Prediction of Protein pKa with Representation Learning

Hatice Gokcan, Olexandr Isayev*

Department of Chemistry, Mellon College of Science,

Carnegie Mellon University, Pittsburgh, PA

[*] Correspondence: olexandr@olexandrisayev.com

Abstract

The behavior of proteins is closely related to the protonation states of the residues. Therefore, prediction and measurement of pKa are essential to understand the basic functions of proteins. In this work, we develop a new empirical scheme for protein pKa prediction that is based on deep representation learning. It combines machine learning with atomic environment vector (AEV) and learned quantum mechanical representation from ANI-2x neural network potential (J. Chem. Theory Comput. 2020, 16, 4192). The scheme requires only the coordinate information of a protein as the input and separately estimates the pKa for all five titratable amino acid types. The accuracy of the approach was analyzed with both cross-validation and an external test set of proteins. Obtained results were compared with the widely used empirical approach PROPKA. The new empirical model provides accuracy with MAEs below 0.5 for all amino acid types. It surpasses the accuracy of PROPKA and performs significantly better than the null model. Our model is also sensitive to the local conformational changes and molecular interactions.

INTRODUCTION

Basic features and the behavior of proteins such as folding or ligand binding, heavily depend on the environmental conditions like local protein environment. Titratable amino acids like aspartic acid (Asp) or histidine (His) are essential in many biological processes[1–5] and can be either protonated or deprotonated depending on the local environment. Thus, the decision of the ionization states via pKa predictions is a prerequisite to understand the protein function. Determination of pKa values via experimental procedures is challenging and the most reliable results for proteins can be obtained only with NMR titrations[6]. This predicament enforces the pKa predictions in proteins by means of theoretical applications[7]. There is a tremendous amount of work on theoretical pKa calculations in the literature. These approaches can be classified into three categories as i) microscopic methods[8,9], ii) macroscopic methods which establish continuum electrostatics[10], and iii) knowledge-based methods that rely on empirical parameters[11,12].

Among three classes of theoretical pKa calculations, microscopic methods such as quantum mechanical (QM) approaches are considered the most reliable ones to compute pKa values of small molecules[13]. The most traditional approach with QM methods is to employ thermodynamic cycles by computing protonation/deprotonation free energies in gas-phase and in solution[14–23]. However, these calculations do not always provide reliable pKa values due to reasons such as the instability of the species in gas-phase or large conformational differences between gas-phase, and in solution[17,24] In the case of the proteins, QM approaches are impractical simply due to the system size and can only be achieved with model systems consist of local protein environment of the residue of interest. Nevertheless, the size of the model and the choice of the local environment can alter the theoretical pKa values[25]. A more practical microscopic method to compute pKa 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[26–28]. Molecular dynamics (MD) based methods such as free energy perturbation[29,30] and constant pH molecular dynamics (CPHMD) simulations[31–41] can provide reliable pKa values for protein residues. Combining enhance sampling techniques with CPHMD simulations can also improve the accuracy of pKa predictions[34,42–47]. Nevertheless, the need of fast and reliable approaches to predict pKa values of protein residues can render the microscopic methods impractical due to the exhaustive computation time.

The macroscopic methods can rely on either numerical Poisson−Boltzmann equation (PBE)[10,48–51] or Generalized Born (GB) technique with analytical approximations to electrostatic energies[52,53]. These methods model the proteins as homogeneous medium with a low dielectric constant while

the environment (solvent) is modeled with a high dielectric constant. The Poisson−Boltzmann equation (PBE) based methods and their variations[54–60]. PBE methods allow modeling the accessibility of the solvent to the titratable residues[61,62] and multiple ionizable residues within the proximity[63,64]. Even though there are different suggestions for the dielectric constant of proteins that varies from 4 to 80[65–73], the appropriate value depends on the polarity of the surrounding residues and the flexibility of the protein [74,75]. This issue can be addressed by taking the flexibility of the protein into account via techniques that involve ensembles of conformers[54,76–82]. 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[70,83–85].

Empirical methods are based on statistical fitting of environmental descriptors and parameters to the three-dimensional structures of proteins. Even though they do not provide physical insights into the molecular determinants of the pKa values, 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[86–88], but PROPKA[11,12] 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 provide new empirical schemes[89,90]. 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 of QM calculations and comparable computational cost with all-atom force fields. These potentials, such as ANI[91–98] and AIMNet[99], can learn the electronic environment of an atom in conjunction with the many-body symmetry functions that arise from the coordinates[100,101]. 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 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 are obtained with a widely used NNP, ANI-2x[96]. The workflow for protein pKa prediction is depicted in Figure 1. In the present work, each amino acid type is treated separately to improve 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[12] tool. The presented approach performs significantly better than null models and improves the current empirical methods for pKa estimations.
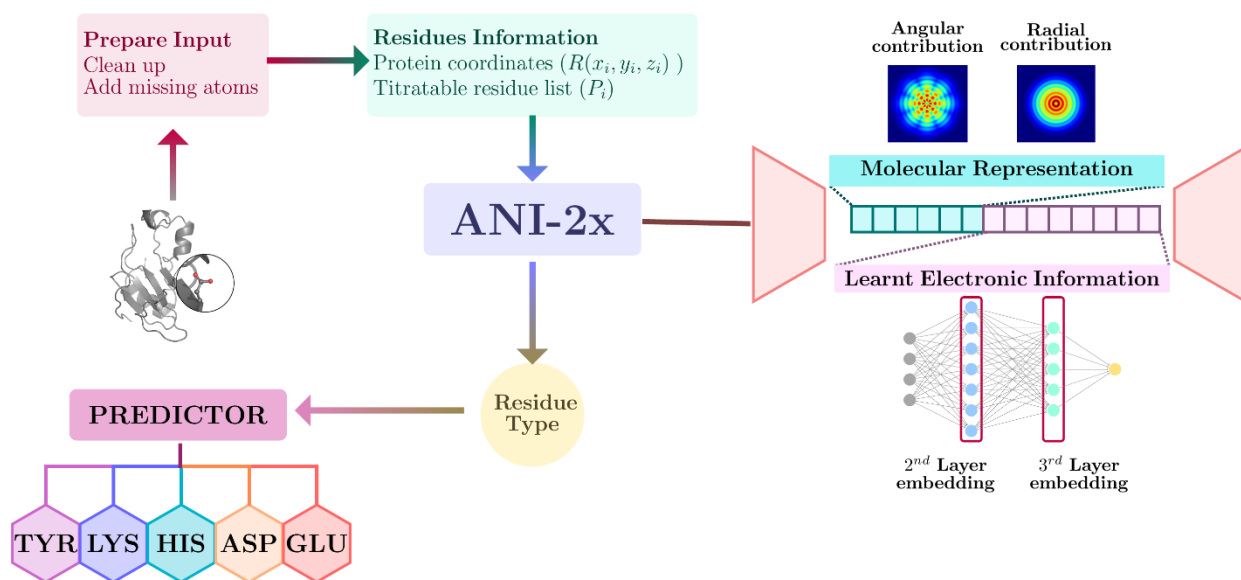

Figure 1. Protein pKa prediction with neural network features obtained with ANI-2x. Each amino acid type has its own predictor.

Reference data for training

The pKa model is trained and tested with two datasets. The first dataset is obtained from PKAD database [102]. This dataset consists of over 1500 experimentally measured pKa values of residues on both wild type (WT) and mutant proteins. The second dataset consists of 337 entries that were extracted from primary literature[103–127]. 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 Supporting Information (Figure 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 Supporting Information (Figure S.2). In conventional PDB files, the crystallographic structures can involve entries 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 pKa 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 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 *tleap* module of AmberTools21[128] using ff14SB force field for proteins[129] and BSC1 force field for DNA[130]. 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 *sander* module of AmberTools21[128].

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 in elsewhere [96,100]. Briefly, In ANI-type NNPs, the environment of the atomic species in the given coordinate system is transformed to atomic environment vectors (AEVs) that contain radial and angular contributions. 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. Additionally, different neural network embeddings were chosen as learned representations. Therefore, 2nd and 3rd layers of atomic neural network embeddings are selected as additional descriptor candidates. In both cases, AEVs and NN embeddings descriptors correspond only to the heavy atoms of titratable residues.

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)[131] process is performed using a random forest regressor (RF)[132] algorithm as implemented in the scikit-learn package[133]. RFE is a technique that allows defining least important features using an importance ranking, and it has been shown that ML models benefit from it[134]. The pseudo-code for RFE is depicted in Figure 2. In each recursive step of the procedure, the feature importance is measured, and a desired number of features are kept ($F^{\dagger}$) by removing less important ones. The new feature list is used to perform training with RFR using 1000 decision trees. A final set of features ($F^{\ddagger}$) 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 RFR using same settings for training in feature elimination process.

```
Input   :
            Training set (pKa)
            Set of features (F = {f1, · · · , f_M})
            Number of features to select (N = {M, · · · , 10, 5})
Output:
            Set of features providing highest accuracy → F‡
for each n in N, do
    Perform RFE and select n features → F†
    Train with RF using F†
    Compute accuracy of model with out of bag predictions → r²_{Ni}
    if r²_{Ni} > r²_{Ni−1} then
        F‡ ← F†
    end
    F = F†
end
```

Figure 2. Pseudo-code for feature selection with recursive feature elimination (RFE).

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 ff14SB force field for proteins[129] using *tleap* module of AmberTools21[128]. 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[135] with a distance between solute and the edge of the box as 12Å, which results in an average box dimension of 66.8 Å x 69.7 Å x 62.3 Å.

Simulations are performed using CUDA version of AMBER20's *pmemd* module[128,136,137]. A time step of 1.0fs is used along with Berendsen temperature coupling [138] and SHAKE algorithm[139] for the bonds involving hydrogen atoms. The particle mesh Ewald summation (PME) technique[140] is employed using a cutoff distance of 8 Å. An 11-step equilibration procedure[141] that consists of harmonic restraints on protein residues and its reduction in each step at 10K, which is followed by the gradual heating of samples to 300K with a gradual harmonic restrain reduction at 300K. A 100ns 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 *cpptraj* module of AMBERTools21 [128]. Clustering is performed using root mean square method as the distance metric for carboxyl group of ASP26 side chain (ASH26 in 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[142,143]. The idea behind representation learning is to learn a mapping that embeds molecular structures as points in a low-dimensional vector space[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 molecular structure.

The pK_a 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 [96]. 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 SwissSidechain database[145]. Figure 3 shows a 2D t-Distributed stochastic neighbor embedding (t-SNE)[146] projection of atomic embeddings for oxygen and nitrogen atoms based on 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 ANI-2x model as an initial set of descriptors.
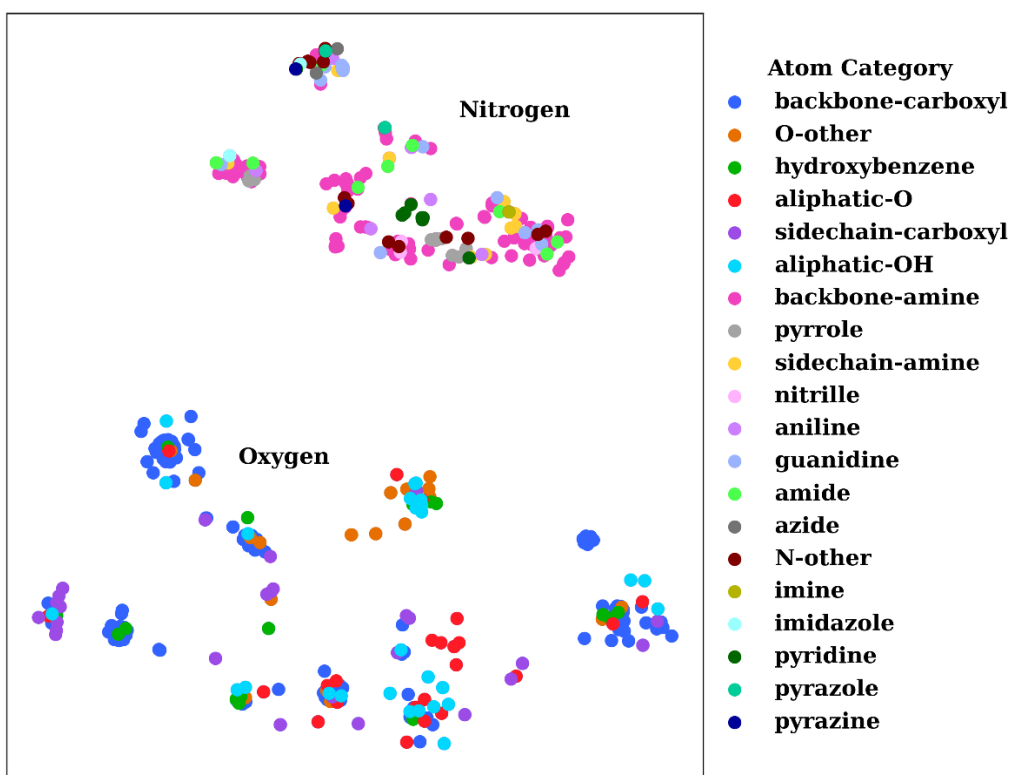

Figure 3. t-Distributed stochastic neighbor embedding (t-SNE) maps depicting similarity of 3rd layer neural network embeddings for oxygen and nitrogen atoms located on structures from SwissSidechain database [145]. The backbones of the corresponding structures ensure zwitterion form with NH_3^+ (backbone-amine) and COO^- (backbone-carboxyl) as backbone groups.

To assess the performance of ML models with ANI-2x descriptors, the available pK$_a$ data is divided into training and test subsets. Different ML algorithms were tested, and the accuracies were analyzed. Results obtained with different procedures are depicted in supporting information (see Figure 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 reliable results with MAEs 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 belong to the backbone atoms are selected as important descriptors. That 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 pK_a predictions require the information regarding the atomic environment of titratable residue and electronic information encoded by the neural network embeddings of the NNP.

First, the model accuracy was accessed with K-fold cross-validation. To compare the accuracy of our model, pK_a values for the whole training dataset are also predicted with Propka 3.1[12]. The results obtained with the ML model, Propka, and the null model for GLU, ASP, and HIS are depicted in Figure 4 (see Supporting Information Figure S.4 for LYS and TYR). It was found that the coefficient of determinations (r^2) for all amino acid types are above 0.6 with ML model (except for LYS, $r^2 = 0.31$) while mean absolute error (MAE) for all amino acid types are below 0.5 pKa units. In 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 ($\widehat{pK_a} = \overline{pK_a}$), 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 ($\Delta pK_a^{water \to protein}$) [11], while the ML model is trained directly for pKa values in protein environment using a relatively larger training set. The number of $\widehat{pK_a}^{error} > 1.0$ is computed for all amino acid types ($N_{error>1.0}$) for experimental pK_a (pK_a^{exp}) values that are 1.0 unit below/above than the pK_a value of the corresponding amino acid in water (pK_a^{water}). The results are depicted in Table 1. We see that the $N_{error>1.0}$ with ML model are 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 has a water to protein pK_a shift that is at least 1.0 unit ($\left| \Delta pK_a^{water \to protein} \right| \geq 1.0$).

Table 1. Number of experimental pKa values that are 1.0 pKa unit lower or larger than the pKa in water (N^{exp}) and the number of prediction errors that are above 1.0 pKa unit ($N_{error>1.0}$).

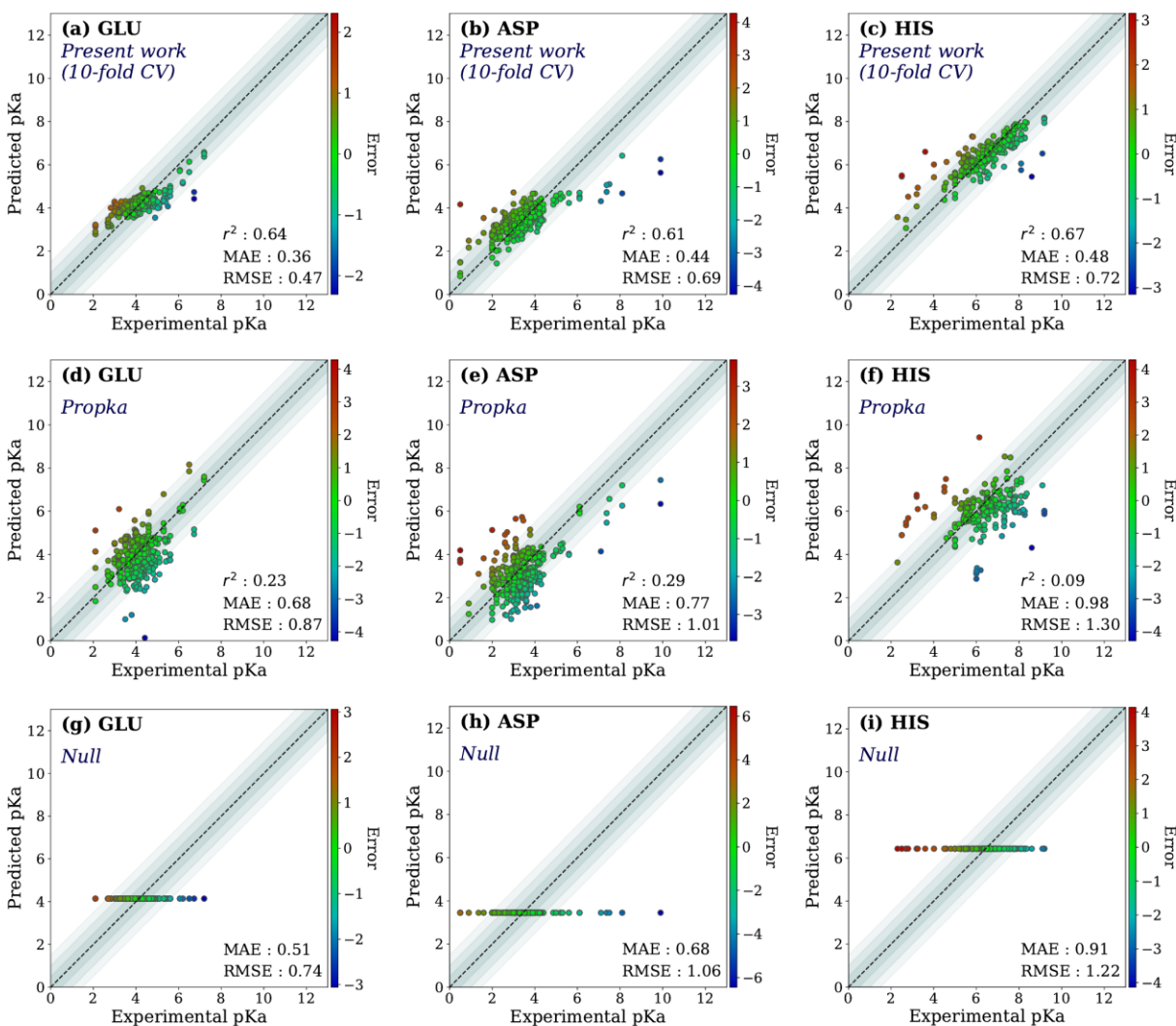| Amino acid	pK$_a$ range	N^{exp}	$N_{error>1.0}^{ML\,Model}$	$N_{error>1.0}^{Propka}$
GLU	pK$_a$<3.5 & pK$_a$>5.5	68	12	21
ASP	pK$_a$<2.8 & pK$_a$>4.8	93	27	35
HIS	pK$_a$<5.5 & pK$_a$>7.5	85	20	55
LYS	pK$_a$<9.5 & pK$_a$>11.5	16	7	8
TYR	pK$_a$<9.0 & pK$_a$>11.0	28	0	8

Figure 4. The accuracy of the predictions of experimental pKa values for a) 10-fold cross validation predictions with ML model for GLU, b) 10-fold cross validation predictions with ML model for ASP, c) 10-fold cross validation predictions with 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.

The ML models were also evaluated with the external test dataset of pK$_a$ values from 33 different proteins that do not appear in the training data. Results for GLU, ASP and HIS amino acids are depicted in Figure 5 (LYS and TYR test results can be found in SI Figure 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 pK$_a$ values for the corresponding amino acid (high $\left|\Delta pK_a^{water \rightarrow protein}\right|$).
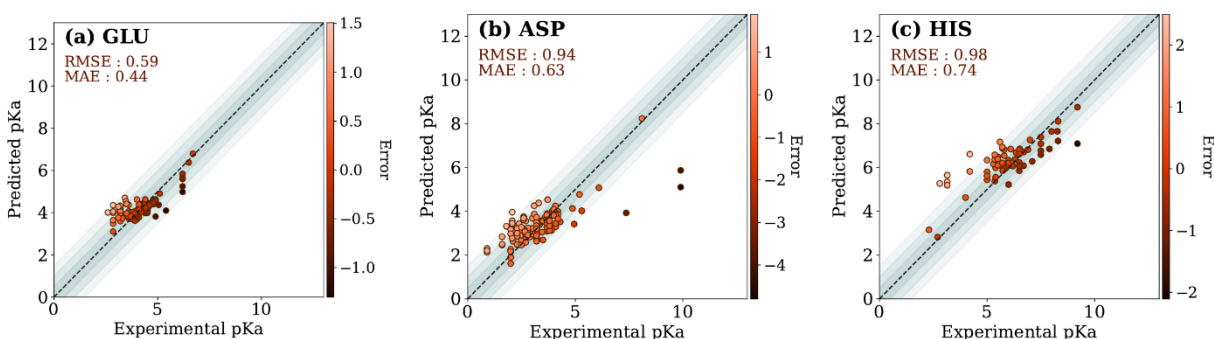

Figure 5. Test set predictions with ML models trained with descriptors obtained with ANI-2x.

Two test set cases are selected to investigate the underlying reason for the errors in certain predictions: GLU7 predictions for hen egg white lysozyme conformers and ASP26 predictions for recombinant human thioredoxin conformer (Figure 6). The hen egg lysozyme white (HEWL) test set comprises seven different crystallographic structures with multiple conformer configurations for GLU7 residue (Figure 6a). In all HEWL conformers, there is at least one positively charged residue within 5Å 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 ($R_{GLU7-LYS1}^{side\,chain} < 3.0$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 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 pK$_a$ unit. The prediction errors for the same residue with CPHMD simulations were reported approximately 0.8 and 1.3 with explicit and implicit solvent respectively[45]. 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 pK$_a$ units for both conformers. The pKa of this residue is reported as 9.9, which indicates that this residue is in 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 100ns long MD simulations, the trajectories are clustered to find the most populated cluster and its representative (Figure 6b). These representatives (negatively charged ASP: MD-ASP26, neutral ASP: MD-ASH26) are then used to predict the pK_a values of ASP26. In the case of neutral ASP residue in MD-ASH26 conformer, the proton on the side chain is removed before the pK_a prediction since the model is trained with negatively charged ASP. It is observed that ASP26 confirmation does 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 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 MD-ASP26 conformer (error=6.18) and reduces more than 1.5 units with MD-ASH26 conformer (error=2.53) relative to the test set conformer. These results point out the conformer sensitivity of the ML model and possible discrepancies between the crystallographic and the experimental conformers that causes the prediction error.
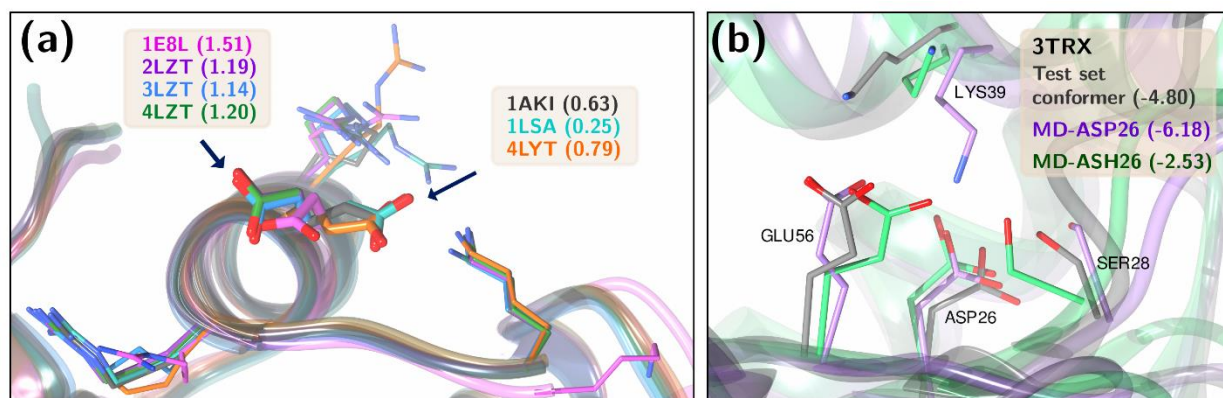

Figure 6. Three-dimensional representations of **a)** hen lysozyme conformers in test set with their PDB IDs. **b)** thioredoxin (PDB ID: 3TRX) conformer in test set (gray), most populated conformer obtained after molecular dynamics simulations with protonated ASP26 (purple), and most populated conformer obtained after molecular dynamics simulations with ASP26 (green). Prediction errors for all cases are depicted within parentheses.

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 Figure 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. A similar pattern is observed for MAE values. Final ML models predict experimental pKa values with MAEs below 0.5, while MAEs obtained with Propka predictions are substantially higher.

Interestingly, Propka pKa predictors 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 pKa predictions with ML models is at least twice as large as that of Propka for all amino acid types.
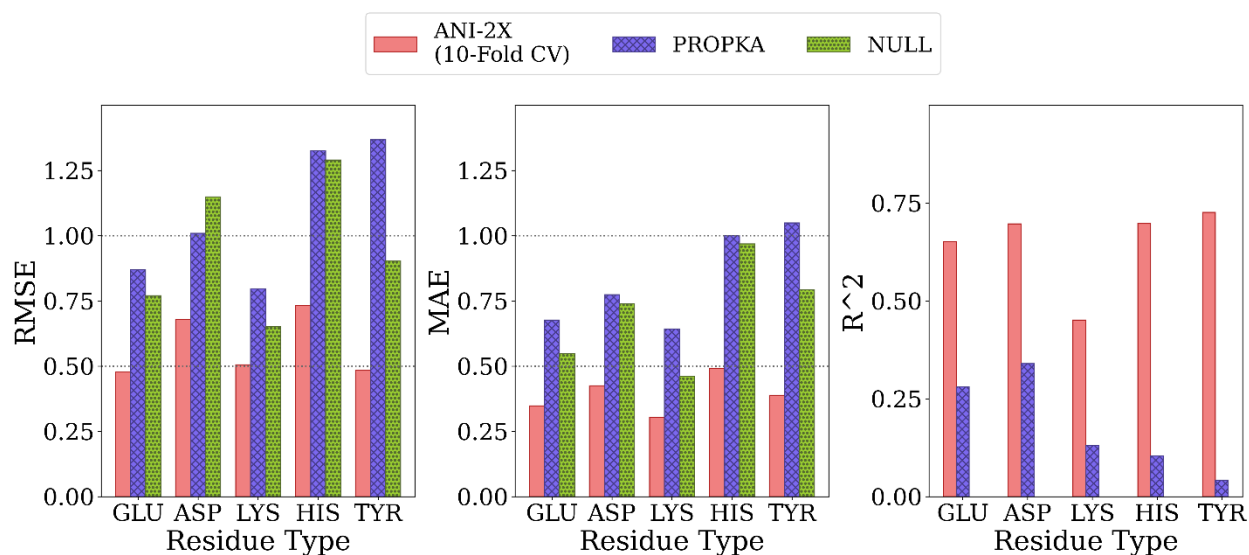

Figure 7. Comparison of final model with propka and null models

Exploring the high dimensional pKa training and test data in terms of similarity is impossible without dimensionality reduction. Thus, t-SNE [146] 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 by the corresponding pKa values. 2D t-SNE maps for GLU and HIS amino acids are given in Figure 8 (see Figure S.6 for LYS and TYR amino acids). Generally, residues with high or low experimental pKa values are well separated except for some outliers, and residues on the same class of proteins form small clusters together. For instance, GLU7 from hen egg-white lysozyme (HEWL) and turkey egg-white lysozyme (TEWL) form clusters a_i (Figure 8a). Among these clusters a_5 involves entries from both species (TEWL PDB IDs: 1LZ3, 135L and HEWL PDB IDs: 1LSA, 1LSE, 1LYS). Clusters are shown with b_i on Figure

8a corresponds to the GLU35 residues on HEWL and TEWL proteins. Other examples of such clusters correspond to GLU2 residues on bovine Ribonuclease A (cluster c, Figure8a), and GLU73 residue on Barnase (clusters d_i, Figure 8a). A similar pattern is observed with HIS amino acid (Figure 8b). Residues in the same class of proteins form small clusters such as cluster $\boldsymbol{a}$ for GLU162 on Bacillus agaradhaerens family 11 xylanase, cluster $\boldsymbol{b}$ for HIS36 on myoglobin from sperm whale and horse, and clusters $\boldsymbol{c_i}$ for HIS72 on bovine tyrosine phosphatase.
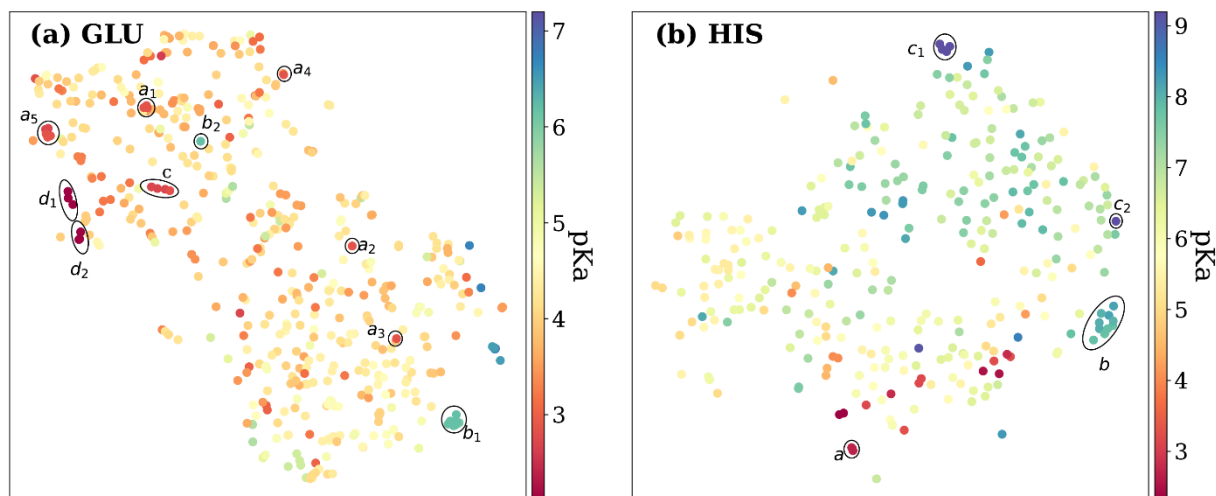

Figure 8. t-Distributed stochastic neighbor embedding (t-SNE) maps depicting similarity of descriptors after recursive feature elimination for (a) GLU residues, (b) HIS residues. Each data point is colored by the corresponding experimental pKa values.

As mentioned before, pKa models are sensitive to conformers, and t-SNE maps show some outliers. An example of such cases can be seen in Figure 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 $\boldsymbol{a}$ on Figure 9) on t-SNE map. This point is in proximity to ASP67 on the tenth type III cell adhesion module of human fibronectin (PDB ID: 1FNA, pK$_a$=4.2), ASP77 on fungal elictor (PDB ID: 1BEG, pK$_a$=2.61), and ASP28 on black rat cell adhesion molecule CD2 (PDB ID: 1HNG, pK$_a$=3.57). The experimental pK$_a$ of ASP26 on human thioredoxin is 9.9 while its neighbors have pK$_a$ values all below pK$_a$=5.0, which results in high prediction error. The positions of residues from MD simulations (MD-ASH26: neutral ASP and MD-ASP26: negatively charged ASP) are shown with arrows $\boldsymbol{b}$ and $\boldsymbol{c}$ on Figure 9. The t-SNE map shows that MD-ASH26 conformer (arrow $\boldsymbol{b}$) is neighboring with thioredoxin from *E.coli* (PDB ID: 2TRX, pKa=7.5). In contrast, MD-ASP26 conformer (arrow $\boldsymbol{c}$) is neighbor to bovine ribonuclease A ASP14 (PDB ID: 3RN3, pKa=2.0). The error of pK$_a$ prediction increases with MD-ASP26

conformer and decreases with MD-ASH26 conformer. These observations point out that the descriptors obtained from ANI-2x NNP can effectively predict pKa of an amino acid by describing its environment. The prediction errors are closely related to the differences in crystal and the experimental conformers.
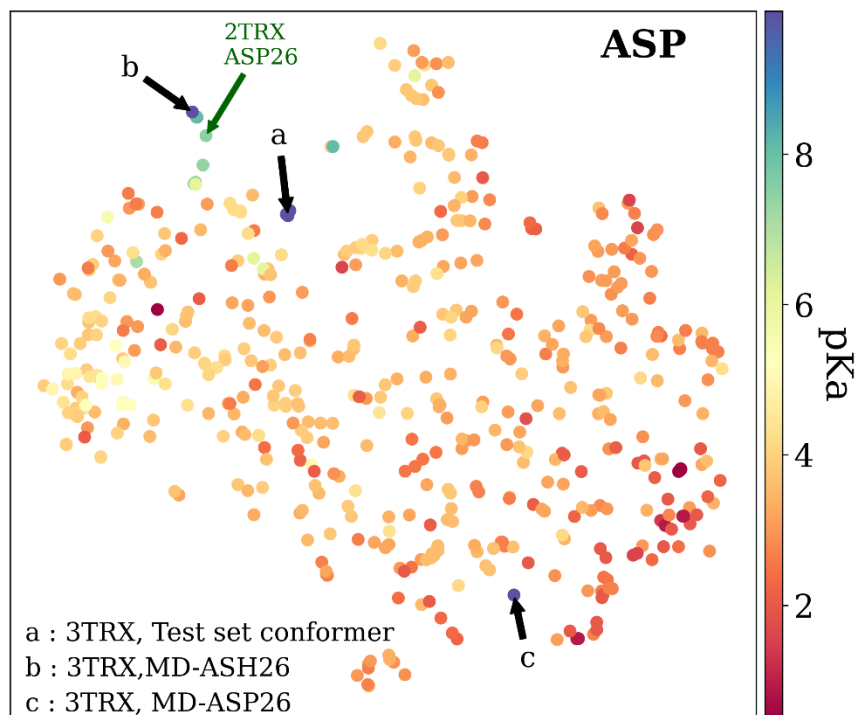

Figure 9. t-Distributed stochastic neighbor embedding (t-SNE) maps depicting similarity of descriptors after recursive feature elimination for ASP residues. Conformers for 3TRX (Test set conformer, conformer obtained from MD simulations with negatively charged ASP26, and neutral ASH26 side chains) are shown with arrows. Each data point is colored by the corresponding experimental pKa values.

CONCLUSION

The presented work demonstrates the capabilities of neural network potentials to provide pKa 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 pKa descriptors can be obtained faster with their employment. ML models presented in this work is the first empirical model that performs significantly better than the null model for all titratable residues.

A new empirical scheme for pKa prediction of amino acids in proteins uses an ML model with descriptors calculated on ANI-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 pKa values. It is found that the adoption of RFE slightly improves the accuracy and yields to the number of features ranging from 25 to 100 in the final model.

The accuracy of the pKa 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 pKa 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.

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 chain). The pKa predictions with these representatives confirm the conformational sensitivity of the ML model. Conceptually, a protein pKa 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.

ACKNOWLEDGEMENTS

REFERENCES

1	A. Warshel, P. K. Sharma, M. Kato and W. W. Parson, *Biochim. Biophys. Acta - Proteins Proteomics*, 2006, 1764, 1647–1676.

2	M. Watari, T. Ikuta, D. Yamada, W. Shihoya, K. Yoshida, S. P. Tsunoda, O. Nureki and H. Kandori, *J. Biol. Chem.*, 2019, **294**, 3432–3443.

3	A. M. Smondyrev and G. A. Voth, *Biophys. J.*, 2002, **83**, 1987–1996.

4	H. Luecke, B. Schobert, J. Stagno, E. S. Imasheva, J. M. Wang, S. P. Balashov and J. K. Lanyi, *Proc. Natl. Acad. Sci.*, 2008, **105**, 16561–16565.

5	N. P. Le, H. Omote, Y. Wada, M. K. Al-Shawi, R. K. Nakamoto and M. Futai, *Biochemistry*, 2000, **39**, 2778–2783.

6	Z. P. Haslak, S. Zareb, I. Dogan, V. Aviyente and G. Monard, *J. Chem. Inf. Model.*, 2021, **61**, 2733–2743.

7	P. G. Seybold and G. C. Shields, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2015, **5**, 290–297.

8	S. Sastre, R. Casasnovas, F. Muñoz and J. Frau, *Theor. Chem. Acc.*, 2013, **132**, 1310.

9	D. Riccardi, P. Schaefer and Q. Cui, *J. Phys. Chem. B*, 2005, **109**, 17715–17733.

10	C. Li, Z. Jia, A. Chakravorty, S. Pahari, Y. Peng, S. Basu, M. Koirala, S. K. Panday, M. Petukh, L. Li and E. Alexov, *J. Comput. Chem.*, 2019, **40**, 2502–2508.

11	M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski and J. H. Jensen, *J. Chem. Theory Comput.*, 2011, **7**, 525–537.

12	C. R. Søndergaard, M. H. M. Olsson, M. Rostkowski and J. H. Jensen, *J. Chem. Theory Comput.*, 2011, **7**, 2284–2295.

13	L. Zanetti-Polzi, I. Daidone and A. Amadei, *J. Phys. Chem. B*, 2020, **124**, 4712–4722.

14	M. Abul Kashem Liton, M. Idrish Ali and M. Tanvir Hossain, *Comput. Theor. Chem.*, 2012, **999**, 1–6.

15   M. Namazian, M. Zakery, M. R. Noorbala and M. L. Coote, *Chem. Phys. Lett.*, 2008, **451**, 163–168.

16   M. D. Liptak, K. C. Gross, P. G. Seybold, S. Feldgus and G. C. Shields, *J. Am. Chem. Soc.*, 2002, **124**, 6421–6427.

17   J. F. Satchell and B. J. Smith, *Phys. Chem. Chem. Phys.*, 2002, **4**, 4314–4318.

18   K. C. Gross, P. G. Seybold, Z. Peralta-Inga, J. S. Murray and P. Politzer, *J. Org. Chem.*, 2001, **66**, 6919–6925.

19   K. C. Gross, P. G. Seybold and C. M. Hadad, *Int. J. Quantum Chem.*, 2002, **90**, 445–458.

20   M. D. Liptak and G. C. Shields, *J. Am. Chem. Soc.*, 2001, **123**, 7314–7319.

21   A. M. Toth, M. D. Liptak, D. L. Phillips and G. C. Shields, *J. Chem. Phys.*, 2001, **114**, 4595–4606.

22   I. E. Charif, S. M. Mekelleche, D. Villemin and N. Mora-Diez, *J. Mol. Struct. THEOCHEM*, 2007, **818**, 1–6.

23   D. Gao, P. Svoronos, P. K. Wong, D. Maddalena, J. Hwang and H. Walker, *J. Phys. Chem. A*, 2005, **109**, 10776–10785.

24   R. Casasnovas, J. Ortega-Castro, J. Frau, J. Donoso and F. Muñoz, *Int. J. Quantum Chem.*, 2014, **114**, 1350–1363.

25   H. Li, A. D. Robertson and J. H. Jensen, *Proteins Struct. Funct. Bioinforma.*, 2004, **55**, 689–704.

26   H. Li, A. W. Hains, J. E. Everts, A. D. Robertson and J. H. Jensen, *J. Phys. Chem. B*, 2002, **106**, 3486–3494.

27   J. H. Jensen, H. Li, A. D. Robertson and P. A. Molina, *J. Phys. Chem. A*, 2005, **109**, 6634–6643.

28   S. C. L. Kamerlin, M. Haranczyk and A. Warshel, *J. Phys. Chem. B*, 2009, **113**, 1253–1272.

29   S. N. Sakipov, J. C. Flores-Canales and M. G. Kurnikova, *J. Phys. Chem. B*, 2019, **123**, 5024–5034.

30   H. Yu, I. M. Ratheal, P. Artigas and B. Roux, *Nat. Struct. Mol. Biol.*, 2011, **18**, 1159–1163.

31   J. Mongan, D. A. Case and J. A. McCammon, *J. Comput. Chem.*, 2004, **25**, 2038–2048.

32   E. J. Arthur, J. D. Yesselman and C. L. Brooks, *Proteins Struct. Funct. Bioinforma.*, 2011, **79**, 3276–3286.

33   Y. Meng and A. E. Roitberg, *J. Chem. Theory Comput.*, 2010, **6**, 1401–1412.

34   J. M. Swails and A. E. Roitberg, *J. Chem. Theory Comput.*, 2012, **8**, 4393–4404.

35   G. B. Goh, B. S. Hulbert, H. Zhou and C. L. Brooks, *Proteins Struct. Funct. Bioinforma.*, 2014, **82**, 1319–1331.

36   J. Khandogin and C. L. Brooks, *Biophys. J.*, 2005, **89**, 141–157.

37   A. M. Baptista, V. H. Teixeira and C. M. Soares, *J. Chem. Phys.*, 2002, **117**, 4184–4200.

38   R. Bürgi, P. A. Kollman and W. F. van Gunsteren, *Proteins Struct. Funct. Bioinforma.*, 2002, **47**, 469–480.

39   M. S. Lee, F. R. Salsbury and C. L. Brooks, *Proteins Struct. Funct. Bioinforma.*, 2004, **56**, 738–752.

40   J. A. Wallace and J. K. Shen, *J. Chem. Theory Comput.*, 2011, **7**, 2617–2629.

41   J. Khandogin and C. L. Brooks, *Biochemistry*, 2006, **45**, 9363–9373.

42   S. L. Williams, C. A. F. De Oliveira and J. Andrew McCammon, *J. Chem. Theory Comput.*, 2010, **6**, 560–568.

43   Y. Meng, D. Sabri Dashti and A. E. Roitberg, *J. Chem. Theory Comput.*, 2011, **7**, 2721–2727.

44   J. Lee, B. T. Miller, A. Damjanović and B. R. Brooks, *J. Chem. Theory Comput.*, 2014, **10**, 2738–2750.

45   J. M. Swails, D. M. York and A. E. Roitberg, *J. Chem. Theory Comput.*, 2014, **10**, 1341–1352.

46   F. L. Barroso daSilva and L. G. Dias, *Biophys. Rev.*, 2017, **9**, 699–728.

47   J. Liu, J. Swails, J. Z. H. Zhang, X. He and A. E. Roitberg, *J. Am. Chem. Soc.*, 2018, **140**, 1639–1648.

48   W. Rocchia, E. Alexov and B. Honig, *J. Phys. Chem. B*, 2001, **105**, 6507–6514.

49   M. Holst, N. Baker and F. Wang, *J. Comput. Chem.*, 2000, **21**, 1319–1342.

50   S. Jo, M. Vargyas, J. Vasko-Szedlar, B. Roux and W. Im, *Nucleic Acids Res.*, 2008, **36**, W270–W275.

51   B. Lu, X. Cheng, J. Huang and J. A. McCammon, *J. Chem. Theory Comput.*, 2009, **5**, 1692–1699.

52   M. Feig and C. L. Brooks, *Curr. Opin. Struct. Biol.*, 2004, **14**, 217–224.

53   M. Feig, A. Onufriev, M. S. Lee, W. Im, D. A. Case and C. L. Brooks, *J. Comput. Chem.*, 2004, **25**, 265–284.

54   J. Warwicker and H. C. Watson, *J. Mol. Biol.*, 1982, **157**, 671–679.

55   M. K. Gilson, A. Rashin, R. Fine and B. Honig, *J. Mol. Biol.*, 1985, **184**, 503–516.

56   N. A. Baker, *Curr. Opin. Struct. Biol.*, 2005, **15**, 137–143.

57   D. Bashford and M. Karplus, *Biochemistry*, 1990, **29**, 10219–10225.

58   M. J. Potter, M. K. Gilson and J. A. McCammon, *J. Am. Chem. Soc.*, 1994, **116**, 10298–10299.

59   T. J. Dolinsky, J. E. Nielsen, J. A. McCammon and N. A. Baker, *Nucleic Acids Res.*, 2004, **32**, W665–W667.

60   V. H. Teixeira, C. A. Cunha, M. Machuqueiro, A. S. F. Oliveira, B. L. Victor, C. M. Soares and A. M. Baptista, *J. Phys. Chem. B*, 2005, **109**, 14691–14706.

61   J. A. Reynolds, D. B. Gilbert and C. Tanford, *Proc. Natl. Acad. Sci.*, 1974, **71**, 2925–2927.

62   J. J. Havranek and P. B. Harbury, *Proc. Natl. Acad. Sci.*, 1999, **96**, 11145–11150.

63   M. K. Gilson, *Proteins Struct. Funct. Bioinforma.*, 1993, **15**, 266–282.

64   C. Lim, D. Bashford and M. Karplus, *J. Phys. Chem.*, 1991, **95**, 5610–5620.

65   E. G. Alexov and M. R. Gunner, *Biochemistry*, 1999, **38**, 8253–8270.

66   V. Z. Spassov, H. Luecke, K. Gerwert and D. Bashford, *J. Mol. Biol.*, 2001, **312**, 203–219.

67   Y. Song, J. Mao and M. R. Gunner, *Biochemistry*, 2003, **42**, 9875–9888.

68   B. Rabenstein, G. M. Ullmann and E.-W. Knapp, *Biochemistry*, 1998, **37**, 2488–2495.

69   Z. Zhu and M. R. Gunner, *Biochemistry*, 2005, **44**, 82–96.

70   R. E. Georgescu, E. G. Alexov and M. R. Gunner, *Biophys. J.*, 2002, **83**, 1731–1748.

71   J. Antosiewicz, J. A. McCammon and M. K. Gilson, *J. Mol. Biol.*, 1994, **238**, 415–436.

72   J. Antosiewicz, J. A. McCammon and M. K. Gilson, *Biochemistry*, 1996, **35**, 7819–7833.

73   L. Sandberg and O. Edholm, *Proteins Struct. Funct. Genet.*, 1999, **36**, 474–483.

74   I. Muegge, P. X. Qi, A. J. Wand, Z. T. Chu and A. Warshel, *J. Phys. Chem. B*, 1997, **101**, 825–836.

75   T. Simonson, J. Carlsson and D. A. Case, *J. Am. Chem. Soc.*, 2004, **126**, 4167–4180.

76   T. J. You and D. Bashford, *Biophys. J.*, 1995, **69**, 1721–1733.

77   P. Beroza and D. A. Case, *J. Phys. Chem.*, 1996, **100**, 20156–20163.

78   G. Kieseritzky and E.-W. Knapp, *Proteins Struct. Funct. Bioinforma.*, 2008, **71**, 1335–1348.

79   P. Barth, T. Alber and P. B. Harbury, *Proc. Natl. Acad. Sci.*, 2007, **104**, 4898–4903.

80   J. Warwicker, *J. Theor. Biol.*, 1986, **121**, 199–210.

81   P. Koehl and M. Delarue, *J. Mol. Biol.*, 1994, **239**, 249–275.

82   C. Cole and J. Warwicker, *Protein Sci.*, 2009, **11**, 2860–2870.

83   E. G. Alexov and M. R. Gunner, *Biophys. J.*, 1997, **72**, 2075–2093.

84   Y. Song, J. Mao and M. R. Gunner, *J. Comput. Chem.*, 2009, **30**, 2231–47.

85   L. Wang, L. Li and E. Alexov, *Proteins Struct. Funct. Bioinforma.*, 2015, **83**, 2186–2197.

86   J. P. Cvitkovic, C. D. Pauplis and G. A. Kaminski, *J. Comput. Chem.*, 2019, **40**, 1718–1726.

87   F. Milletti, L. Storchi and G. Cruciani, *Proteins Struct. Funct. Bioinforma.*, 2009, **76**, 484–495.

88   K. P. Tan, T. B. Nguyen, S. Patel, R. Varadarajan and M. S. Madhusudhan, *Nucleic Acids Res.*, 2013, **41**, W314–W321.

89   C. X. Zhou, W. M. Grumbles and T. R. Cundari, *ChemRxiv*, ,
       DOI:10.26434/chemrxiv.12646772.

90   V. Sinha, J. J. Laan and E. A. Pidko, *Phys. Chem. Chem. Phys.*, 2021, **23**, 2557–2567.

91   J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.

92   J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev and A. E. Roitberg, *J. Chem. Phys.*, 2018,
       148, 241733.

93   J. S. Smith, O. Isayev and A. E. Roitberg, *Sci. Data*, 2017, **4**, 170193.

94   J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak,
       O. Isayev and A. E. Roitberg, *Nat. Commun.*, 2019, **10**, 1–8.

95   J. S. Smith, R. Zubatyuk, B. Nebgen, N. Lubbers, K. Barros, A. E. Roitberg, O. Isayev
       and S. Tretiak, *Sci. Data*, 2020, **7**, 1–10.

96   C. Devereux, J. S. Smith, K. K. Huddleston, K. Barros, R. Zubatyuk, O. Isayev and A. E.
       Roitberg, *J. Chem. Theory Comput.*, 2020, **16**, 4192–4202.

97   X. Gao, F. Ramezanghorbani, O. Isayev, J. S. Smith and A. E. Roitberg, *J. Chem. Inf.
       Model.*, 2020, **60**, 3408–3415.

98   J. M. Stevenson, L. D. Jacobson, Y. Zhao, C. Wu, J. Maple, K. Leswing, E. Harder and R.
       Abel, 2019, 1–20.

99   R. Zubatyuk, J. S. Smith, J. Leszczynski and O. Isayev, *Sci. Adv.*, 2019, **5**, eaav6490.

100   H. Gokcan and O. Isayev, *WIREs Comput. Mol. Sci.*, , DOI:10.1002/wcms.1564.

101   T. Zubatiuk and O. Isayev, *Acc. Chem. Res.*, 2021, **54**, 1575–1585.

102   S. Pahari, L. Sun and E. Alexov, *Database (Oxford).*, 2019, **2019**, 1–7.

103   H. Webb, B. M. Tynan-Connolly, G. M. Lee, D. Farrell, F. O'Meara, C. R. Søndergaard,
       K. Teilum, C. Hewage, L. P. McIntosh and J. E. Nielsen, *Proteins Struct. Funct.
       Bioinforma.*, 2011, **79**, 685–702.

104   S. Xiao, V. Patsalo, B. Shan, Y. Bi, D. F. Green and D. P. Raleigh, *Proc. Natl. Acad. Sci.*,
       2013, **110**, 11337–11342.

105   K. Bartik, C. Redfield and C. M. Dobson, *Biophys. J.*, 1994, **66**, 1180–1184.

106   S. Kuramitsu and K. Hamaguchi, *J. Biochem.*, 1980, **87**, 1215–9.

107   T. Takahashi, H. Nakamura and A. Wada, *Biopolymers*, 1992, **32**, 897–909.

108   F. Inagaki, T. Miyazawa, H. Hori and N. Tamiya, *Eur. J. Biochem.*, 1978, **89**, 433–42.

109   Y.-H. Kao, C. A. Fitch, S. Bhattacharya, C. J. Sarkisian, J. T. J. Lecomte and B. García-Moreno E., *Biophys. J.*, 2000, **79**, 1637–1654.

110   D. Bashford, D. A. Case, C. Dalvit, L. Tennant and P. E. Wright, *Biochemistry*, 1993, **32**, 8045–8056.

111   L. Yu and S. W. Fesik, *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.*, 1994, **1209**, 24–32.

112   W. Schaller and A. D. Robertson, *Biochemistry*, 1995, **34**, 4714–4723.

113   L. Swint-Kruse and A. D. Robertson, *Biochemistry*, 1995, **34**, 4724–4732.

114   M. Betz, F. Löhr, H. Wienk and H. Rüterjans, *Biochemistry*, 2004, **43**, 5820–5831.

115   E. Arbely, T. J. Rutherford, T. D. Sharpe, N. Ferguson and A. R. Fersht, *J. Mol. Biol.*, 2009, **387**, 986–992.

116   Y. Oda, T. Yamazaki, K. Nagayama, S. Kanaya, Y. Kuroda and H. Nakamura, *Biochemistry*, 1994, **33**, 5275–5284.

117   G. Zhang, A. S. Mazurkie, D. Dunaway-Mariano and K. N. Allen, *Biochemistry*, 2002, **41**, 13370–13377.

118   W. R. Baker and A. Kintanar, *Arch. Biochem. Biophys.*, 1996, **327**, 189–199.

119   S. FUJII, K. AKASAKA and H. HATANO, *J. Biochem.*, 1980, **88**, 789–796.

120   Y.-J. Tan, M. Oliveberg, B. Davis and A. R. Fersht, *J. Mol. Biol.*, 1995, **254**, 980–992.

121   E. Arbely, T. J. Rutherford, H. Neuweiler, T. D. Sharpe, N. Ferguson and A. R. Fersht, *J. Mol. Biol.*, 2010, **403**, 313–327.

122   J. D. Forman-Kay, G. M. Clore and A. M. Gronenborn, *Biochemistry*, 1992, **31**, 3442–3452.

123   M. M. Zhou, J. P. Davis and R. L. Van Etten, *Biochemistry*, 1993, **32**, 8479–8486.

124 P. A. Tishmack, D. Bashford, E. Harms and R. L. Van Etten, *Biochemistry*, 1997, **36**, 11984–11994.

125 V. Dillet, R. L. Van Etten and D. Bashford, *J. Phys. Chem. B*, 2000, **104**, 11321–11333.

126 M. D. Joshi, A. Hedberg and L. P. Mcintosh, *Protein Sci.*, 2008, **6**, 2667–2670.

127 D. V. Laurents, B. M. P. Huyghues-Despointes, M. Bruix, R. L. Thurlkill, D. Schell, S. Newsom, G. R. Grimsley, K. L. Shaw, S. Treviño, M. Rico, J. M. Briggs, J. M. Antosiewicz, J. M. Scholtz and C. N. Pace, *J. Mol. Biol.*, 2003, **325**, 1077–1092.

128 D. A. Case, H. M. Aktulga, K. Belfon, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham, III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, G. Giambasu, M. K. Gilson, H. Gohlke, A. W. Goetz, R. Harris, S. Izadi, S. A. Izmailov, C. Jin, K. Kasavajhala, M. C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K. M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K. A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, N. R. Skrynnikov, J. Smith, J. Swails, R. C. Walker, J. Wang, H. Wei, R. M. Wolf, X. Wu, Y. Xue, D. M. York, S. Zhao and P. A. Kollman, 2021.

129 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J. Chem. Theory Comput.*, 2015, **11**, 3696–3713.

130 I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goñi, A. Balaceanu, G. Portella, F. Battistini, J. L. Gelpí, C. González, M. Vendruscolo, C. A. Laughton, S. A. Harris, D. A. Case and M. Orozco, *Nat. Methods*, 2016, **13**, 55–58.

131 I. Guyon, J. Weston, S. Barnhill and V. Vapnik, *Mach. Learn.*, 2002, **46**, 389–422.

132 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.

133 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

134 V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan and B. P. Feuston, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1947–1958.

135 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem.*

Phys., 1983, **79**, 926–935.

136 A. W. Götz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand and R. C. Walker, *J. Chem. Theory Comput.*, 2012, **8**, 1542–1555.

137 R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand and R. C. Walker, *J. Chem. Theory Comput.*, 2013, **9**, 3878–3888.

138 H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684–3690.

139 J.-P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, *J. Comput. Phys.*, 1977, **23**, 327–341.

140 U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *J. Chem. Phys.*, 1995, **103**, 8577–8593.

141 B. N. Koleva, H. Gokcan, A. A. Rizzo, S. Lim, K. Jeanne Dit Fouque, A. Choy, M. L. Liriano, F. Fernandez-Lima, D. M. Korzhnev, G. A. Cisneros and P. J. Beuning, *Biophys. J.*, 2019, **117**, 587–601.

142 S. Jaeger, S. Fulle and S. Turk, *J. Chem. Inf. Model.*, 2018, **58**, 27–35.

143 X. Li and D. Fourches, *J. Cheminform.*, 2020, **12**, 27.

144 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268–276.

145 D. Gfeller, O. Michielin and V. Zoete, *Nucleic Acids Res.*, 2012, **41**, D327–D332.

146 L. Van Der Maaten and G. Hinton, *J. Mach. Learn. Res.*, 2008, **9**, 2579–2625.