Protein pK_a prediction with machine learning

Zhitao Cai,[†] Fangfang Luo,^{†,‡} Yongxian Wang,[†] Enling Li,[†] and Yandong Huang^{*,†}

[†]*College of Computer Engineering, Jimei University, Xiamen 361021, China*

[‡] *Joint first author.*

E-mail: yandonghuang@jmu.edu.cn

Abstract

Protein pK_a prediction is essential for the investigation of pH-associated relationship between protein structure and function. In this work, we introduce a deep learning based protein pK_a predictor DeepKa, which is trained with 12809 pK_a 's derived from continuous constant pH molecular dynamics (CpHMD) simulations of 279 soluble proteins. Here the CpHMD implemented in the Amber molecular dynamics package has been employed (Huang, Harris, and Shen J. Chem. Inf. Model. 2018, 58, 1372-1383). Notably, to deal with the finite-size effect, grid charges are proposed to represent electrostatics, but rather the previously used atomic charges. We show that the prediction accuracy by DeepKa is close to that by CpHMD benchmarking simulations, validating DeepKa as an efficient protein pK_a predictor. In addition, the training dataset created in this study can be applied to the development of machine learning based protein pK_a predictors in future. Finally, the new grid charge representation is general and applicable to other topics, like the protein-ligand binding affinity prediction.

Introduction

Solution pH plays a key role in many biological events, such as proton-coupled transports of ions¹ or small molecules² across the cellular membrane, enzyme catalyses that require

proton donors (acids) and nucleophiles (bases) in the active sites to carry out the catalytic functions³ and pH-gradient driven ATP synthesis in the cellular energy metabolism.⁴ pK_a on the other hand measures how tight a proton is held by an ionizable site and the resulting microscopic protonation and deprotonation equilibria at a certain pH condition. Thus the predictions of pK_a 's are essential to understand the molecular mechanisms of pH-mediated processes in biology and chemistry. pK_a 's can be determined by experiments, typically NMR for biomolecules.^{5,6} However, NMR experiments are often expensive and time consuming, making theoretical methods more favored.

A fast way to compute pK_a 's is coming up with a simple empirical function, such as PropKa that calculates pK_a shifts contributed by desolvation, hydrogen bonding and charge-charge interaction respectively.⁷⁻¹⁰ However, the construction of such an empirical function requires the knowledge of pK_a 's determinants. In addition, measured pK_a 's are needed to determine model coefficients. On the other hand, in the framework of a classic force field, like CHARMM¹¹ and Amber,¹² the pK_a shift can be achieved by estimating the free energy perturbation (FEP) of transferring an ionizable group from aqueous solution to a biomolecule.¹³ For instance, FEP can be approximated numerically with the Poisson-Boltzmann (PB) based methods, such as MCCE,¹⁴ H++,¹⁵ and DelphiKa.¹⁶⁻¹⁸ PB based approaches assume that the structure and the charge distribution of a biomolecule are fixed. As a consequence, the calculated pK_a value of an ionizable group might be biased to a specific conformation, which is known as microscopic pK_a .

Alternatively, constant pH molecular dynamics (CpHMD) simulations have been developed to calculate FEP and the resulting pK_a 's. CpHMD methods can be divided into two classes. One is the discrete protonation-state CpHMD that periodically updates the protonation states stochastically in the Monte-Carlo step followed by a short molecular dynamics simulation.¹⁹ The other is the continuous protonation-state CpHMD that propagates simultaneously the spacial and titration coordinates in the framework of λ dynamics.^{20,21} During CpHMD simulations, molecular conformations are coupled with protonation states. Simul-

taneously interactions of titration events are elucidated. Thus macroscopic pK_a 's can be obtained with CpHMD simulations, in accord with the apparent pK_a 's measured by experiments. Recently CpHMD simulations have been proved applicable to those challenging systems, such as enzymes²²⁻²⁴ and membrane proteins that are very dynamic,²⁵⁻²⁷ demonstrating high robustness. However the price comes that CpHMD simulations are time consuming when compared with other state-of-the-art methods. Thus a cheap method is demanded that offers comparable precision with that by a CpHMD method.

Unlike the knowledge based approaches above, machine learning (ML) methods build algorithms on sample data to accomplish complex tasks.²⁸ For instance, ML methods have been successful in predicting pK_a 's for aliphatic amines,²⁹ protein-ligand binding affinity,^{30,31} and protein secondary/tertiary structure.³²⁻³⁴ Like the protein-ligand binding affinity, the pK_a value of an ionizable group is sensitive to the varying of protein environment.²² Thus, exploiting the complex protein environment patterns is the primary task to model a protein pK_a predictor. A valid training dataset is fundamental to a ML model. However, protein pK_a dataset is quite lack. At present, experimental pK_a values of 1350 ionizable residues in 119 proteins are available in the pK_a database PKAD,³⁵ which is far not enough for a rational modeling with ML.

In this work, a deep learning based protein pK_a predictor DeepKa has been developed and evaluated. The training dataset HH279 that includes 12809 pK_a 's has been created based on continuous protonation-state CpHMD simulations of 279 proteins.^{36,37} The deep learning architecture proposed by Steniewska-Dziubinska and coworkers has been applied with minor modification.³⁰ To reduce the computational cost, instead of the entire protein, a cubic box³⁰ or in another study a sphere³⁸ is defined as model input. Such a truncation scheme excludes the long-range electrostatics beyond the cutoff. In physics, energy fluctuations could be observed at the cutoff,³⁹ which may result in artificial pK_a shift. Noting that electrostatics is the major contributor to pK_a shifts,¹³ it's of importance to eliminate the cutoff-induced finite-size effect. In this study, a general solution has been proposed to smooth the charge

distribution and the resulting electrostatic energy. Finally, based on the PKAD database mentioned above, the test dataset HH69 that contains 472 pK_a 's was created to evaluate DeepKa. To find out the theoretical limit, CpHMD simulations of benchmark proteins in HH69 were carried out. The prediction accuracy by DeepKa will be compared with that by CpHMD simulations to examine the predictive power of the present deep learning model. Finally, the proposed grid charge representation will be validated.

Methods and Materials

Feature representation

The protein environment of an ionizable group should be transformed and encoded to make it readable by a neural network. Instead of the entire protein, a 20-Å cubic box or grid was defined to present the protein environment.³⁰ The center of the box in this study is the titratable site of interest. As a result, the minimal distance from the center to the edge of the box is 10-Å, which has been used as the cut-off distance of non-bonded interactions in a periodic system.⁴⁰ Then heavy atoms were mapped to the 3D grid with 1-Å resolution.³⁰ As a consequence, a 4D tensor was created where the first three dimensions correspond to the Cartesian coordinates and the last dimension contains 19 features that describe an atom. As listed below, the first 17 features resemble the ones proposed by Steniewska-Dziubinska and coworkers in their Pafnucy model:³⁰

- 9 bits (one-hot or all null) that encode atom types, including *B*, *C*, *N*, *O*, *P*, *S*, *Se*, *halogen* and *metal*
- 1 integer (1, 2, or 3) that indicates atom hybridization
- 1 integer that counts the number of bonded heavy atoms
- 1 integer that counts the number of bonded hetero atoms

- 5 bits (1 if present) that encode the five properties defined with SMARTS patterns, namely *hydrophobic*, *aromatic*, *acceptor*, *donor* and *ring*⁴¹
- 1 float with grid charge
- 1 bit (1 if belonging to the titratable residue at the center of the box)

In the present work, grid charges were applied to preventing potential artifacts at the cutoff. In specific, atomic charges were spread over the grid with B-spline interpolation algorithm, which will be elucidated in details below. If more than one atom is assigned to a grid point, which is rare for a 1-Å grid unit, features from all colliding atoms were added.³⁰ As to the grid points that have no atom assigned to, all features were set zero.³⁰ The 19 features above were scaled with the z-score normalization.

Charge spreading

In this work, each atomic charge in a protein was distributed to $n \times n \times n$ grid points with the B-spline interpolation algorithm where n denotes the interpolation order that equals 4, the default value used by smoothed particle-mesh Ewald summation in CHARMM package.^{39,42} Here we take a two-dimensional mesh with two identical atomic charges as an example (Fig. 1a). As illustrated in Fig. 1b, each charge is spread to 4×4 grid points with the B-spline interpolation. The weights of the two atomic charges at a grid point are determined individually by the B-spline coefficients which can be formulated as below for one dimension.⁴²

$$M_2(u) = \begin{cases} 1 - |u - 1|, & 0 \leq u \leq 2 \\ 0, & u < 0 \text{ or } u > 2 \end{cases} \quad (1)$$

$$M_n(u) = \frac{u}{n-1} M_{n-1}(u) + \frac{n-u}{n-1} M_{n-1}(u-1) \quad (2)$$

where u is a real number that denotes the scaled fractional coordinate and the B-spline coefficients obey the normalization condition.

Figure 1: The schematic diagram of spreading two identical atomic charges to a 2D grid. (a) Atomic charges A and B are represented by two blue solid circles. (b) B-spline interpolation algorithm is utilized to distribute charges A and B over the grid with the interpolation order of 4. Open circles indicate charges A and B have been spread to the grid. Red and green solid circles represent the distributions of charges A and B respectively over the grid. The grid points with accumulated contributions by two charges are in gray color. The radii of circles indicate the accumulated weights of atomic charges at grid points. The 2D grid is truncated in (c) and (d). Charge B is removed in (c) and the grid charges that belong to the truncated area (gray) are removed in (d).

From Fig. 1b, one can see that the distribution of an atomic charge over the grid is diffusive. The contributions from two atomic charges at a grid point are accumulated (gray solid circles in Fig. 1b). To explain the potential artifact with atomic charges, the grid is truncated. As a result, charge B is fully excluded even though it’s close in position to the boundary (Fig. 1c). On the contrary, charge B can be partly captured when represented with grid charges (Fig. 1d).

Dataset preparation

Following the pafnucy model proposed by Stepniewska-Dziubinska and coworkers, atomic features were calculated with Open Babel.⁴³ In particular, the atomic charges were assigned based on Amber ff14SB force field¹² by UCSF Chimera.⁴⁴ In this work the atomic charges

were spread over the grid. The preprocessing steps were applied to the proteins in both training (HH279) and test (HH69) datasets to generate the inputs of the deep learning model. SWISS-MODEL was utilized to build the positions of missing residues in the protein crystal structures.⁴⁵ At present, four titratable residue types, namely Asp, GLu, His and Lys, are included in HH279 and HH69.

Training dataset HH279 The training set HH279 generated in this work contains 12809 pK_a values for 3600 Asp, 4171 Glu, 1419 His and 3619 Lys in 279 proteins. These proteins were selected randomly from the TR6614 which was used previously as the training set for protein secondary structure prediction.⁴⁶ The predictive power of a CpHMD method is governed mainly by the solvation energy that can be elucidated by implicit or explicit solvent model. Accordingly implicit,^{20,21,36} hybrid^{47,48} or explicit^{49,50} solvent based CpHMD methods have been developed and implemented in two popular molecular dynamics engines, CHARMM³⁹ and Amber.⁵¹ In theory, the recently proposed particle-mesh Ewald (PME) based explicit-solvent CpHMD scheme is applicable to any protein that a classical force field can describe, including membrane proteins.⁵⁰ However, explicit-solvent CpHMD is time consuming and slow to reach pK_a convergence.⁵⁰ Thus pK_a 's in HH279 were computed with the GBNeck2-based implicit-solvent CpHMD method,³⁶ which had been implemented in the GPU-accelerated *pmemd* program of the Amber package.^{37,51} GBNeck2-based implicit-solvent is limited to soluble proteins, thus there is no membrane protein in the current HH279 dataset. CpHMD simulation details can be found below.

Test dataset HH69 Most pK_a 's in the test set HH69 were extracted from the PKAD database that collects from literature the experimentally measured pK_a values of ionizable groups in proteins.³⁵ Original PKAD contains pK_a data of 1350 residues in 157 wild-type proteins. The residues with pK_a values outside the experimental pH ranges were excluded. Besides, if a residue is measured more than once and the pK_a 's are similar, only one measurement is used to avoid data duplication. It's noticed that the absolute pK_a shifts of

Lys are all smaller than 2.0 in PKAD. To examine the predictive power of DeepKa for Lys residues with high pK_a shifts, the measured pK_a 's of 3 Lys residues in 3 SNase mutants (PDB ID: 3RUZ, 4HMI and 3C1E) were added to the test set.⁵² Notably, the absolute pK_a shifts of the 3 Lys residues are larger than 2.0. To remove the possible overlaps in amino acid sequences, global sequence identity between any two proteins from HH279 and HH69 respectively was calculated with FASTA package (version 36.3.8h). One protein (PDB ID: 1UBQ) was found overlapped with the protein (PDB ID: 1WRD) in HH269 and therefore excluded from the test set. Finally, the test set HH69 contains 472 pK_a 's for 151 Asp, 178 Glu, 75 His and 68 Lys in 69 proteins. In contrast to previous CpHMD researches,^{36,37,50} much more benchmark pK_a 's have been collected, enabling definitive conclusions regarding the GBNeck2-based implicit-solvent CpHMD method.

CpHMD simulation protocols

Structure preparation The initial structures of proteins in both HH279 and HH69 were obtained from the protein data bank (PDB) by providing the PDB codes. For the proteins in HH279, the first chain in the crystal structure or the first conformation in the NMR models was selected. All proteins were acetylated at N-terminus (ACE) and amidated at C-terminus (CT2). Disulfide bonds if any were built and missing hydrogens were added.⁵³ Residue name HIS in PDB files was substituted with HSP that denotes the protonated state of His. Restraining the heavy atoms with a harmonic force constant of 50 kcal/mol.Å², hydrogen atoms were relaxed by 50 steps of steepest descent and then 10 steps of Newton-Raphson minimization in GBSW implicit solvent with the ionic strength of 0.15 M.⁵⁴ Here the CHARMM22/CMAP additive force field was utilized to represent proteins.^{11,55} To mimic proton buffers, dummy hydrogens bonded to the carboxyl oxygens of Asp/Glu were added and placed in the *syn* position. The structures were minimized following the scheme mentioned above. The fore-mentioned steps were accomplished with CHARMM program (version c45a1).³⁹ The resulting PDB files were then converted into Amber style PDB files with the

MMTSB tool set.⁵⁶ The residue names ASP, GLU and HSP were replaced manually by AS2, GL2 and HIP respectively, such that Asp/Glu/His can be identified as ionizable by Amber. If two Cys residues form a disulfide bond, their residue names were renamed too, which is from CYS to CYX. The interaction between two dummy hydrogens were excluded to remove artificial effect of nearby dummy hydrogens. In addition, the intrinsic Born radii of the His hydrogens was reduced from the GB-Neck2 default value of 1.3 to 1.17 Å.³⁷ Providing the PDB files as inputs, the parameter and coordinate files were generated with the LEaP utility in Amber where proteins were represented by ff14SB Amber force field,¹² in consistent with the force field utilized by the grid charge representation mentioned above. Finally, with the same harmonic force constant of 50 kcal/mol.Å² applied to heavy atoms, hydrogen atoms were relaxed by 100 steps of steepest descent followed by 200 steps of Newton-Raphson minimization in GBNeck2 implicit solvent (igb=8) with the ionic strength of 0.15 M and the *mbondi3* intrinsic Born radii.

Equilibration After the minimization, each structures undergoes four equilibration simulations where the restraints on heavy atoms reduced gradually from 5 to 0 kcal/mol.Å². The GPU-accelerated GBNeck2 implicit-solvent based CpHMD model was selected (iphmd=1) and Asp, Glu, His and Lys were set titratable.^{36,37} The pH value of 7 that represents neutral aqueous solution was applied during the equilibration stage. Langevin algorithm with the collision frequency of 1.0 p⁻¹ was utilized for constant temperature (300 K) simulation (ntt=3). SHAKE algorithm was used to constrain the bonds involving hydrogen atoms.⁵⁷ As a result, time step of 2 fs can be used. The running length for each stage is 4 ps.

Production pH-based asynchronous replica-exchange protocol was utilized to improve the sampling of conformations as well as protonation states.^{47,58} It has been demonstrated that most pK_a 's are converged within 2 ns per replica, which is supported by the block analysis on the training dataset HH279 (Fig. S1). Thus, unless otherwise noted each replica was run for 2 ns and the first 0.2 ns was discarded for later pK_a calculation. Exchanges were

attempted every 1000 MD steps or 2 ps. Replica pH ranges from 0.0 to 12.5 with the interval of 0.5. As a result, there are 26 replicas in total and the accumulative running length is 52 ns for each protein. The pH-based asynchronous replica-exchange CpHMD simulations were carried out on two NVIDIA RTX 2080TI graphics processing units (GPUs) each with 11 Gigabyte (GB) memory.

p*K*_a calculation The unprotonated fraction *S* at a pH condition can be calculated as below.

$$S = \frac{N^{unpro}}{N^{unpro} + N^{pro}} \quad (3)$$

where N^{unpro} and N^{pro} denote the counts of unprotonated ($\lambda > 0.8$) and protonated ($(\lambda < 0.2)$) states respectively. The variable λ that ranges from 0 to 1 is the titration coordinate of an ionizable group. Then residue-specific p*K*_a's were computed by fitting *S* at different pH conditions to the generalized Henderson-Hasselbalch equation:

$$S = \frac{1}{1 + 10^{h(pK_a - pH)}}, \quad (4)$$

where the Hill coefficient *h* is the slope of the titration curve that represents the coupling between ionizable groups.

Deep learning architecture

Convolutional neural network (CNN) has been succeeded in different areas. In particular, CNN was validated recently by Stepniewska-Dziubinska and coworkers in predicting protein-ligand binding affinity that plays a key role in drug design.³⁰ In fact, the p*K*_a value of an ionizable site reflects the binding affinity of a proton to this site at a certain pH condition. Thus, the deep learning architecture proposed by Stepniewska-Dziubinska and coworkers was applied in this work,³⁰ aiming at discovering protein environment patterns of proton

Figure 2: The 3D convolutional neural network architecture of DeepKa. The titratable carboxyl oxygens of an Asp residue is highlighted with red balls. The input structure is the 20-Å open red cubic box centralized at the Asp. The four solid boxes represent 4D tensors with the grid sizes of $21 \times 21 \times 21$, $11 \times 11 \times 11$, $6 \times 6 \times 6$, and $3 \times 3 \times 3$ respectively. The corresponding channel counts of the grids are 19, 64, 128 and 256 respectively. In the dense block, the input layer (gray) includes 6912 neurons and the following three hidden layers (blue) have 1000, 500 and 200 neurons respectively. The output layer with one single node (green) is the predicted pK_a .

binding sites and the resulting pK_a 's. As illustrated in Fig. 2, a cubic box centralized at the titratable site of interest was cropped from a protein. The cubic box was then transformed to a 4D tensor as the input of the CNN (Fig. 2 input tensor). Followed are three convolutional layers with 64, 128 and 256 5-Å cubic filters respectively. For each convolutional layer, a 3D convolution was coupled with a max pooling, between which the action function and exponential linear units (ELU) followed by a batch normalization (BN) were used to speed up network convergence. Here the max pooling uses a 2-Å cubic patch. The BN layers were not considered by Stepniewska-Dziubinska and coworkers in their model. In addition, ReLU was employed in their model instead of the present ELU. In this study ReLU would result in dead neurons, which is known as dead ReLU obstacle that can be solved by ELU. The output 4D tensor of the convolution block is a 3×3 grid and the forth vector contains 256 channels. The tensor was then flattened into a 1D array with 6912 neurons. This array was fed to four consecutive dense layers. To overcome overfitting and vanishing, ELU function and BN

operations were applied to the dense layers too. Finally, the messages in the 200-neuron array were integrated by a single node where the pK_a value was predicted.

Implementation details

First of all the Kaiming algorithm embedded in the Pytorch environment was used to initialize the training.⁵⁹ More precisely, the weights and biases were initialized with uniform distribution. The network was then trained with the Adam optimizer. The learning rate was set 10^{-3} . To reduce overfitting, dropout was applied to the dense layers. A large mini batch size of 128 was used, which is favored in the presence of BN layers. To reduce unexpected sensitivity to the orientation of a protein structure, the cubic box was rotated before entering the network.³⁰ There are 24 combinations of 90 rotations around three axes for a cubic box. Finally 720 epochs in total were carried out to make sure each orientation was visited randomly for 30 times in average. After three rounds of training, the model with the lowest root mean square deviation from HH69 was selected as the final model. The experiments were implemented under the PyTorch (version 1.8.1) environment and the model was trained on a single NVIDIA RTX 2080TI GPU with 11 GB memory.

Performance evaluation

Root mean square deviation (RMSD) or Mean absolute error (MAE) is often used to measure the overall accuracy, whereas Pearson’s correlation coefficient (R) is indispensable for investigating the robustness of a model. The reference pK_a ’s of Asp, GLu, His and Lys applied in this work are 3.67, 4.25, 6.54⁶⁰ and 10.40⁶¹ respectively. pK_a shifts from the reference values tell to what extent ionizable sites are affected by proteins. Up to 83.0% of pK_a ’s shifts in HH69 are lower than 1.0, which would mislead the discussion below. Thus the subset of HH69, namely HH69S, was utilized to assess the model. HH69S contains 167 pK_a ’s for 46 Asp, 60 Glu, 31 His and 30 Lys in 52 proteins (Tab. S2). In HH69S the pK_a shifts below 1.5 are less populated and simultaneously the normal distribution is satisfied.

Results and discussion

Protein pK_a prediction accuracy

Figure 3: Pairwise comparison of the pK_a 's (a, c and e) and pK_a shifts (b, d and f) from DeepKa, experiment (Expt) and CpHMD simulations (CpHMD). Data for Asp, Glu, His and Lys residues in HH69S, the subset of HH69, are colored pink, gray, green, and blue, respectively. The diagonal ($y = x$) and linear regression lines are colored gray (dashed) and black (solid), respectively. The open circles are data for the residues in HH69. RMSD, regression slope (m), and R are displayed.

Unless otherwise noted, the following analyses are based on the subset of test dataset HH69, namely HH69S. Nevertheless, pK_a 's in HH69 are displayed in Fig. 3 as background. As illustrated in Fig. 3, both RMSD and R values by the proposed DeepKa model (Fig. 3a and b) agree with those by the benchmark CpHMD simulations (Fig. 3c and d). However,

from the linear regression slopes, one can see that the pK_a shifts are underestimated systematically by both DeepKa (Fig. 3b) and CpHMD (Fig. 3d). In fact, it’s the intrinsic error of GBNeck2-based implicit solvent applied by CpHMD that desolvation penalties are underestimated.³⁶ Now that the present training set is created by GBNeck2-based CpHMD simulations, it’s acceptable that the systematic error of CpHMD scheme is inherited by deep learning. To get rid of such errors, more accurate CpHMD approach, like PME-based CpHMD, is demanded.⁵⁰

Notably, the regression slope m by DeepKa (Fig. 3b) is 0.24 lower than that by CpHMD (Fig. 3d), implying other errors arising from deep learning. Thus, to examine the prediction accuracy of DeepKa with respect to CpHMD, the correlation of pK_a ’s or pK_a shifts produced by DeepKa and CpHMD simulations was computed where the intrinsic error of CpHMD can be cancelled out. From the RMSD, m and R values, one can see that the pK_a ’s or pK_a shifts by DeepKa are closer to those by CpHMD simulations (Fig. 3e and f) than the measured values (Fig. 3a and b), which is normal as DeepKa is trained with the pK_a ’s calculated by CpHMD simulations. At the same time these quantities, especially the regression slope m in Fig. 3f, still have measurable deviations from the ideal values of 0.0, 1.0 and 1.0 for RMSD, m and R respectively that correspond to two identical data sets.

From Fig. 3b and f, one can see that the pK_a ’s of Glu residues with large shifts that span a range of 2 to 4 are underestimated by DeepKa. Thus, it’s of interest to examine the predicted pK_a ’s of Asp, Glu, His and Lys individually. In addition to the RMSD, R and regression slope m , mean absolute error (MAE) and maximum absolute deviation (Max) are considered too. Now that R and m were calculated for individual residue types, the analyses on pK_a ’s are equivalent to those on pK_a shifts. Indeed, as summarized in Tab. 1, m or R values for Glu are most underestimated when compared with those for another three residue types. It’s found from HH279 that 66.7%, 81.9%, 65.0% and 86.0% of pK_a shifts for Asp, Glu, His and Lys respectively are smaller than 1.0 (Tab. S1), which might lead to the prediction biased to the reference pK_a ’s, and therefore reduced correlation with experiment.

Table 1: Comparisons of Asp, Glu, His and Lys pK_a 's by DeepKa and CpHMD simulations and experiment

Res	MAE	RMSD	Max	R	m
DeepKa vs Experiment					
All	0.78	1.02	3.09	0.71	0.41
Asp	0.75	0.93	3.09	0.76	0.40
Glu	0.90	1.12	2.50	0.53	0.24
His	0.65	0.87	2.83	0.75	0.63
Lys	0.71	0.90	2.68	0.84	0.51
CpHMD vs Experiment					
All	0.74	1.02	4.27	0.73	0.65
Asp	0.93	1.24	4.27	0.67	0.54
Glu	0.78	1.06	2.94	0.62	0.50
His	0.57	0.80	2.49	0.80	0.77
Lys	0.53	0.72	2.35	0.89	0.91
DeepKa vs CpHMD					
All	0.64	0.85	2.31	0.76	0.49
Asp	0.78	0.95	2.10	0.77	0.51
Glu	0.61	0.86	2.31	0.58	0.32
His	0.59	0.78	1.87	0.76	0.67
Lys	0.50	0.75	2.00	0.95	0.56

According to the solvent accessible surface area (SASA) calculations of the four residue types in HH279,⁶²⁻⁶⁴ most are solvent accessible (Fig. S2), which is responsible for the abundance of pK_a 's nearby the reference values. Encouragingly, though the distribution of pK_a 's for Lys, like that for Glu, is concentrated around the reference pK_a 's, the measured pK_a down shifts of the three Lys residues in the 3 SNase mutants are captured by DeepKa, enabling higher R value of 0.84 for Lys. Noting that the three Lys residues are all highly buried, we suggest that the pK_a down shifts for Lys be dominated by desolvation, which facilitates the learning of large down shifts based on a small sample. Once the three data points are removed from Fig. 3a, the correlation coefficient reduces significantly from 0.84 to 0.28. On the other hand, the prediction of pK_a up shifts for Glu is much worse by DeepKa. In contrast to Lys, a buried Glu would involve in complex hydrogen bonding network which leads to pK_a down shifting and therefore compensate to some extent desolvation-induced pK_a up shifting.²² Concentrated distribution of training pK_a 's around reference values would result

in low sensitivity of DeepKa to the varying of protein environment. Thus, more sufficient pK_a 's for buried Glu in the training dataset are required to improve the predicting accuracy. The training set HH279 proposed in this work obeys the population of protein pK_a 's in nature. Thus, simply adding protein samples to the training set is not likely to change the distribution of pK_a 's and the resulting accuracy. It's possible to improve the accuracy via manipulating the distribution of pK_a 's for training. However, this is beyond the scope of the present work and will be investigated in the future.

Figure 4: Comparison of the pK_a error histograms for (a) Asp, (b) Glu, (c) His and (d) Lys by DeepKa (pink) and CpHMD simulations (gray). ΔpK_a is the deviation of calculated pK_a 's by CpHMD/DeepKa from measured values.

It's noticed that a large proportion of points for Asp in Fig. 3c are below the diagonal line. Thus it's of interest to calculate the histograms of pK_a differences between DeepKa/CpHMD and experiment. Here the test dataset HH69 is utilized. As illustrated in Fig. 4, the distributions by DeepKa are similar to those by CpHMD except for Asp (Fig. 4a). In specific, a systematic movement of about 0.5 relative to the distribution by CpHMD is observed. Notably the histograms (Fig. 4a and b) for Asp and Glu by the CpHMD simulations performed in this study are consistent with the counter plots of Figure 4d and e in the previous research that implemented the GPU-accelerated GBNeck2-based implicit-solvent CpHMD.³⁷ Two histograms are centralized around -0.25 and 0.25 respectively, which would lead to sim-

ilar RMSDs. Noting that the present model parameters were determined by RMSD alone, such a deviation is allowed. Remarkably, extensive CpHMD simulations performed in this work have verified that the GBNeck2-based implicit-solvent CpHMD yields accurate prediction of pK_a 's for His (Fig. 4c) and Lys (Fig. 4d), which is not for sure because of the lack of benchmark proteins in the previous research.³⁷

Validation of grid charge representation

In this work, atomic charges in a protein were substituted by grid charges to represent electrostatics, the principal contributor to pK_a shifts.²² As illustrated in Fig. 5a, the atoms in the SNase mutant V74K (PDB ID: 3RUZ) are colored based on the atomic charges that are discrete in space. When the atomic charges are distributed over the grid, one can see from Fig. 5b that the charge density is smoothed.

To assess the new grid charge representation, as a test atomic charges were applied to DeepKa. Comparing panel a or b in Fig. 6 with that in Fig. 3, it's found that two charge representations provide similar prediction accuracy. In addition, from Fig. 6c and d, the grid charge based model reproduces the pK_a 's by the atomic charge based model, validating the grid charge representation. It's not a surprise as the availability of charge spreading by B-spline interpolation algorithm to estimate the electrostatics of a periodic system has been verified extensively by PME-based explicit-solvent molecular dynamics simulations. Noting that the minimal distance from the ionizable site at the center to the margin of the cubic box is set 10Å in the present work, most electrostatic interactions that the ionizable site can sense have been included such that the finite-size effect can be ignored for most cases. Under the circumstances, the grid charge representation won't surpass the atomic charge representation. A more extensive study of the grid charge representation will be done in the future.

Figure 5: (a) Atomic and (b) grid charge representations of SNase mutant V74K (PDB ID: 3RUZ). (a) Each atom is represented with point model and colored according to the atomic charge (q^{atom}) assigned by the Amber ff14SB force field.¹² The side chain of K74 is highlighted with stick and ball model and colored yellow. The crystal structure is displayed as well with NewCartoon model. (b) Atomic charges are distributed over the grid with the B-spline interpolation algorithm where the interpolation order equals 4. A 20-Å cubic box centralized at the sidechain nitrogen of K74 is cropped. The grid charges inside the box are colored red. To make the plot clean, the sign is not displayed. The absolute values of grid charges ($|q^{grid}|$) are coupled with the radii of circles.

Concluding remarks

In this work, the training dataset HH279 that includes 12809 protein pK_a 's has been created based on continuous constant pH molecular dynamics (CpHMD) simulations of 279 soluble proteins. To the best of our knowledge, HH279 is the first training dataset available for machine learning of protein pK_a 's. Besides, the test dataset HH69 that contains 472 measured pK_a 's extracted from the database PKAD has been created to benchmark a machine

Figure 6: Comparison of the pK_a 's and pK_a shifts (a-b) from atomic charge based DeepKa to experiment and (c-d) from grid charge based DeepKa to atomic charge based DeepKa. The subset of HH69, namely HH69S, are colored cyan. The diagonal ($y = x$) and linear regression lines are colored gray (dashed) and black (solid), respectively. The gray circles are data for the residues in HH69. RMSD, regression slope (m), and R are displayed.

learning model trained by HH279. Based on the two datasets, a deep learning based protein pK_a predictor DeepKa has been developed. Encouragingly, the prediction accuracy yielded by DeepKa is close to that by CpHMD benchmark simulations, validating DeepKa as an efficient protein pK_a predictor.

Instead of previously used atomic charges, grid charges have been proposed to represent electrostatics. The new grid charge representation is general and has been applied to the prediction of protein-ligand binding affinity that also requires accurate description of electrostatics, especially when the ligand is charged (data not published).

The success of DeepKa has in turn demonstrated that HH279 in combined with HH69 enables sustained development of protein pK_a predictors in future via machine learning. In the mean time, the training dataset HH279 will be continuously updated to improve the sampling of protein active site pK_a 's.

Acknowledgement

Y.H. acknowledges the support of the National Natural Science Foundation of China (11804114) and the National Key R&D Program of China (2018YFD0901004)

Supporting Information Available

Supplemental figures and tables including the statistics of the training and test datasets, the distribution of solvent accessible surface areas of residues in the training dataset, and the pK_a 's convergence analysis of CpHMD simulations.

Data Availability Statement

The code of DeepKa and the relevant data can be downloaded from <https://gitlab.com/yandonghuang/deepka>.

References

- (1) Lee, C.; Kang, H. J.; von Ballmoons, C.; Newstead, S.; Uzdavinyis, P.; Dotson, D. L.; Iwata, S.; Beckstein, O.; Cameron, A. D.; Drew, D. A two-domain elevator mechanism for sodium/proton antiport. *Nature* **2013**, *501*, 573–577.
- (2) Qian, H.; Wu, X.; Du, X.; Yao, X.; Zhao, X.; Lee, J.; Yang, H.; Yan, N. Structural basis of low-pH-dependent lysosomal cholesterol egress by NPC1 and NPC2. *Cell* **2020**, *182*, 1–14.
- (3) Li, Z.; Zhang, X.; Wang, Q.; Li, C.; Zhang, N.; Zhang, X.; Xu, B.; Ma, B.; Schrader, T. E.; Coates, L.; Kovalevsky, A.; Huang, Y.; Wan, Q. Understanding the pH-Dependent Reaction Mechanism of a Glycoside Hydrolase Using High-Resolution X-ray and Neutron Crystallography. *ACS Catalysis* **2018**, *8*, 8058–8069.

- (4) Singharoy, A. et al. Atoms to phenotypes: Molecular design principles of cellular energy metabolism. *Cell* **2019**, *179*, 1098–1111.
- (5) Xiao, S.; Patsalob, V.; Shana, B.; Bia, Y.; Greena, D. F.; Raleigh, D. P. Rational modification of protein stability by targeting surface sites leads to complicated results. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 11337—11342.
- (6) Webb, H.; Tynan-Connolly, B. M.; Lee, G. M.; Farrell, D.; O’Meara, F.; Søndergaard, C. R.; Teilum, K.; Hewage, C.; McIntosh, L. P.; Nielsen, J. E. Remeasuring HEWL pK_a values by NMR spectroscopy: Methods, analysis, accuracy, and implications for theoretical pK_a calculations. *Proteins* **2011**, *79*, 685—702.
- (7) Sondergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; ; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pK_a Values. *J. Chem. Theory Comput.* **2011**, *7*, 2284–2295.
- (8) Olsson, M. H. M.; Sondergaard, C. R.; Rostkowski, M.; ; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK_a Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.
- (9) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very fast prediction and rationalization of pK_a values for protein-ligand complexes. *Proteins* **2008**, *73*, 765–783.
- (10) Li, H.; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and rationalization of protein pK_a values. *Proteins* **2005**, *61*, 704–721.
- (11) A. D. MacKerell, J. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- (12) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

- (13) Bashford, D.; Karplus, M. pK_a 's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. *Biochemistry* **1990**, *29*, 10219—10225.
- (14) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. Combining conformational flexibility and continuum electrostatics for calculating pK_a s in proteins. *Biophysical journal* **2002**, *83*, 1731–1748.
- (15) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Research* **2012**, *40*, W537–W541.
- (16) Pahari, S.; Sun, L.; Basu, S.; Alexov, E. DelPhiPKa: Including salt in the calculations and enabling polar residues to titrate. *Proteins* **2018**, *86*, 1277—1283.
- (17) Pahari, S.; Sun, L.; Basu, S.; Alexov, E. DelPhiPKa web server: predicting pK_a of proteins, RNAs and DNAs. *Bioinformatics* **2016**, *32*, 614—615.
- (18) Wang, L.; Li, L.; Alexov, E. Predictions for Proteins, RNAs and DNAs with the Gaussian Dielectric Function Using DelPhiPKa. *Proteins* **2015**, *83*, 2186—2197.
- (19) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. Constant-pH molecular dynamics using stochastic titration. *J. Chem. Phys.* **2002**, *117*, 4184–4200.
- (20) Lee, M. S.; Freddie R. Salsbury, J.; III, C. L. B. Constant-pH molecular dynamics using continuous titration coordinates. *Proteins* **2004**, *56*, 738–752.
- (21) Khandogin, J.; III, C. L. B. Constant pH molecular dynamics with proton tautomerism. *Biophys. J.* **2005**, *89*, 141–157.
- (22) Huang, Y.; Yue, Z.; Tsai, C.-C.; Henderson, J. A.; Shen, J. Predicting Catalytic Proton Donors and Nucleophiles in Enzymes: How Adding Dynamics Helps Elucidate the Structure-Function Relationships. *J. Phys. Chem. Lett.* **2018**, *9*, 1179–1184.

- (23) Verma, N.; Henderson, J. A.; Shen, J. Proton-Coupled Conformational Activation of SARS Coronavirus Main Proteases and Opportunity for Designing Small-Molecule Broad-Spectrum Targeted Covalent Inhibitors. *J. Am. Chem. Soc.* **2020**, *142*, 21883–21890.
- (24) Liu, R.; Yue, Z.; Tsai, C.-C.; Shen, J. Assessing Lysine and Cysteine Reactivities for Designing Targeted Covalent Kinase Inhibitors. *J. Am. Chem. Soc.* **2019**, *141*, 6553–6560.
- (25) Huang, Y.; Chen, W.; Dotson, D. L.; Beckstein, O.; Shen, J. Mechanism of pH-dependent activation of the sodium-proton antiporter NhaA. *Nature communications* **2016**, *7*, 12940.
- (26) Henderson, J. A.; Huang, Y.; Beckstein, O.; Shena, J. Alternative proton-binding site and long-distance coupling in Escherichia coli sodium–proton antiporter NhaA. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 25517–25522.
- (27) Vo, Q. N.; Mahinthichaichan, P.; Shen, J.; Ellis, C. R. How μ -opioid receptor recognizes fentanyl. *Nature communications* **2021**, *12*, 984.
- (28) Shi, Q.; Chen, W.; Huang, S.; Wang, Y.; Xue, Z. Deep learning for mining protein data. *Briefings in Bioinformatics* **2021**, *22*, 194–218.
- (29) Lu, Y.; Anand, S.; Shirley, W.; Gedeck, P.; Kelley, B. P.; Skolnik, S.; Rodde, S.; Nguyen, M.; Lindvall, M.; ; Jia, W. Prediction of pK_a using machine learning methods with rooted topological torsion fingerprints: application to aliphatic amines. *J. Chem. Inf. Model.* **2019**, *59*, 4706–4719.
- (30) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. *Bioinformatics* **2018**, *34*, 3666–3674.

- (31) Wee, J.; Xia, K. Ollivier Persistent Ricci Curvature-Based Machine Learning for the Protein–Ligand Binding Affinity Prediction. *J. Chem. Inf. Model.* **2021**, *61*, 1617–1626.
- (32) Lyu, Z.; Wang, Z.; Luo, F.; Shuai, J.; Huang, Y. Protein secondary structure prediction with a reductive deep learning method. *Front. Bioeng. Biotechnol.* **2021**, *9*, 687426.
- (33) Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**,
- (34) Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**,
- (35) Pahari, S.; Sun, L.; Alexov, E. PKAD: a database of experimentally measured pKa values of ionizable groups in proteins. *Database* **2019**, 1–7.
- (36) Huang, Y.; Harris, R. C.; ; Shen, J. Generalized Born Based Continuous Constant pH Molecular Dynamics in Amber: Implementation, Benchmarking and Analysis. *J. Chem. Inf. Model.* **2018**, *58*, 1372–1383.
- (37) Harris, R. C.; ; Shen, J. GPU-Accelerated Implementation of Continuous Constant pH Molecular Dynamics in Amber: pKa Predictions with Single-pH Simulations. *J. Chem. Inf. Model.* **2019**, *59*, 4821–4832.
- (38) Milletti, F.; Storchi, L.; Cruciani, G. Predicting protein pK_a by environment similarity. *Proteins* **2009**, *76*, 484–495.
- (39) Brooks1, B. et al. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
- (40) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hessa, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-levelparallelism from laptops to supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.

- (41) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. De-CAF—Discrimination, Comparison, Alignment Tool for 2D PHarmacophores. *Molecules* **2017**, *21*, 1128.
- (42) Essmann, U.; Perera, L.; Berkowitz, M. L. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (43) O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: an open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
- (44) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of computational chemistry* **2004**, *25*, 1605–1612.
- (45) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; Lepore, R.; ; Schwede, T. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* **2018**, *46*, W296–W303.
- (46) Zhang, B.; Li, J.; Lu, Q. Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC Bioinformatics* **2018**, *19*, 293.
- (47) Wallace, J. A.; Shen, J. K. Continuous constant pH molecular dynamics in explicit solvent with pH-based replica exchange. *J. Chem. Theory Comput.* **2011**, *7*, 2617–2629.
- (48) Huang, Y.; Henderson, J. A.; Shena, J. Continuous constant pH molecular dynamics of transmembrane proteins. *Methods in Molecular Biology* **2021**, *2302*, 275–287.
- (49) Chen, W.; Wallace, J. A.; Yue, Z.; Shen, J. K. Introducing titratable water to all-atom molecular dynamics at constant pH. *Biophys. J.* **2013**, *105*, L15–L17.

- (50) Huang, Y.; Chen, W.; Wallace, J. A.; Shen, J. All-Atom Continuous Constant pH Molecular Dynamics With Particle Mesh Ewald and Titratable Water. *J. Chem. Theory Comput.* **2016**, *12*, 5411–5421.
- (51) Case, D. et al. AMBER 2019. *University of California, San Francisco* **2019**,
- (52) Isom, D. G.; Castañeda, C. A.; Cannon, B. R.; E., B. G.-M. Large shifts in pKa values of lysine residues buried inside a protein. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 5260–5265.
- (53) Brünger, A. T.; Karplus, M. Polar hydrogen positions in proteins: Empirical energy placement and neutron diffraction comparison. *Proteins* **1988**, *4*, 148–156.
- (54) Im, W.; Lee, M. S.; Brooks, C. L. I. Generalized Born model with a simple smoothing function. *J. Comput. Chem.* **2003**, *24*, 1691–1702.
- (55) Jr., A. D. M.; Feig, M.; III, C. L. B. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Phys. Chem. B* **2004**, *25*, 1400–1415.
- (56) Feig, M.; Karanicolas, J.; III, C. L. B. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graphics Model.* **2004**, *22*, 377–395.
- (57) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (58) Henderson, J. A.; Verma, N.; Harris, R. C.; Liu, R.; Shen, J. Assessment of proton-coupled conformational dynamics of SARS and MERS coronavirus papain-like pro-

- teases: Implication for designing broad-spectrum antiviral inhibitors. *J. Chem. Phys.* **2020**, *153*, 115101.
- (59) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* **2016**, 770–778.
- (60) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. pK values of the ionizable groups of proteins. *Protein Science* **2006**, *15*, 1214–1218.
- (61) Nozaki, Y.; Tanford, C. Examination of titration behavior. *Methods Enzymol* **1967**, *11*, 715–734.
- (62) Huang, Y.; Shuai, J. Induced Dipoles Incorporated into All-Atom Zn Protein Simulations with Multiscale Modeling. *J. Phys. Chem. B* **2013**, *117*, 6138–6148.
- (63) Weiser, J.; Shenkin, P. S.; Still, W. C. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J. Comput. Chem.* **1999**, *20*, 217–230.
- (64) Weiser, J.; Weiser, A. A.; Peter S. Shenkin, W. C. S. Neighbor-list reduction: Optimization for computation of molecular van der Waals and solvent-accessible surface areas. *J. Comput. Chem.* **1998**, *19*, 797–808.

TOC Graphic