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 and validated with the $pK_a$ values 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 avoid discontinuities at the boundary, grid charges are proposed to represent protein electrostatics. 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 and validation sets created in this study can be applied to the development of machine learning based protein $pK_a$ predictors in future. Finally, the grid charge representation is general and applicable to other topics, such as the protein-ligand binding affinity prediction.

1 Introduction

Solution pH plays a key role in many biological events, such as proton-coupled transports of ions\(^1\) or small molecules\(^2\) across the cellular membrane, enzyme catalyses that require
proton donors (acids) and nucleophiles (bases) in the active sites to carry out the catalytic functions and pH-gradient driven ATP synthesis in the cellular energy metabolism. \( pK_a \) on the other hand measures how tight a proton is hold by an ionizable site and the resulting microscopic protonation and deprotonation equilibra 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. 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. 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 CHARMM and Amber, 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. For instance, FEP can be approximated numerically with the Poisson-Boltzmann (PB) based methods, such as MCCE, H++, and DelphipKa. 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. The other is the continuous protonation-state CpHMD that propagates simultaneously the spatial and titration coordinates in the framework of \( \lambda \) dynamics. 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 that are very dynamic, such as enzymes$^{22-24}$ and membrane proteins,$^{25-27}$ demonstrating high robustness. Apart from the CpHMD methods, a hybrid Rosetta-MCCE protocol provides the $pK_a$'s where the conformational changes coupled to ionization states are considered too.$^{28}$ However the price comes that either CpHMD simulations or the hybrid Rosetta-MCCE protocol is 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.$^{29}$ For instance, ML methods have been successful in predicting $pK_a$'s for aliphatic amines,$^{30}$ protein-ligand binding affinity,$^{31,32}$ and protein secondary/tertiary structure.$^{33-35}$ Similar to the protein-ligand binding affinity, the $pK_a$ value of an ionizable group is governed by the 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 157 proteins are available in the $pK_a$ database PKAD,$^{36}$ 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 and validation sets that include 11368 and 1441 $pK_a$’s respectively have been created based on continuous protonation-state CpHMD simulations of 279 soluble proteins.$^{37,38}$ The deep learning architecture proposed by Steniewska-Dziubinska and coworkers has been applied with minor modification.$^{31}$ To reduce the computational cost, typically a cubic box$^{31}$ or in another study a sphere$^{39}$ is defined as model input, instead of the entire protein. Such a truncation scheme excludes the protein electrostatics beyond the
cutoff. In physics, energy fluctuations could be observed at the cutoff,\textsuperscript{40} which may result in artificial pK\textsubscript{a} shift. Noting that electrostatics is the major contributor to pK\textsubscript{a} shifts,\textsuperscript{13} it’s of importance to eliminate the cutoff-induced discontinuities. 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 set that contains 167 pK\textsubscript{a}’s was created to evaluate DeepKa. To find out the theoretical limit, CpHMD simulations of benchmark proteins in the test set were carried out. The prediction accuracy by DeepKa will be compared with that by CpHMD simulations as well as the PropKa method to examine the predictive power of the present deep learning model. Finally, the proposed grid charge representation will be validated.

2 Methods and Materials

2.1 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.\textsuperscript{31} The center of the box is the titratable site of interest. As a result, the minimal distance from the center to the edge of the box is 10-Å. Then heavy atoms were mapped to the 3D grid with 1-Å resolution.\textsuperscript{31} Discretized Cartesian coordinates of grid points that provide the spatial information are set as the model input. As a consequence, a 4D tensor was created where the first three dimensions correspond to the Cartesian coordinates and the last dimension contains 20 features that describe a grid point. As listed below, the first 17 features resemble the ones proposed by Steniewska-Dziubinska and coworkers in their Pafnucy model:\textsuperscript{31}

- 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)

1 integer that indicates the ionizable residue type at the center of the box (0, 1, 2 and 3 correspond to Asp, Glu, Lys and His respectively)

In this work grid charges are proposed to preventing discontinuities at the cutoff. In specific, atomic charges assigned by a force field were spread over the grid with B-spline interpolation algorithm, which will be elucidated in detail 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.\(^{31}\) As to the grid points that have no atom assigned to, all features were set zero,\(^{31}\) except for the ionizable residue type feature. These grid points will be recognized as water molecules by the model. The last two features indicate the titratable residue at the center of the cubic box and the corresponding residue type respectively. The 20 features were scaled with the z-score normalization.

2.2 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.\(^{40,42}\) Here we take a two-dimensional mesh with two identical atomic charges as an example (Fig.
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).

1a). As illustrated in Fig. 1b, each charge is spread to $4 \times 4$ 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.\(^{42}\)

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

\[
M_n(u) = \frac{u}{n-1}M_{n-1}(u) + \frac{n-u}{n-1}M_{n-1}(u-1) \tag{2}
\]

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

2.3 Dataset preparation

Following the pafnucy model proposed by Stepniewska-Dziubinska and coworkers, atomic features were calculated with Open Babel. In particular, the atomic charges were assigned based on Amber ff14SB force field by UCSF Chimera. In this work the atomic charges were spread over the grid. SWISS-MODEL was utilized to build the positions of missing residues in the protein crystal structures. At present, four titratable residue types, namely Asp, GLu, His and Lys, are considered.

Training and validation sets PHMD252 and PHMD27 The pKa values for training and validation were calculated with CpHMD simulations. The training set PHMD252 contains 11368 pKa values that correspond to 3237 Asp, 3672 Glu, 1268 His and 3191 Lys in 252 proteins. The validation set PHMD27 contains 1441 pKa values for 363 Asp, 499 Glu, 151 His and 428 Lys in 27 proteins. The proteins in PHMD252 and PHMD27 were selected randomly from the dataset TR6614 which was used previously as the training set for protein secondary structure prediction. 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, hybrid or explicit solvent based CpHMD methods have been developed and implemented in two popular molecular dynamics engines, CHARMM and Amber. In theory, the particle-mesh Ewald (PME) based explicit-solvent CpHMD scheme is applicable to any protein that a classical force field can describe, includ-
ing membrane proteins. However, explicit-solvent CpHMD methods are time consuming and slow to reach $pK_a$ convergence. Thus $pK_a$ values in both PHMD252 and PHMD27 were computed with the GBNek2-based implicit-solvent CpHMD method, which had been implemented in the GPU-accelerated pmemd program of the Amber package. GBNek2-based implicit-solvent is limited to soluble proteins, thus there is no membrane protein in the current training and validation sets. CpHMD simulation details can be found below.

**Test set EXP69S** The test set EXP69S is the subset of dataset EXP69. The $pK_a$ values in EXP69 were extracted from the PKAD database that collects from literature the experimentally measured $pK_a$ values of ionizable groups in proteins. 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. Notably, the absolute $pK_a$ shifts of the 3 Lys residues are larger than 2.0. To remove the possible duplication of a protein in two datasets, global sequence identity between any two proteins from PHMD252/PHMD27 and EXP69 respectively was calculated with FASTA package (version 36.3.8h). From the sequence identity calculations, it’s found all are below 30% except for three of 100%, 98.7% and 43.2% respectively. The protein (PDB ID: 1UBQ) which is 100% overlapped in sequence with the protein (PDB ID: 1WRD) in PHMD252 was excluded from EXP69. Finally, the dataset EXP69 contains 472 $pK_a$ values for 151 Asp, 178 Glu, 75 His and 68 Lys in 69 proteins. It’s found the distributions of $pK_a$ values around model $pK_a$'s are dense, which would lead to overestimated accuracy and underestimated correlation between calculated and measured $pK_a$ values. Thus the subset of EXP69, namely EXP69S that satisfies normal distribution of $pK_a$ values, was utilized as
the test set to evaluate the model. In specific, the p\(K_a\) value was divided into five windows, namely \([0,0.5), [0.5,1.0), [1.0,1.5), [1.5,2.0)\) and \([2.0,\infty)\). To balance the populations among the windows and reach the normal distribution, some data points in the regions of \([0,0.5)\) and \([0.5,1.0)\) have been selected randomly and removed. The number distribution of p\(K_a\)’s in the resulting EXP69S can be found from Tab. S1. Finally, EXP69S contains 167 p\(K_a\)’s for 46 Asp, 60 Glu, 31 His and 30 Lys in 52 proteins. Notably, the p\(K_a\) values around model p\(K_a\)’s are less populated (Tab. S1).

In contrast to previous CpHMD researches,\(^{37,38,50}\) much more benchmark p\(K_a\)’s have been collected, enabling definitive conclusions regarding the GBNeck2-based implicit-solvent CpHMD method.

### 2.4 CpHMD simulation protocols

**Structure preparation** The initial structures of proteins in PHMD252, PHMD27 and EXP69 were obtained from the protein data bank (PDB) by providing the PDB codes. For the proteins in PHMD252 and PHMD27, 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\(\cdot\)\(\text{Å}^2\), 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/CMAF 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).\(^{40}\) 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.\textsuperscript{38} 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,\textsuperscript{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.Å\textsuperscript{2} 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 mbondi\textsuperscript{3} 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.Å\textsuperscript{2}. The GPU-accelerated GBNeck2 implicit-solvent based CpHMD model was selected (iphmd=1) and Asp, Glu, His and Lys were set titratable.\textsuperscript{37,38} 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\textsuperscript{-1} was utilized for constant temperature (300 K) simulation (ntt=3). SHAKE algorithm was used to constrain the bonds involving hydrogen atoms.\textsuperscript{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.\textsuperscript{47,58} From the block analysis as illustrated in Fig. S1, it’s evident that most pK\textsubscript{a}’s are converged within 2 ns per replica, consistent with that by Harris and Shen.\textsuperscript{38} Thus, unless otherwise noted each replica was run for 2 ns and the first 0.2 ns was discarded for later pK\textsubscript{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.

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

$$S = \frac{N^{\text{unpro}}}{N^{\text{unpro}} + N^{\text{pro}}}$$

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

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

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

### 2.5 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.$^{31}$ In fact, the pK$_a$ 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,$^{31}$ aiming at discovering protein environment patterns of proton binding sites and the resulting pK$_a$’s. As illustrated in Fig. 2, a cubic box centralized at the
Figure 2: The 3D convolutional neural network architecture of DeepKa. The titratable carboxy al oxygen s 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 and the channel counts per grid point are 19, 64, 128 and 256 respectively. In the dense block, the input layer includes 6912 neurons and the following three hidden layers have 1000, 500 and 200 neurons respectively. The output layer with one single node (green) is the predicted $pK_a$.

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 $3 \times 3$ 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.

### 2.6 Implementation details

First of all the Kaiming algorithm embedded in the PyTorch environment was used to initialize the training.\(^5\) 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.\(^3\) There are 24 combinations of 90° rotations around three axes for a cubic box. The availability of 90° rotations has been demonstrated previously.\(^3\) It should be noted that the selection of 90° is a balance between computational cost and speed. A smaller degree of rotation may reduce the effect of orientation, but the computational cost will increase. 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 and validation, the model with the lowest root mean square deviation from the validation set PHMD27 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.

### 2.7 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. \( pK_a \) shifts from the reference values tell to what extent ionizable sites are affected by proteins. The reference \( pK_a \) values for Asp, Glu, His and Lys are 3.67, 4.25, 6.54\(^6\) and 10.40\(^6\) respectively. In the following section, \( pK_a \) shifts will be employed to examine the ability of DeepKa in capturing the protein environment of
3 Results and discussion

3.1 Protein pK\(_a\) prediction accuracy

![Graphs showing pairwise comparison of pK\(_a\) and pK\(_a\) shifts]

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 the test set EXP69S, the subset of EXP69, 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 EXP69. RMSD, regression slope (m), and R are displayed.

Unless otherwise noted, the following analyses are based on the test set EXP69S. Nevertheless, pK\(_a\) values in EXP69 are displayed too 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\textsubscript{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.\textsuperscript{37} 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.\textsuperscript{50}

Notably, the regression slope m by DeepKa (Fig. 3b) is 0.23 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\textsubscript{a}’s or pK\textsubscript{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\textsubscript{a}’s or pK\textsubscript{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 model parameters are established by the pK\textsubscript{a}’s calculated with 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 the regression slope m in Fig. 3f, one can see that the pK\textsubscript{a} shifts are generally underestimated by DeepKa when compared with those by CpHMD simulations. As shown in Tab. 1, four ionizable residue types were examined 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\textsubscript{a}’s are equivalent to those on pK\textsubscript{a} shifts. Indeed, as summarized in Tab. 1, m or R values for Glu are most underestimated by both DeepKa and CpHMD when compared with those for another three residue types. It’s found from PHMD252 that 67.1\%, 81.6\%, 64.8\% and 85.8\% of pK\textsubscript{a} shifts for Asp, Glu, His and Lys respectively are smaller than
Table 1: Comparisons of $pK_a$ values by DeepKa, CpHMD simulations, PropKa and experiment.

<table>
<thead>
<tr>
<th></th>
<th>Res</th>
<th>MAE</th>
<th>RMSD</th>
<th>Max R</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepKa vs Experiment</td>
<td>All</td>
<td>0.81</td>
<td>1.05</td>
<td>3.27</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Asp</td>
<td>0.79</td>
<td>1.08</td>
<td>2.91</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Glu</td>
<td>0.94</td>
<td>1.15</td>
<td>3.27</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>His</td>
<td>0.78</td>
<td>1.01</td>
<td>3.15</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>Lys</td>
<td>0.63</td>
<td>0.84</td>
<td>2.71</td>
<td>0.84</td>
</tr>
<tr>
<td>CpHMD vs Experiment</td>
<td>All</td>
<td>0.74</td>
<td>1.02</td>
<td>4.27</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Asp</td>
<td>0.93</td>
<td>1.24</td>
<td>4.27</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Glu</td>
<td>0.78</td>
<td>1.06</td>
<td>2.94</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>His</td>
<td>0.57</td>
<td>0.80</td>
<td>2.49</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Lys</td>
<td>0.53</td>
<td>0.72</td>
<td>2.35</td>
<td>0.89</td>
</tr>
<tr>
<td>PropKa vs Experiment</td>
<td>All</td>
<td>0.98</td>
<td>1.37</td>
<td>5.99</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>Asp</td>
<td>0.88</td>
<td>1.21</td>
<td>5.75</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>Glu</td>
<td>0.90</td>
<td>1.37</td>
<td>5.28</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>His</td>
<td>1.17</td>
<td>1.69</td>
<td>5.99</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>Lys</td>
<td>1.07</td>
<td>1.24</td>
<td>2.19</td>
<td>0.73</td>
</tr>
<tr>
<td>DeepKa vs CpHMD</td>
<td>All</td>
<td>0.58</td>
<td>0.79</td>
<td>2.55</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Asp</td>
<td>0.61</td>
<td>0.82</td>
<td>2.55</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Glu</td>
<td>0.57</td>
<td>0.82</td>
<td>2.30</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>His</td>
<td>0.66</td>
<td>0.83</td>
<td>2.11</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Lys</td>
<td>0.45</td>
<td>0.66</td>
<td>2.31</td>
<td>0.94</td>
</tr>
</tbody>
</table>
1.0 (Tab. S1), which might lead to the prediction biased to the reference pK\textsubscript{a}'s, and therefore reduced correlation with experiment. According to the solvent accessible surface area (SASA) calculations of the four residue types in PHMD252\textsuperscript{,62-64} most are solvent accessible (Fig. S2), which is responsible for the abundance of pK\textsubscript{a}'s nearby the reference values. Encouragingly, though the distribution of pK\textsubscript{a}'s for Lys is concentrated around the reference pK\textsubscript{a}'s, like that for Glu, the measured pK\textsubscript{a} down shifts of the three Lys residues that belong to the 3 SNase mutants are captured by DeepKa, resulting in higher R value of 0.84 for Lys. Noting that the three Lys residues are all highly buried, we suggest that the pK\textsubscript{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, the correlation coefficient for Lys reduces significantly from 0.84 to 0.38. On the other hand, the prediction of pK\textsubscript{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\textsubscript{a} down shifting and therefore compensate to some extent desolvation-induced pK\textsubscript{a} up shifting.\textsuperscript{22} Concentrated distribution of training pK\textsubscript{a}'s around reference values would result in low sensitivity of DeepKa to the varying of protein environment. Thus, more sufficient pK\textsubscript{a}'s for the buried residues in the training dataset are desired to improve the predicting accuracy. The training and validation sets proposed in this work obey the population of protein pK\textsubscript{a}'s in nature. Thus, simply adding protein samples to the training set is not likely to change the distribution of pK\textsubscript{a}'s and the resulting accuracy. It’s possible to improve the accuracy via manipulating the distribution of pK\textsubscript{a}'s in PHMD252 and PHMD27. However, this is beyond the scope of the present work and will be investigated in the future.

In addition to CpHMD, the proposed DeepKa is compared to another model PropKa mentioned above. PropKa has become popular due to its high calculation speed. Like DeepKa, the pK\textsubscript{a} calculations of one protein can be done in seconds by PropKa. From Tab. 1, one can see that the prediction accuracy by DeepKa is generally better than that by PropKa, especially for His. From the correlation plots as illustrated in Fig. S3, one can see
that some data points by PropKa largely deviate from the measured values, which is rare for either DeepKa or CpHMD. In fact, 29 out of 167 residues in EXP69S have to be excluded, as their pK_a’s predicted by PropKa are negative, which is absent for either DeepKa or CpHMD.

![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.](image)

To further investigate the accuracy and robustness of DeepKa, the histograms of pK_a differences between DeepKa/CpHMD and experiment have been computed. The dataset EXP69 that contains more pK_a values was utilized. As illustrated in Fig. 4, the distributions by DeepKa are similar to those by CpHMD. The distributions for Glu, His and Lys by CpHMD simulations are concentrated around zero, demonstrating high robustness of CpHMD model. On the other hand, the distributions for the three residue types are centralized around -0.5 by DeepKa, revealing a systematic error in DeepKa. As illustrated in Fig. 4a, such a systematic error is made by CpHMD for Asp, whereas DeepKa gives more symmetric distribution around zero for Asp. Notably, the trend above is consistent with that in Tab. 1.

From the statistics for PHMD252 (Tab. S1), one can see that the counts of pK_a values with the absolute shifts larger than 1.0 for Asp, Glu, His and Lys are 1065, 676, 446 and 454 respectively where Asp has the most. The training data with large pK_a shifts are
indispensable for capturing the diversity of protein environments, which might to some extent explain the robustness of DeepKa for Asp. Now that DeepKa as well as CpHMD is examined by measured pK\textsubscript{a} values, it’s possible that DeepKa outperforms CpHMD in some circumstances.

3.2 Validation of grid charge representation

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 (\(q_{\text{atom}}\)) assigned by the Amber ff14SB force field.\textsuperscript{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 (\(|q_{\text{grid}}|\)) are coupled with the radii of circles.
In this work, the atomic charges assigned by force field were substituted by grid charges to represent protein electrostatics, the principal contributor to pK\textsubscript{a} shifts. 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 by B-spline interpolation algorithm, one can see from Fig. 5b that the charge density is smoothed.

![Image](image.png)

Figure 6: Comparison of the pK\textsubscript{a}'s and pK\textsubscript{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 EXP69, namely the test set EXP69S, 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 EXP69. RMSD, regression slope (m), and R are displayed.

To assess the new grid charge representation, as a test atomic charges were applied to DeepKa. In specific, the atomic charges assigned by force field were mapped to the nearest grids in the cubic box. A cutoff of 20 Å was used similar to the grid charge representation. Replacing the grid charges with atomic charges, the model was trained, validated and tested too by PHMD252, PHMD27 and EXP69S respectively. 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. Noting that the minimal distance from the ionizable site at the center to the margin of the cubic box is set 10 Å in the present work, most electrostatic interactions that the ionizable site can sense have been included such that the boundary 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 regarding the grid charge representation will be done in the future. For instance, a soluble protein can be solvated in a water box where the minimal distance from the protein to the margin of the box is 10 Å. Then explicit water molecules will fill the grid points not occupied by protein atoms in the cropped cubic box, which may improve the charge distribution on the surface of the protein and the resulting electrostatics. In addition, the size of the cropped cubic box can be adjusted as a model parameter to investigate the effect of the cutoff distance on the performance.

4 Concluding remarks

In this work, the training (PHMD252) and validation (PHMD27) sets that include 12809 protein pK\(_a\) values have been created based on continuous constant pH molecular dynamics (CpHMD) simulations of 279 soluble proteins. To the best of our knowledge, PHMD252 is the first training dataset available for machine learning of protein pK\(_a\)'s. Based on PHMD252 and PHMD27, a deep learning based protein pK\(_a\) predictor DeepKa has been established. The DeepKa is examined by the test set EXP69S that contains 167 measured pK\(_a\)'s, most of which are from the PKAD database. Encouragingly, the prediction accuracy yielded by DeepKa is close to that by CpHMD simulations. In addition Deepka performs better than another protein pK\(_a\) predictor PropKa which is as fast as DeepKa, validating DeepKa as an efficient protein pK\(_a\) predictor. The success of DeepKa tells that the protein pK\(_a\) datasets created in this study enable future developments of protein pK\(_a\) predictors by machine learning. In the mean time, the training and validation sets PHMD252 and PHMD27 will
keep updating to improve the population of high $pK_a$ shifts.

Instead of previously used atomic charges, grid charges have been proposed to represent electrostatics. The 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).

Finally, extensive CpHMD simulations performed in this work have verified that, in addition to Asp and Glu, the GBNeck2-based implicit-solvent CpHMD is able to yield accurate prediction of $pK_a$’s for His and Lys, which is not for sure because of the lack of benchmark proteins in previous researches.

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


TOC Graphic

![Graph showing DeepKa and grid charge correlation with calculated and experimental pKₐ values.](image)