. Formally, the update of ij pair representation is denoted as

$$\boldsymbol{q}_{ij}^0 = \boldsymbol{p}_{ij} M, \ \boldsymbol{q}_{ij}^{l+1} = \boldsymbol{q}_{ij}^l + \left\{ \frac{Q_i^{l,h}(K_j^{l,h})^T}{\sqrt{d}} \middle| h \in [1, H] \right\} \tag{6}$$

where H is the number of attention heads, d is the dimension of hidden representations, $Q_i^{l,h}$ is the Query of the i-th atom in the l-th layer h-th head, $K_j^{l,h}$ is the Key of the j-th atom in the l-th layer h-th head, $M \in \mathbb{R}^{D \times H}$ is the projection matrix to make the representation the same shape as multihead Query-Key product results. Besides, pair-to-atom communication, using the pair representation as the bias term in self-attention, helps leverage spatial information in the atom representation. Formally, the self-attention with pair-to-atom communication is denoted as

$$\begin{aligned} &\text{Attention}(Q_i^{l,h}, K_j^{l,h}, V_j^{l,h}) \\ &= \text{softmax}\left(\frac{Q_i^{l,h}(K_j^{l,h})^T}{\sqrt{d}} + \boldsymbol{q}_{ij}^{l-1,h} \right) V_j^{l,h} \end{aligned} \tag{7}$$

where $V_j^{l,h}$ is the Value of the j-th atom in the l-th layer h-th head.

A simple SE(3)-equivariance head is also added to Uni-Mol to enable the model to directly output coordinates. This plays a role in the 3D position recovery task in pretraining. The design of the SE(3)-equivariance head is denoted as

$$\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i + \sum_{j=1}^{n} \frac{(\boldsymbol{x}_i - \boldsymbol{x}_j) e_{ij}}{n}, \ e_{ij} = \text{ReLU}((\boldsymbol{q}_{ij}^L - \boldsymbol{q}_{ij}^0)\boldsymbol{U})\boldsymbol{W} \tag{8}$$

where n is the number of total atoms, $\boldsymbol{x}_i \in \mathbb{R}^3$ is the input coordinate of the i-th atom, and $\hat{\boldsymbol{x}}_i \in \mathbb{R}^3$ is the output coordinate of the i-th atom, $\boldsymbol{U} \in \mathbb{R}^{H \times H}$ and $\boldsymbol{W} \in \mathbb{R}^{H \times 1}$ are the projection matrices to convert pair representation to a scalar.

For model input, unlike standard Uni-Mol, Uni-pK_a takes three inputs. Along with atom types and coordinates, we also consider the influence of atom charges, as they are 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. We consider states with formal charge values of 0, 1, and −1 for discrete charges, 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 multilayer perceptron (MLP) to obtain their initial representations.

5.5. Pretraining

Inspired by the success of large language models in natural language processing[70−73] and computer vision,[74] 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. 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 data set as "low fidelity data" for a weakly supervised pretraining of pK_a models.[21,22,24] 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.

5.5.1. Weakly Supervised Task: pK_a-Prediction. First, 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,[70] we introduce a special atom called `[CLS]`. Its coordinates represent the center of the molecule. We used 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. Enabled by the proton ensemble theory, this output will be interpreted as the predicted $\beta\Delta_f G_m^{\ominus}$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^A, g_2^A, ⋯ for the microstates in A macrostate and g_1^B, g_2^B, ⋯ for the microstates in B macrostate, then the final loss function of a single data point is a combination of mean square error loss and the macro-pK_a-free-energy formula 2:

$$\mathcal{L}_{\text{pK}_a}(\boldsymbol{g}^A, \boldsymbol{g}^B; \text{pK}_a) = \frac{1}{2}\left[\text{pK}_a + \log_{10} \frac{\sum_i e^{-g_i^B}}{\sum_i e^{-g_i^A}} \right]^2 \tag{9}$$

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:

https://doi.org/10.1021/jacsau.4c00271
JACS Au 2024, 4, 3451−3465

1. For each data point, Uni-Mol in Uni-pK_a outputs free energy vectors g^A and g^B for the microstates in macrostates A and B.
2. These are used to compute the total Boltzmann-weighted partition functions of A and B microstates based on eq 2.
3. The loss function (eq 9) compares this logarithmic partition-function ratio to the reported macro-pK_a through a mean squared error term.
4. 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 such as scaling would break the physical meaning of the outputs. However, translation by a value of t maintains interpretation as pH-dependent free energies, $\beta\Delta_f G_m(\cdot\ ;\ \mathrm{pH} = t)$, as proved in Supporting Information: Theoretical Details.

5.5.2. Self-Supervised Tasks. Additionally, we introduced 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 `[MASK]` token prediction by replacing masked atom types with a `[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 $\mathcal{L}_{\mathrm{atom}}$:

$$\mathcal{L}_{\mathrm{atom}} = -\sum_{i\in\mathcal{M}} \log P(y_i|x_i) \tag{10}$$

where $\mathcal{M}$ denotes the set of masked atoms, y_i is the true type of atom i, and x_i is the input containing the `[MASK]` token.

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 prediction of 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 `[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 $\mathcal{L}_{\mathrm{charge}}$:

$$\mathcal{L}_{\mathrm{charge}} = -\sum_{i\in\mathcal{M}} \log P(c_i|x_i) \tag{11}$$

where $\mathcal{M}$ denotes the set of masked atoms, c_i is the true charge of atom I, and x_i is the input containing the `[MASK]` token. 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. For continuous charges, since they cannot be directly masked like discrete charges, we simply add their representations to the atom representations.

Furthermore, we aim for the model to learn 3D structural information within molecules. Therefore, we retained the 3D position 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 the 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 $\mathcal{L}_{\mathrm{dist}}$.

$$\mathcal{L}_{\mathrm{coord}} = \sum_{i\in\mathcal{M}} \mathrm{smooth}_l(\hat{\mathbf{r}}_i - \mathbf{r}_i) \tag{12}$$

$$\mathcal{L}_{\mathrm{dist}} = \sum_{i\in\mathcal{M}} \sum_{j} \mathrm{smooth}_l(\hat{d}_{i,j} - d_{i,j}) \tag{13}$$

where $\mathcal{M}$ denotes the set of masked atoms, $\hat{\mathbf{r}}_i$ and $\mathbf{r}_i$ are the predicted and true coordinates of atom i, respectively, and $\hat{d}_{i,j}$ and $d_{i,j}$ are the predicted and true pairwise distances between masked atom i and all other atoms j, respectively. These two losses also constitute part of the original Uni-Mol loss. The smooth $\mathcal{L}_l$ loss is used to ensure stability during training, and its formula is as follows:

$$\mathrm{smooth}_{l_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{14}$$

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

$$\mathcal{L}_{\mathrm{all}} = \mathcal{L}_{\mathrm{p}K_a} + 2\mathcal{L}_{\mathrm{charge}} + \mathcal{L}_{\mathrm{atom}} + 2\mathcal{L}_{\mathrm{coord}} + \mathcal{L}_{\mathrm{dist}} \tag{15}$$

5.6. 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 fine-tuned on the Dwar-iBonD data set using the loss function (9).

For aiding model convergence, the pK_a target is translated by the average of the data set 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.

■ ASSOCIATED CONTENT

Data Availability Statement

Relevant data sets can be obtained from https://www.aisquare.com/datasets/. The code is open source at https://

https://doi.org/10.1021/jacsau.4c00271
JACS Au 2024, 4, 3451−3465

github.com/dptech-corp/Uni-pKa. Users can utilize Uni-pK_a for predicting and ranking the protonation states of molecules under various pH conditions via https://bohrium.dp.tech/apps/uni-pka.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacsau.4c00271.

Theoretical details of the protonation ensemble framework; origins, reconstruction methods, organization, and statistics of the data sets; technical details of molecular preprocessing and model training; and additional test results (PDF)

AUTHOR INFORMATION

Corresponding Authors

Zhifeng Gao − *DP Technology, Beijing 100089, China*; Email: gaozf@dp.tech

Hang Zheng − *DP Technology, Beijing 100089, China*; Email: zhengh@dp.tech

Authors

Weiliang Luo − *Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; DP Technology, Beijing 100089, China*; orcid.org/0009-0005-6150-2797

Gengmo Zhou − *DP Technology, Beijing 100089, China; Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China*

Zhengdan Zhu − *DP Technology, Beijing 100089, China*; orcid.org/0000-0001-9260-6226

Yannan Yuan − *DP Technology, Beijing 100089, China*

Guolin Ke − *DP Technology, Beijing 100089, China*

Zhewei Wei − *Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.4c00271

Author Contributions

‖W.L. and G.Z. contributed equally to this work. CRediT: **Weiliang Luo** conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing-original draft; **Gengmo Zhou** data curation, formal analysis, investigation, methodology, software, validation, visualization, writing-original draft; **Zhengdan Zhu** conceptualization, resources, validation, writing-review & editing; **Yannan Yuan** software, validation, visualization; **Guolin Ke** funding acquisition, project administration, resources, supervision; **Zhewei Wei** funding acquisition, project administration, resources, supervision; **Zhifeng Gao** methodology, resources, validation, writing-original draft, writing-review & editing; **Hang Zheng** conceptualization, funding acquisition, project administration, resources, supervision, validation, writing-original draft, writing-review & editing.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Fengmei Chen for retouching the illustrations and Haihui Lan for providing a critical reading of the manuscript.

REFERENCES

(1) Wang, H.; Tianfan, F.; Yuanqi, D.; Gao, W.; Huang, K.; Liu, Z.; Chandak, P.; Liu, S.; Van Katwyk, P.; Deac, A.; et al. Scientific discovery in the age of artificial intelligence. *Nature* **2023**, *620* (7972), 47−60.

(2) Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Stefan Bringuier, L.; Brinson, C.; Choudhary, K.; Circi, D.; et al. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery* **2023**, *2* (5), 1233−1250.

(3) Rodrigues, T. The good, the bad, and the ugly in chemical and biological data for machine learning. *Drug Discovery Today: Technol.* **2019**, *32*, 3−8.

(4) Nandy, A.; Duan, C.; Kulik, H. J. Audacity of huge: overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery. *Curr. Opin. Chem. Eng.* **2022**, *36*, No. 100778.

(5) Frey, N. C.; Soklaski, R.; Axelrod, S.; Samsi, S.; Gomez-Bombarelli, R.; Coley, C. W.; Gadepally, V. Neural scaling of deep chemical models. *Nat. Mach. Intell.* **2023**, *5* (11), 1297−1305.

(6) Karniadakis, G. E.; Kevrekidis, I. G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3* (6), 422−440.

(7) Zhang, X.; Wang, L.; Helwig, J.; Luo, Y.; Fu, C.; Xie, Y.; Liu, M.; Lin, Y.; Xu, Z.; Yan, K. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423* **2023**.

(8) Hasan, M. H.; McCrum, I. pKa as a predictive descriptor for electrochemical anion adsorption. *Angew. Chem. Int. Ed.* **2024**, *63*, No. e202313580.

(9) Yang, J.-D.; Xue, J.; Cheng, J.-P. Understanding the role of thermodynamics in catalytic imine reductions. *Chem. Soc. Rev.* **2019**, *48* (11), 2913−2926.

(10) Craig, M. J.; Garcia-Melchor, M. High-throughput screening and rational design to drive discovery in molecular water oxidation catalysis. *Cell Rep. Phys. Sci.* **2021**, *2* (7), No. 100492.

(11) Chang, E. D.; Town, R. M.; Owen, S. F.; Hogstrand, C.; Bury, N. R. Effect of water pH on the uptake of acidic (ibuprofen) and basic (propranolol) drugs in a fish gill cell culture model. *Environ. Sci. Technol.* **2021**, *55* (10), 6848−6856.

(12) Manallack, D. T.; Prankerd, R. J.; Yuriev, E.; Oprea, T. I.; Chalmers, D. K. The significance of acid/base properties in drug discovery. *Chem. Soc. Rev.* **2013**, *42* (2), 485−496.

(13) Chen, W.; Deng, Y.; Russell, E.; Yujie, W.; Abel, R.; Wang, L. Accurate calculation of relative binding free energies between ligands with different net charges. *J. Chem. Theory Comput.* **2018**, *14* (12), 6346−6358.

(14) 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.* **2018**, *15* (1), 424−435.

(15) Jialu, W.; Kang, Y.; Pan, P.; Hou, T. Machine learning methods for pK_a prediction of small molecules: Advances and challenges. *Drug Discovery Today* **2022**, *27*, No. 103372.

(16) Hammett, L. P. Linear free energy relationships in rate and equilibrium phenomena. *Trans. Faraday Soc.* **1938**, *34*, 156−165.

(17) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: a software program for pK_a prediction and protonation state generation for drug-like molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681−691.

(18) Mansouri, K.; Cariello, N. F.; Korotcov, A.; Tkachenko, V.; Grulke, C. M.; Sprankle, C. S.; Allen, D.; Casey, W. M.; Kleinstreuer, N. C.; Williams, A. J. Open-source QSAR models for pKa prediction using multiple machine learning approaches. *J. Cheminform.* **2019**, *11* (1), 60.

(19) Baltruschat, M.; Czodrowski, P. Machine learning meets pK_a. *F1000Res.* **2020**, *9*, 113.

(20) Yang, Q.; Li, Y.; Yang, J.-D.; Liu, Y.; Zhang, L.; Luo, S.; Cheng, J.-P. Holistic prediction of the pK_a in diverse solvents based on a

machine-learning approach. *Angew Chem. Int. Ed.* **2020**, *59* (43), 19282−19291.
(21) Pan, X.; Wang, H.; Li, C.; Zhang, J. Z. H.; Ji, C. MolGpka: A web server for small molecule pK_a prediction using a graph-convolutional neural network. *J. Chem. Inf. Model.* **2021**, *61* (7), 3159−3165.
(22) Mayr, F.; Wieder, M.; Wieder, O.; Langer, T. Improving small molecule pK_a prediction using transfer learning with graph neural networks. *Front. Chem.* **2022**, *10*, No. 866585.
(23) Xiong, J.; Li, Z.; Wang, G.; Zunyun, F.; Zhong, F.; Tingyang, X.; Liu, X.; Huang, Z.; Liu, X.; Chen, K.; et al. Multi-instance learning of graph neural networks for aqueous pK_a prediction. *Bioinformatics* **2022**, *38* (3), 792−798.
(24) Jialu, W.; Wan, Y.; Zhenxing, W.; Zhang, S.; Cao, D.; Hsieh, C.-Y.; Hou, T. Mf-SuP-pK_a: multi-fidelity modeling with subgraph pooling mechanism for pK_a prediction. *Acta Pharma. Sin. B* **2023**, *13* (6), 2572−2584.
(25) Johnston, R. C.; Yao, K.; Kaplan, Z.; Chelliah, M.; Leswing, K.; Seekins, S.; Watts, S.; Calkins, D.; Elk, J. C.; Jerome, S. V.; Repasky, M. P.; Shelley, J. C. Epik: pK_a and protonation state prediction through machine learning. *J. Chem. Theory Comput.* **2023**, *19* (8), 2380−2388.
(26) 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.; et al. pK_a measurements for the SAMPL6 prediction challenge for a set of kinase inhibitor-like fragments. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 1117−1138.
(27) Lee, A. C.; Crippen, G. M. Predicting pK_a. *J. Chem. Inf. Model.* **2009**, *49* (9), 2013−2033.
(28) Rupp, M.; Korner, R.; Tetko, I. V. Predicting the pKa of small molecules. *Comb. Chem. High Throughput Screen.* **2011**, *14* (5), 307−327.
(29) Leeson, L. J.; Krueger, J. E.; Nash, R. A. Concerning the structural assignment of the second and third acidity constants of the tetracycline antibiotics. *Tetrahedron Lett.* **1963**, *4* (18), 1155−1160.
(30) Gunner, M. R.; Murakami, T.; Rustenburg, A. S.; Işık, M.; Chodera, J. D. Standard state free energies, not pK_as, are ideal for describing small molecule protonation and tautomeric states. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 561−573.
(31) Brönsted, J. N. Einige bemerkungen über den begriff der säuren und basen. *Recl. Trav. Chim. Pays-Bas* **1923**, *42* (8), 718−728.
(32) Lowry, T. M. The uniqueness of hydrogen. *J. Soc. Chem. Ind.* **1923**, *42* (3), 43−47.
(33) Edsall, J. T.; Wyman, J. *Biophysical chemistry. Vol. 1, Themrodynamics, electrostatics, and the biological significance of the properties of matter*; Academic Press, 1958.
(34) Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; Ke, G. Uni-Mol: A universal 3D molecular representation learning framework. In *Paper presented at the Eleventh International Conference on Learning Representations*, 2023, https://openreview.net/forum?id=6K2RM6wVqKu.
(35) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, **2017**, Vol. *30*, https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
(36) Gaulton, A.; Bellis, L. J.; Patricia Bento, A.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40* (D1), D1100−D1107.
(37) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: an open-source program for chemistry aware data visualization and analysis. *J. Chem. Inf. Model.* **2015**, *55* (2), 460−473.
(38) *Internet Bond-energy Databank (pKa and BDE)−iBonD Home Page*, 2017. https://ibond.las.ac.cn/.
(39) *Prediction of dissociation constant using microconstants*. https://docs.chemaxon.com/display/docs/attachments/attachments_1814016_1_Prediction_of_dissociation_constant_using_microconstants.pdf, accessed on Jun 3, 2024.
(40) Masunov, A. ACD/I-Lab 4.5: an internet service review. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (4), 1093−1095.
(41) Farajtabar, A.; Gharib, F. Solvent effect on protonation constants of salicylic acid in mixed aqueous organic solutions of dmso. *Monatsh. Chem.* **2010**, *141*, 381−386.
(42) Teitelbaum, A. B.; Derstuganova, K. A.; Shishkina, N. A.; Kudryavtseva, L. A.; Bel'skii, V. E.; Ivanov, B. E. Tautomerism in the ortho-aminomethylphenols. *Bull. Acad. Sci. USSR Div. Chem. Sci.* **1980**, *29*, 558−562.
(43) Nelson, D. L.; Lehninger, A. L.; Cox, M. M. *Lehninger principles of biochemistry*; Macmillan, 2008.
(44) https://colab.research.google.com/github/mayrf/pkasolver/blob/main/notebooks/pka_prediction.ipynb accessed on Jun 3, 2024.
(45) Bochevarov, A. D.; Watson, M. A.; Greenwood, J. R.; Philipp, D. M. Multiconformation, density functional theory-based pK_a prediction in application to large, flexible organic molecules with diverse functional groups. *J. Chem. Theory Comput.* **2016**, *12* (12), 6001−6019.
(46) Yu, H. S.; Watson, M. A.; Bochevarov, A. D. Weighted averaging scheme and local atomic descriptor for pK_a prediction based on density functional theory. *J. Chem. Inf. Model.* **2018**, *58* (2), 271−286.
(47) 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.
(48) Li, M.; Zhang, H.; Chen, B.; Yan, W.; Guan, L. Prediction of pKa values for neutral and basic drugs based on hybrid artificial intelligence methods. *Sci. Rep.* **2018**, *8* (1), 3991.
(49) Chen, B.; Zhang, H.; Li, M. Prediction of pK(a) values of neutral and alkaline drugs with particle swarm optimization algorithm and artificial neural network. *Neural Comput. Appl.* **2019**, *31*, 8297−8304.
(50) Hunt, P.; Hosseini-Gerami, L.; Chrien, T.; Plante, J.; Ponting, D. J.; Segall, M. Predicting pK_a using a combination of semi-empirical quantum mechanics and radial basis function methods. *J. Chem. Inf. Model.* **2020**, *60* (6), 2989−2997.
(51) Sinha, V.; Laan, J. J.; Pidko, E. A. Accurate and rapid prediction of pK_a of transition metal complexes: semiempirical quantum chemistry with a data-augmented approach. *Phys. Chem. Chem. Phys.* **2021**, *23* (4), 2557−2567.
(52) Raddi, R. M.; Voelz, V. A. Stacking gaussian processes to improve *pK$_a$* predictions in the sampl7 challenge. *J. Comput.-Aided Mol. Des.* **2021**, *35* (9), 953−961.
(53) Tielker, N.; Eberlein, L.; Güssregen, S.; Kast, S. M. The SAMPL6 challenge on predicting aqueous pK_a values from EC-RISM theory. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 1151−1163.
(54) Ristić, M. M.; Petković, M.; Milovanović, B.; Belić, J.; Etinski, M. New hybrid cluster-continuum model for pKa values calculations: Case study of neurotransmitters' amino group acidity. *Chem. Phys.* **2019**, *516*, 55−62.
(55) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **1980**, *36* (22), 3219−3228.
(56) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **2019**, *15* (3), 1652−1671.
(57) Belding, L.; Stoyanov, P.; Dudding, T. Synthesis, theoretical analysis, and experimental pK_a determination of a fluorescent, nonsymmetric, in−out proton sponge. *J. Org. Chem.* **2016**, *81* (1), 6−13.
(58) Margetić, D.; Ishikawa, T.; Kumamoto, T. Exceptional superbasicity of bis (guanidine) proton sponges imposed by the bis (secododecahedrane) molecular scaffold: A computational study. *Eur. J. Org. Chem.* **2010**, *34* (2010), 6563−6572.
(59) Arnett, E. M.; Harrelson, J. A. Ion pairing and reactivity of enolate anions. 7. A spectacular example of the importance of

https://doi.org/10.1021/jacsau.4c00271
JACS Au 2024, 4, 3451−3465

rotational barriers: the ionization of Meldrum's acid. *J. Am. Chem. Soc.* **1987**, *109* (3), 809−812.
(60) Wang, X.; Houk, K. N. Theoretical elucidation of the origin of the anomalously high acidity of Meldrum's acid. *J. Am. Chem. Soc.* **1988**, *110* (6), 1870−1872.
(61) Byun, K.; Mo, Y.; Gao, J. New insight on the origin of the unusual acidity of Meldrum's acid from ab initio and combined QM/MM simulation study. *J. Am. Chem. Soc.* **2001**, *123* (17), 3974−3979.
(62) Nakamura, S.; Hirao, H.; Ohwada, T. Rationale for the acidity of Meldrum's acid. consistent relation of C-H acidities to the properties of localized reactive orbital. *J. Org. Chem.* **2004**, *69* (13), 4309−4316.
(63) Wahl, O.; Sander, T. Tautobase: An open tautomer database. *J. Chem. Inf. Model.* **2020**, *60* (3), 1085−1089.
(64) 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.
(65) 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.
(66) Bergazin, T. D.; Tielker, N.; Zhang, Y.; Mao, J.; Gunner, M. R.; Francisco, K.; Ballatore, C.; Kast, S. M.; Mobley, D. L. Evaluation of log *P*, pK_a, and log *D* predictions from the SAMPL7 blind challenge. *J. Comput.-Aided Mol. Des.* **2021**, *35* (7), 771−802.
(67) Nandkeolyar, A.; Bahr, M. N.; Mobley, D. L. Insights from the SAMPL8 physical properties blind prediction challenge. *Biophys. J.* **2023**, *122* (3), 423a.
(68) Baevski, A.; Auli, M. Adaptive input representations for neural language modeling. *arXiv preprint arXiv:1809.10853* **2018**.
(69) Shuaibi, M.; Kolluru, A.; Das, A.; Grover, A.; Sriram, A.; Ulissi, Z.; Zitnick, C. L. Rotation invariant graph neural networks using spin convolutions. *arXiv preprint arXiv:2106.09575* **2021**.
(70) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding, June 2019. In *Paper presented at proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*; Human Language Technologies, 2019; Vol. *1* (Long and Short Papers).
(71) Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.*Improving language understanding by generative pre-training*, 2018, https://openai.com/research/language-unsupervised.
(72) Radford, A.; Jeffrey, W.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. et al. *Language models are unsupervised multitask learners*, 2019, https://openai.com/index/better-language-models/.
(73) Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. et al. Language models are few-shot learners. In *Advances in neural information processing systems*, **2020**; Vol. 33, pp. 1877−1901
(74) Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. In *Paper presented at the International Conference on Learning Representations*, 2021.

https://doi.org/10.1021/jacsau.4c00271
JACS Au 2024, 4, 3451−3465