Introduction

Basic features and the behavior of proteins, such as folding or ligand binding, heavily depend on the environmental conditions like the local protein environment. Titratable amino acids like aspartic acid (Asp) or histidine (His) are essential in many biological processes and can be either protonated or deprotonated depending on the local environment. Thus, determination of the ionization states via \( pK_a \) predictions is a prerequisite to understand the protein function. Determination of \( pK_a \) values via experimental procedures is challenging and the most reliable results for proteins can be obtained only with NMR titrations. This predicament enforces the \( pK_a \) predictions in proteins by means of theoretical applications.

There is a tremendous amount of work on theoretical \( pK_a \) calculations in the literature. These approaches can be classified into three categories as (i) microscopic methods, (ii) macroscopic methods which establish continuum electrostatics, and (iii) knowledge-based methods that rely on empirical parameters.

Among the three classes of theoretical \( pK_a \) calculations, microscopic methods such as quantum mechanical (QM) or quantum mechanics/molecular mechanics (QM/MM) approaches are considered the most reliable ones to compute \( pK_a \) values of small molecules. The most traditional approach with QM methods is to employ thermodynamic cycles by computing protonation/deprotonation free energies in the gas-phase and in solution. However, these calculations do not always provide reliable \( pK_a \) values due to reasons such as the instability of the species in the gas-phase or large conformational differences between the gas-phase and in solution. In the case of the proteins, QM approaches are impractical simply due to the system size and can only be achieved with model systems consisting of the local protein environment of the residue of interest. Nevertheless, size of the model and the choice of the local environment can alter the theoretical \( pK_a \) values.

A more practical microscopic method to compute \( pK_a \) values is the hybrid quantum mechanics/molecular mechanics (QM/MM) approach, in which the titratable residue is modeled at a quantum level. At the same time, the remaining media is treated with molecular mechanics. Molecular dynamics (MD) based methods such as free energy perturbation and constant pH molecular dynamics (CPHMD) simulations can provide reliable \( pK_a \) values for protein residues. Combining enhanced sampling techniques with CPHMD simulations can also improve the accuracy of \( pK_a \) predictions. Nevertheless, the need for fast and reliable approaches to predict \( pK_a \) values of protein residues can render the microscopic methods impractical due to the exhaustive computation time.

Macroscopic methods rely on either the numerical Poisson–Boltzmann equation (PBE) or the Generalized Born (GB) technique with analytical approximations to electrostatic energies. These methods model the proteins as a homogeneous medium with a low dielectric constant while the environment (solvent) is modeled with a high dielectric constant. The PBE based methods and their variations can allow modeling the accessibility of the solvent to the titratable residues and predicting the stability of the species in the gas-phase or large conformational differences between the gas-phase and in solution.
multiple ionizable residues within the proximity. Even though there are different suggestions for the dielectric constant of proteins that varies from 4 to 80, the appropriate value depends on the polarity of the surrounding residues and the flexibility of the protein. This issue can be addressed by taking the flexibility of the protein into account via techniques that involve ensembles of conformers. An example of such an approach is the Multi-Conformation Continuum Electrostatic (MCCE) method which has been shown to successfully predict pKa values of several protein residues with different force fields.

Empirical methods are based on statistical fitting of environmental descriptors and parameters to the three-dimensional structures of proteins. Their sufficiently accurate predictions for most cases combined with their low computational cost make them widespread and favorable. There are a variety of empirical tools with comparable accuracies, but PROPKA is the most widely used for protein pKa predictions. Conceptually, PROPKA computes the change of the amino acid pKa value from water to a protein environment. In this tool, the environmental perturbation is expressed as the sum of perturbation contributions from a protein environment.

Recent studies with machine learning (ML) algorithms for pKa estimations of transition metal complexes have provided new empirical schemes. These approaches combine the pattern recognition capabilities of ML algorithms with the atomistic and molecular features that are obtained with a QM tool. However, this scheme can only be practical for proteins if molecular descriptors are obtained with low computational cost, such as neural network potentials (NNPs). Over the last decade, NNPs have been shown to provide accuracy approaching that of QM calculations and comparable computational cost with all-atom force fields. These potentials, such as ANI and AIMNet, can learn the electronic environment of an atom in conjunction with the many-body symmetry functions that arise from the coordinates. Using this learned information and combining it with the structural fingerprints that depend on the coordinates, NNPs can predict target molecular properties such as energy and forces. Thus, NNPs can be utilized to obtain information that stems from the atomic environment, and this information can be used to train ML models for protein pKa estimations.

In this context, we developed an empirical scheme for protein pKa predictions that employs ML algorithms for five amino acid types (ASP, GLU, HIS, LYS, and TYR). We rely on representation learning, i.e., learning representation of the data by automatically extracting useful information when the ML model is trained. We used ANI atomistic neural network architecture that learns molecular representation end-to-end, i.e., directly from atomic coordinates. This molecular representation reduces the dimensionality of a molecular structure into a compact vector format that encodes important quantum mechanical information.

**Methods**

Our model provides predictions via the atomic environment and the learned electronic information that is obtained with a widely used NNP, ANI-2x. The workflow for protein pKa prediction is depicted in Fig. 1. In the present work, each amino acid type is treated separately to improve the accuracy by ensuring different molecular features for different amino acid types. Models are trained and tested over hundreds of experimental pKa values, and the accuracy is also compared with the widely used PROPKA tool. The presented approach performs

![Fig. 1](image-url)
significantly better than null models and improves the current empirical methods for pK_a estimations.

Reference data for training

The pK_a model is trained and tested with two datasets. The first dataset is obtained from the PKAD database. This dataset consists of over 1500 experimentally measured pK_a values of residues on both wild type (WT) and mutant proteins. The second dataset consists of 337 entries that were extracted from the primary literature. Mutation of a residue on a protein can cause significant conformational changes that alter the amino acids’ electronic environment in proximity to the mutation site. However, not all mutant proteins have crystallographic structures deposited to the databanks. Extensive conformational sampling must be performed to account for the conformational alteration due to the mutations. Since conformational sampling is out of the scope of this study, all mutant protein entries were excluded from datasets. Our model is trained only for WT proteins. This selection results in training and test datasets containing entries from 186 WT PDB structures. The distribution of the pK_a values in training and test datasets can be found in the ESI (Fig. S.1). For this initial proof of principle model, only five titratable residues (GLU, ASP, LYS, HIS and TYR) are selected as targets for pK_a predictions.

Data curation

Crystallographic structures of 187 WT proteins are obtained from the PDB. A flowchart for data preparation prior to the training can be found in the ESI (Fig. S.2). In conventional PDB files, the crystallographic structures can involve entities other than proteins and nucleotides, such as ligands, mobile counterions, metal ions, or water molecules. It is important to state that the presence of a co-factor or a ligand can alter the pK_a of residues within a protein. However, any entry other than proteins and nucleotides is removed from PDB structures due to two reasons. First, the number of atomic species that are defined in a neural network potential (NNP) is currently limited to nonmetals. This limitation prevents inclusion of HETATM entries that can have atomic species that NNP does not define. Second, the conditions in the experimental procedures for pK_a determination and the crystallographic data preparation can be different. PDB entries correspond to constrained structures obtained using either X-ray or neutron diffractions, requiring specific strategies to achieve crystallographic packing. For example, many PDB entries tend to contain mobile counterions due to the packing procedures and these ions mainly do not exist in experimental pK_a determinations.

After the clean-up of PDB entries, missing heavy atoms and H atoms are added with the leap module of AmberTools21 using the ff14SB force field for proteins and BSCF force field for DNA. For titratable protein residues, standard protonation states are assumed. To prevent any possible steric clashes after the addition of missing atoms, very short gas-phase minimizations (250 steps of steepest descent followed by a conjugate gradient up to 500 steps in total) are performed using the sander module of AmberTools21.

Descriptor calculations

Minimized structures are used as inputs for NNP to compute all descriptors. A detailed description of ANI neural network potential and corresponding descriptors can be found elsewhere. Briefly, in ANI-type NNP, the environment of the atomic species in the given coordinate system is transformed into atomic environment vectors (AEVs) that contain radial and angular contributions (see Fig. 1). Since the pK_a of amino acids in proteins are sensitive to the neighborhood environment, naturally, AEVs were chosen as candidates for pK_a descriptors. This representation includes structural information on both bonded and non-bonded interactions of any given atom within the default ANI cutoff distance (r_cut = 5.2 Å). In addition to AEVs, neural network embeddings were chosen as learned representations. Therefore, 2nd and 3rd layers of atomic neural network embeddings are selected as additional descriptor candidates.

Feature importance and training

We observed that many features in the overall descriptor were redundant or highly correlated. To eliminate the redundant features, a three-step filtering procedure is adopted. First, noninformative features (values of 0.0) for all reference data are removed. Second, correlation of the features is computed, and highly correlated features (correlation coefficient > 0.95) are eliminated. Third, a recursive feature elimination (RFE) process is performed using a random forest regressor (RF) algorithm as implemented in the scikit-learn package. RFE is a technique that allows defining the least important features using an importance ranking, and it has been shown that ML models benefit from it. The pseudo-code for RFE is depicted in Fig. 2. In each recursive step of the procedure, the feature importance is measured, and a desired number of features are kept (F_k) by removing less important ones. The new feature list is used to perform training with RF using 1000 decision trees. A final set of features (F_k) is defined by the local model that has the best coefficient of determination for predictions over out-of-bag samples. After obtaining the final set of features, a 10-fold cross-validation (CV) is performed with RF using same settings for training in the feature elimination process.

Molecular dynamics simulations and clustering

Two different ionization states of ASP26 (neutral: ASH, and negatively charged: ASP) on human thioredoxin conformer (PDB ID: 3TRX) are considered. Topology and coordinate files are built with the default ionization states for residues in the ff14SB force field for proteins using the leap module of AmberTools21. The samples are neutralized using Na⁺ counter ions: 4Na⁺ for the sample containing neutral ASP, and 5 Na⁺ for the sample containing negatively charged ASP. To provide salt concentration, 5 Na⁺ and 5 Cl⁻ counter ions are added to the samples. Waters in the original crystal structure are deleted, and the samples are solvated using TIP3P water molecules with a distance between the solute and the edge of
the box being 12 Å, which results in an average box dimension of 66.8 Å × 69.7 Å × 62.3 Å.

Simulations are performed using the CUDA version of AMBER20’s pmemd module.\textsuperscript{128,136,137} A time step of 1.0 fs is used along with Berendsen temperature coupling\textsuperscript{138} and SHAKE algorithm\textsuperscript{139} for the bonds involving hydrogen atoms. The particle mesh Ewald summation (PME) technique\textsuperscript{140} is employed using a cutoff distance of 8 Å. We carried out an 11-step equilibration procedure\textsuperscript{141} that consists of harmonic restraints on protein residues and its reduction in each step at 10 K, which is followed by the gradual heating of samples to 300 K with a gradual harmonic restrain reduction at 300 K. A 50 ns long production simulation is performed using equilibrated samples for both samples. Production trajectories are used to cluster the frames using a hierarchical agglomerative (bottom-up) approach as implemented in the cpptraj module of AMBERTools21.\textsuperscript{128} Clustering is performed using the root mean square method as the distance metric for the carboxyl group of the ASP26 side chain (ASH26 in the case of neutral ASP). It is finalized when the minimum distance between the clusters is larger than 1.5 Å. The best cluster representatives are selected using the lowest cumulative distance to all the other frames in the same cluster.

**Results and discussion**

There has been a surge of approaches looking to learn a representation that directly encodes information about molecules.\textsuperscript{142,143} The idea behind representation learning is to learn a mapping that embeds molecular structures as points in a low-dimensional vector space.\textsuperscript{144} The goal is to optimize this mapping so that relationships in the embedding space reflect the similarities between objects. After optimizing the embedding space, the learned embeddings can be used as feature inputs for downstream machine learning tasks. The key distinction between representation learning and traditional descriptor calculations is how they treat the molecular structure problem. Descriptors treat this problem as a pre-processing step, using domain knowledge and hand-crafted rules to extract molecular information. In contrast, representation learning treats this problem as a machine learning task, using a purely data-driven approach to learn embeddings that encode a molecular structure.

The pKa of an amino acid on a protein can be affected by different environmental features such as amino acids in proximity or solvent access. The surrounding amino acids can be encoded through so-called atomic environment vectors (AEVs) which can be obtained with popular atomistic neural network potentials like ANI.\textsuperscript{145} Even though the presence of the solvent cannot be modeled with the current ANI-2x implementation, the gas-phase electronic-structure contributions can be addressed with neural network embeddings. These embeddings would provide information regarding the electronic environment of the titratable residue.

To show the utility of the representation learning, we first performed a simple exercise. We extracted 3D structures for 171 natural and non-natural amino acids from the SwissSidechain database.\textsuperscript{146} Fig. 3 shows a 2D t-distributed stochastic neighbor embedding (t-SNE)\textsuperscript{147} projection of atomic embeddings for oxygen and nitrogen atoms based on the 3rd (top) layer neural network. Naturally, oxygen and nitrogen atoms show two distinctly different clusters corresponding to each element.

Inside the oxygen cluster, titratable groups like sidechain carboxyls, aliphatic and aromatic alcohols are spread out. This is possible due to the very different environments modulated by non-natural amino acids. We hypothesized that the difference in embedding vectors should reflect the acid–base properties of these groups too. Therefore, these embedding vectors could be used as descriptors for empirical pKa prediction. For the sake of completeness, we will consider all possible descriptors, i.e., AEV, and 2nd and 3rd layer neural network embeddings obtained with the ANI-2x model as an initial set of descriptors.

To assess the performance of ML models with ANI-2x descriptors, the available pKa data are divided into training and test subsets. Different ML algorithms were tested, and the
accuracies were analyzed. Results obtained with different procedures are depicted in the ESI (see Fig. S.3).† We observed that linear regression (LR) and support vector machines (SVMs) with linear kernel yielded similar results. Training with the RF provided more accurate results with MAEs of about 0.5, while the inclusion of recursive feature elimination (RFE) improved the accuracy even further. RFE resulted in a feature space of about 10 to 100 descriptors for amino acids. We observed that the features belong to the side chains and the features that belong to the backbone atoms are selected as important descriptors. This can be related to the learned inductive (through-bond) effects. Feature elimination revealed that even though most of the descriptors from the initial feature list are eliminated, all the feature classes are preserved in the final feature list. These results indicate that pKa predictions require the information regarding the atomic environment of titratable residues and electronic information encoded by the neural network embeddings of the NNP.

First, the model accuracy was assessed with k-fold cross-validation. To compare the accuracy of our model, pKa values for the whole training dataset are also predicted with PROPKA 3.1.† The results obtained with the ML model, PROPKA, and the null model for GLU, ASP, and HIS are depicted in Fig. 4 (see ESI Fig. S.4† for LYS and TYR). It was found that the coefficients of determination ($r^2$) for all amino acid types are above 0.6 with the ML model (except for LYS, $r^2 = 0.31$) while mean absolute errors (MAEs) for all amino acid types are below 0.5 pKa units. In the case of PROPKA, predictions have $r^2 < 0.3$ and MAE > 0.6 with GLU and ASP being the most reliable predictions. Interestingly, PROPKA yields similar or less reliable results relative to the null model ($\tilde{pKa} = \bar{pKa}$), especially for HIS, LYS and TYR. These results might be due to the PROPKA computation scheme which considers the shift of the pKa value for the amino acid from water to protein ($\Delta pK_a^\text{water→protein}$), while the ML model is trained directly for pKa values in the protein environment using a relatively larger training set. The number of $N_{\text{error}} > 1.0$ is computed for all amino acid types ($N_{\text{error}} > 1.0$) for experimental pKa ($pK_a^{\text{exp}}$) values that are 1.0 unit below/above the pKa value of the corresponding amino acid in water ($pK_a^{\text{water}}$). The results are depicted in Table 1. We see that the $N_{\text{error}} > 1.0$ with the ML model is about twice smaller than with PROPKA for all amino acid types. These results indicate that ML model predictions are more reliable for all amino acid types that have a water to protein pKa shift which is at least 1.0 unit ([$|\Delta pK_a^{\text{water→protein}}| \geq 1.0$]).

The ML models were also evaluated with the external test dataset of pKa values from 33 different proteins that do not appear in the training data. Results for GLU, ASP and HIS amino acids are depicted in Fig. 5 (LYS and TYR test results can be found in ESI Fig. S.5†). We found that ML models for all amino acid types provide predictions with MAE < 1.0, where GLU and LYS yield better predictions (MAE < 0.5) relative to the other amino acids. The higher MAE values, especially in the case of ASP are related to outliers that have very high/low experimental pKa values for the corresponding amino acid (high $|\Delta pK_a^{\text{water→protein}}|$).

A similar evaluation was performed with DelPhiPKa† using the external test set. Only 23 proteins were completed due to the extended run time over one week. Calculations are performed using default runtime parameters that are provided by the DelPhiPKa program. The RMSE/MAE values for predictions of 281 pKa values with DelPhiPKa (present work) are computed as
1.03(0.76)/0.74(0.56), 1.17(0.60)/0.90(0.45), 1.38 (0.88)/0.96(0.67), 1.33 (0.49)/1.06(0.40), and 0.98 (0.87)/0.82(0.76) for ASP, GLU, HIS, LYS and TYR respectively. It should be noted that all calculations are performed sequentially on a Linux computer with the runtime of \( \frac{127}{C_24} \) s/residue for DelPhiPKa and \( \frac{0.2}{C_24} \) s per residue for the ML model presented in this work. These results indicate that the ML model not only provides more reliable results but also runs about 500 times faster.

Two test set cases are selected to investigate the underlying reason for the errors in certain predictions: GLU7 predictions and \(~0.2\) s per residue for the ML model presented in this work. These results indicate that the ML model not only provides more reliable results but also runs about 500 times faster.

Fig. 4 The accuracy of the predictions of experimental pKₐ values for (a) 10-fold cross-validation predictions with the ML model for GLU, (b) 10-fold cross-validation predictions with the ML model for ASP, (c) 10-fold cross-validation predictions with the ML model for HIS, (d) GLU using PROPKA, (e) ASP using PROPKA, (f) HIS using PROPKA, (g) GLU with null model, (h) ASP with null model, (i) HIS with null model.

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>pKₐ range</th>
<th>( N^{exp} )</th>
<th>( N^{error &gt; 1.0}_{\text{ML Model}} )</th>
<th>( N^{error &gt; 1.0}_{\text{Propka}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLU</td>
<td>pKₐ &lt; 3.5 &amp; pKₐ &gt; 5.5</td>
<td>68</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td>ASP</td>
<td>pKₐ &lt; 2.8 &amp; pKₐ &gt; 4.8</td>
<td>93</td>
<td>27</td>
<td>35</td>
</tr>
<tr>
<td>HIS</td>
<td>pKₐ &lt; 5.5 &amp; pKₐ &gt; 7.5</td>
<td>85</td>
<td>20</td>
<td>55</td>
</tr>
<tr>
<td>LYS</td>
<td>pKₐ &lt; 9.5 &amp; pKₐ &gt; 11.5</td>
<td>16</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>TYR</td>
<td>pKₐ &lt; 9.0 &amp; pKₐ &gt; 11.0</td>
<td>28</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>
for hen egg white lysozyme conformers and ASP26 predictions for recombinant human thioredoxin conformer (Fig. 6). The hen egg lysozyme white (HEWL) test set comprises seven different crystallographic structures with multiple conformer configurations for the GLU7 residue (Fig. 6a). In all HEWL conformers, there is at least one positively charged residue within 5 Å of GLU7; ARG5 in all conformers, LYS1 in every conformer except 1E8L, and Arg14 for all conformers except 1E8L, 1LSA, and 4LYT. It is observed that GLU7 in three conformers (1AKI, 1LSA, and 4LYT) is in close proximity to LYS1, promoting a H-bond interaction ($R_{	ext{side-chain}} \text{GLU7-lys1} < 3.0$ Å). In the other four HEWL conformers, there is no H-bond interaction between these residues since GLU7 is rotated to the opposite direction of the LYS1 residue. Interestingly, the prediction errors for the conformers with GLU7–LYS1 side chain interaction are lower than 1.0 while the prediction errors for the conformers that do not contain this interaction are higher than 1.0 pKₐ unit. The prediction errors for the same residue with CPHMD simulations were reported to be approximately 0.8 and 1.3 with the explicit solvent and implicit solvent respectively. These results indicate that the model is highly sensitive to the conformational states of the residues and provides similar results with CPHMD simulations.

Another test case is the ASP26 on recombinant human thioredoxin (PDB IDs: 3TRX and 4TRX). Here we see prediction errors of more than 4.0 pKₐ units for both conformers. The pKₐ of this residue is in the neutral form. Thus, the effect of different ASP26 states (charged and neutral) on thioredoxin is investigated with conformers obtained from molecular dynamics (MD) simulations. Since there is no distinctive conformational difference between two thioredoxin crystallographic structures, simulations were performed only with 3TRX. After 50 ns long MD simulations, the trajectories are clustered to find the most populated cluster and its representative (Fig. 6b). These representatives (negatively charged ASP: MD-ASP26, neutral ASP: MD-ASH26) are then used to predict the pKₐ values of ASP26. In the case of the neutral ASP residue in the MD-ASH26 conformer, the proton on the side chain is removed before the pKₐ prediction since the model is trained with negatively charged ASP. It is observed that the ASP26 conformations do not alter drastically, but the conformations of three surrounding residues (SER28, LYS39, GLU56) are affected with different ionization states of ASP. In both test set and MD-ASP26 conformers, LYS39 and GLU56 share a hydrogen bond, while this interaction does not exist in the MD-ASH26 conformer.

Additionally, the hydrogen bond interactions between ASP26 and SER28 in both test set and MD-ASP26 conformers are not observed in MD-ASH26. Instead, SER28 in MD-ASH26 forms a hydrogen bond interaction with GLU56. Predictions with the ML model reveal that the error increases with the MD-ASP26 conformer (error = 6.18) and reduces more than 1.5 units with the MD-ASH26 conformer (error = 2.53) relative to the test set conformer. These results point out the conformer sensitivity

![Fig. 6](image-url) Three-dimensional representations of (a) hen lysozyme conformers in the test set with their PDB IDs. (b) thioredoxin (PDB ID: 3TRX) conformer in the test set (gray), most populated conformer obtained after molecular dynamics simulations with protonated ASP26 (purple), and the most populated conformer obtained after molecular dynamics simulations with ASP26 (green). Prediction errors for all cases are depicted within parentheses.
of the ML model and possible discrepancies between the crystallographic and the experimental conformers that cause the prediction error.

Final ML models are trained using both the training and the test datasets following the same procedure for feature elimination and tests with 10-fold cross-validation. The accuracy of the predictions is compared with PROPKA and the null model. All results are depicted in Fig. 7. The RMSE values for all amino acid types are computed below 1.0 with ML models, while PROPKA predictions, except for ASP, yield higher RMSE values than the null models. Our model is found more accurate for GLU, ASP, HIS, and LYS residues relative to DelPhiPKa benchmarks without salt concentration. When the salt concentration is included in DelPhiPKa benchmarks, accuracies for LYS and ASP are comparable. Both benchmarks use a different dataset consisting 752 residues on 82 proteins. A similar pattern is observed for MAE values. Final ML models predict experimental $p_K_a$ values with MAEs below 0.5, while MAEs obtained with PROPKA predictions are substantially higher.

Interestingly, PROPKA have MAEs similar to or even worse than the null models. To our knowledge, the model presented in this work is the first empirical model that performs statistically significantly better than the null model for all titratable residues. Finally, the coefficient of determination for $p_K_a$ predictions with ML models is at least twice as large as that of PROPKA for all amino acid types.

Exploring the high dimensional $p_K_a$ training and test data in terms of similarity is impossible without dimensionality reduction. Thus, t-SNE is used to reduce the high dimensional data by transforming it into two-dimensional similarity maps. Such visualization allowed us to align similar residues and cross-reference them with the corresponding $p_K_a$ values. 2D t-SNE maps for GLU and HIS amino acids are given in Fig. 8 (see Fig. 5). Differences point out that the descriptors obtained from ANI-2x NNP references in the crystal and the experimental conformers. $t$-SNE maps show some outliers. An example of such cases can be seen in Fig. 9, which depicts the $t$-SNE map for ASP amino acids. For instance, ASP26 in recombinant human thioredoxin conformer in the test set (PDB ID: 3TRX) is an outlier (arrow a on Fig. 9) on the $t$-SNE map. This point is in proximity to ASP67 on the tenth type III cell adhesion module of human fibronectin (PDB ID: 1FNQ, $p_K_a = 4.2$), ASP77 on fungal elicitor (PDB ID: 1BEG, $p_K_a = 2.61$), and ASP28 on black rat cell adhesion molecule CD2 (PDB ID: 1HNG, $p_K_a = 3.57$). The experimental $p_K_a$ of ASP26 on human thioredoxin is 9.9 while its neighbors have $p_K_a$ values all below $p_K_a = 5.0$, which results in a high prediction error. The positions of residues from MD simulations (MD-ASH26: neutral ASP and MD-ASP26: negatively charged ASP) are shown with arrows b and c on Fig. 9. The $t$-SNE map shows that the MD-ASH26 conformer (arrow b) is neighboring with thioredoxin from $E. coli$ (PDB ID: 2TRX, $p_K_a = 7.5$). In contrast, the MD-ASP26 conformer (arrow c) is a neighbor to bovine ribonuclease A ASP14 (PDB ID: 3RN3, $p_K_a = 2.0$). The error of $p_K_a$ prediction increases with the MD-ASP26 conformer and decreases with the MD-ASH26 conformer. These observations point out that the descriptors obtained from ANI-2x NNP can effectively predict the $p_K_a$ of an amino acid by describing its environment. The prediction errors are closely related to the differences in the crystal and the experimental conformers.

![Fig. 7 Comparison of the final model with PROPKA and null models.](image-url)
Conclusion

The presented work demonstrates the capabilities of neural network potentials to provide pK_a descriptors for knowledge-based methods. The learned representation can be used to describe the chemical environment of amino acids in proteins. As the neural network potentials emerge as an alternative to the all-atom potentials, reliable pK_a descriptors can be obtained faster with their employment. The ML model presented in this work is the first empirical model that performs significantly better than the null model and PROPKA. The MAEs for all amino acid types are found below 0.5, again for the initial and the final models. In the case of PROPKA, the calculated RMSEs are over 1.0 except for GLU and LYS residues which are still over 0.7. The computed MAEs for PROPKA predictions (all above 0.6) show that PROPKA performs almost on par – if not worse – with the null model.

Further evaluations with an external test set not included in training data show a slight increase in RMSEs and MAEs. Among the external test set, two cases are selected to explore the principal reason for errors. The conformational differences of GLU7 on HEWL structures and their respective prediction errors indicate that the ML model is sensitive to the conformational differences. The latter case involves representative structures for ASP26 on recombinant human thioredoxin that are obtained with MD simulations (both with neutral and ionized ASP26 side chains). The pK_a predictions with these representatives confirm the conformational sensitivity of the ML model. Conceptually, a protein pK_a predictor should be sensitive to conformational alterations. Two test cases demonstrate the capability of the ML model in distinguishing different conformational states. Therefore, the errors obtained with the presented models are closely related to the conformational discrepancies between the crystal (fixed) and experimental (flexible) structures.

As with any model, the present approach has limitations. Some of them, such as the absence of Cys and Ser, can be overcome by adding more training data, and mining pK_a values 2x NNP. The quantum mechanical information, which depends on the local chemical environment, is obtained from the top layers of neural network embeddings. These descriptors are used for training with the RF model to predict pK_a values. It is found that the adoption of RFE slightly improves the accuracy and yields the number of features ranging from 25 to 100 in the final model.

The accuracy of the pK_a estimations is accessed via 10-fold CV, and the results are compared with the null models and PROPKA predictions. It is found that the model presented in this work performs better than the null model and PROPKA. The RMSE of the pK_a predictions is below 0.7 except for HIS (0.72) with both the initial and the final models. The MAEs for all amino acid types are found below 0.5, again for the initial and the final models. In the case of PROPKA, the calculated RMSEs are over 1.0 except for GLU and LYS residues which are still over 0.7. The computed MAEs for PROPKA predictions (all above 0.6) show that PROPKA performs almost on par – if not worse – with the null model.

Fig. 8 t-Distributed stochastic neighbor embedding (t-SNE) maps depicting the similarity of descriptors after recursive feature elimination for (a) GLU residues, (b) HIS residues. Each data point is colored using the color code corresponding to the experimental pK_a values.

Fig. 9 t-Distributed stochastic neighbor embedding (t-SNE) maps depicting the similarity of descriptors after recursive feature elimination for ASP residues. Conformers for 3TRX (test set conformer, conformer obtained from MD simulations with negatively charged ASP26, and neutral ASH26 side chains) are shown with arrows. Each data point is colored using the color code corresponding to the experimental pK_a values.
from the primary literature. Future work will aim to extend the present model for coenzyme and cofactor effects. The current ANI descriptor has only biogenic elements and has not parametrized for metals, therefore all HETATM entries in PDB files are ignored. There is a set of limitations that would require the development of a new approach, for instance inclusion of the ionic strength or different solvents into NN descriptors.

Data availability

The code and ML models are available at https://github.com/isayevlab/pKa-ANI.

Author contributions

O. I. conceived the idea. H. G. carried out method implementation and performed all calculations. All authors critically contributed to the design of the project, analysis of results, and writing of the manuscript. O. I. supervised and acquired funding for the project.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors acknowledge Dr Adrian Roitberg for his invaluable insights and discussions. We acknowledge support from NSF CHE-2041108. This publication resulted in part from research supported by the Office of Naval Research (ONR) through the Energetic Materials Program (MURI grant no. N00014-21-1-2476). We also acknowledge the Extreme Science and Engineering Discovery Environment (XSEDE) award CHE200122, which is supported by NSF grant number ACI-1053575. This research is part of the Frontera computing project at the Texas Advanced Computing Center. Frontera is made possible by the National Science Foundation award OAC-1818253. We gratefully acknowledge the support and hardware donation from NVIDIA Corporation and express our special gratitude to Jonathan Lefman.

References

143 X. Li and D. Fournes, J. Cheminf., 2020, 12, 27.