A fast and interpretable deep learning approach for accurate electrostatics-driven pKa predictions in proteins

Pedro Reis (pdreis@ciencias.ulisboa.pt)
BioISI - Biosystems & Integrative Sciences Institute https://orcid.org/0000-0003-3563-6239
Marco Bertolini
Bayer AG
Floriane Montanari
Bayer AG
Walter Rocchia
Istituto Italiano di Tecnologia https://orcid.org/0000-0003-2480-7151
Miguel Machuqueiro
BioISI - Biosystems & Integrative Sciences Institute https://orcid.org/0000-0001-6923-8744
Djork-Arné Clevert
Bayer AG

Article

Keywords: pKa shifts, deep learning, accuracy

Posted Date: March 23rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-949180/v2

A fast and interpretable deep learning approach for accurate electrostatics-driven $\mathrm{p}K_\mathrm{a}$ predictions in proteins

Pedro B.P.S. Reis[1,2,3,*], Marco Bertolini[1], Floriane Montanari[1], Walter Rocchia[3], Miguel Machuqueiro[2, *], and Djork-Arné Clevert[1, *]

[1]Bayer A.G., Machine Learning Research, Berlin, Germany
[2]BioISI - Biosystems & Integrative Sciences Institute, Faculty of Sciences, University of Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
[3]CONCEPT Lab, Istituto Italiano di Tecnologia, Via E. Melen 83, 16152 - Genova, Italy
[*]e-mail:

ABSTRACT

Existing computational methods to estimate $\mathrm{p}K_\mathrm{a}$ values in proteins rely on theoretical approximations and lengthy computations. In this work, we use a data set of 6 million theoretically determined $\mathrm{p}K_\mathrm{a}$ shifts to train deep learning models that are shown to rival the physics-based predictors. These neural networks managed to assign proper electrostatic charges to chemical groups, and learned the importance of solvent exposure and close interactions, including hydrogen bonds. Although trained only using theoretical data, our pKAI+ model displays the best accuracy on a test set of $\sim$750 experimental values. Inference times allow speedups of more than 1000 times faster than physics-based methods. By combining speed, accuracy and a reasonable understanding of the underlying physics, our models provide a game-changing solution for fast estimations of macroscopic $\mathrm{p}K_\mathrm{a}$ from ensembles of microscopic values as well as for many downstream applications such as molecular docking and constant-pH molecular dynamics simulations.

Main

Many biological processes are triggered by changes in the ionization state of key amino acid side-chains[1,2]. Experimentally, the titration behavior of a molecule can be measured using potentiometry or by tracking free energy changes across a pH range. For individual sites, titration curves can be derived from infrared

or NMR spectroscopy. Detailed microscopic information can be quickly and inexpensively obtained with computational methods, and several *in silico* pK_a calculations have become widely used to provide insights about structural and functional properties of proteins[3–5].

In Poisson–Boltzmann-based (PB) methods, the solvent is implicitly described while proteins are represented by point charges in a low dielectric medium[3,4,6,7]. These continuum electrostatics (CE) methods assume that the pK_{half} (the proton binding affinity for a chemical group in a given conformation) is a good estimate for the macroscopic pK_a value. This assumption holds when the protein structure is sufficiently representative of the conformational ensembles corresponding to both protonation states. Experimentally determined structures exhibit conformations at a minimum energy state, which, in turn, is related to a specific protonation state. However, biomolecular systems can explore different energy basins, which may exhibit alternative protonation states. Energy minima can be affected by experimental conditions, such as temperature, ionic strength and pH. Inaccuracies in pK_a predictions due to limited conformational rearrangements can be reduced by increasing the protein dielectric constant from its default value (2-4), which only accounts for electronic polarisation. The dielectric constant can be used as an empirical parameter mimicking the effect of the response mechanisms to the local electric field that are not explicitly taken into account in the model[8–12]. A more computationally expensive approach is to explicitly include protein motion by sampling conformers via Monte Carlo (MC) or molecular dynamics (MD) simulations and applying conformational averaging[4,13–15]. Finally, by coupling the sampling of protonation states at given pH and conformations, constant-pH MD methods[16–20] provide greater insight into pH-dependent processes[21–25].

As larger data sets of experimental pK_a values have become available, a new class of purely empirical methods has been developed. These models rely on statistical fits of empirical parameters weighting the different energetic contributions into simplified functions. PROPKA[5] is arguably the most popular of such methods[26], and has been shown to perform competitively even when compared to higher-level theory methods[6,27]. The empirical methods are much faster than the physics-based ones although at the cost of providing less microscopic insights, and their predictive power is unknown on mutations and/or proteins dissimilar to those composing the training set.

The accuracy of most predictors is bound to the estimation of the same quantity, the so-called ΔpK_a. This is the free energy of transferring the ionizable residue from the solvent to the protein, compared to its neutral counterpart. Since pK_a values for all amino acids in water have been experimentally determined, the $pK_a^{solvent}$ term can be fixed and, in practice, it can also be adjusted to incorporate systematic errors. The ΔpK_a can be regarded as a sum of mostly electrostatic contributions stemming from the residue microenvironment. Therefore, an accurate prediction of pK_a values for a given conformation requires a

correct description of the residue interactions with the surrounding protein charges and with the solvent.

At their core, deep learning (DL) models are complex non-linear empirical functions fitted to best map input variables to output properties. Considering chemical properties, such as pK_a values, which are dictated by molecular configurations, and provided that enough examples are presented, it is possible to train a model to map this relationship without the need to solve non-linear equations in 3D or to sort through the massive space of possible states.

In this paper, we have developed two DL-based pK_a predictors: pKAI and pKAI+, for pK_{half} and experimental pK_a values, respectively. These models have been trained on a database with ∼6 million pK_a values estimated from ∼50 thousand structures using a continuum electrostatics method, PypKa[6]. pKAI+ displays an unrivaled performance at predicting experimental pK_a values on a ∼750 members data set. Also, pKAI exhibits an accuracy comparable to the PB-based predictor used to generate the training set while being approximately 10–1000× faster. By applying explainable artificial intelligence (XAI) analysis, we show that these simple models are able to implicitly model most of the required energetic contributions such as Coulomb interactions, desolvation and hydrogen-bonding. Therefore, the presented models feature the best characteristics of CE-based methods – accuracy and interpretability – with the speed provided by empirical approaches.

Results

The main goal of pKAI is to mimic the pK_a predictive ability of PB-based methods with a significant computational performance improvement. Our training set is comprised of pK_a values calculated using PypKa on a large number of proteins taken from the Protein Data Bank[28]. An elaborate data split was performed to minimize data leakage from the training set to the validation and test sets (see Methods). pKAI was designed to be a simple and interpretable model using the minimum structural features that still capture the electrostatic environment surrounding every titratable residue. The model has been trained on ΔpK_a values rather than on absolute values. The pK_a shift is in fact a more appropriate quantity to learn, less dependent on the chemical peculiarities of individual amino-acids and more sensitive to the local electrostatic environment. For example, residues that share a common side-chain chemical group (such as glutamate and aspartate sharing a carboxylic acid) are influenced by the same environment in a similar way.

We wanted our model to capture the electrostatic dependence between the environment of a residue and its consequent pK_a shift while keeping the input layer as small as possible (see Methods). By ignoring all carbon and hydrogen atoms, we are greatly reducing the dimensionality of our input layer, while retaining

most of the information regarding charged particles. There is of course a significant loss of topological information, although much can be inferred from the positions of the included atoms. In fact, there is no performance gain when adding solvent exposure measurements (e.g. SASA, residue depth) to the environment embedding. Considering that solvent exposure entails topological information and that the model is not able to extract additional information from it, we conclude that it was already estimating, to some degree, these molecular properties (see Model Explainability subsection).

pKAI: predicting theoretical pK_a

The performance of the model on the test set is reported in Supplementary Table S1 and in Figure 1A. The null model used for comparison consists of the reference pK_a value in water for each residue type, and corresponds to 0 in the ΔpK_a scale. Overall, pKAI reproduces the PB-based data with a MAE value of 0.31 and a RMSE of 0.52. However, in this case, we are only predicting theoretical values with a well-defined relation between structure and pK_{half} (pK value of a single conformation). Experimental pK_a estimation is a much more complex task since the pK_{half} values corresponding to the different conformations spanned by the protein should be weighted according to their occurrence probability at equilibrium. The performance of pKAI is impressive considering the high complexity of the dependence between pK_a and the site electrostatic environment, illustrated by the high RMSE value of the Null model (1.89). Some residues are easier to predict (e.g. LYS and termini residues) while others are more challenging (e.g. CYS and TYR). This can be explained by their solvent exposure distribution (Figure 1B): well-solvated residues exhibit small ΔpK_a values while more buried ones are more affected by the desolvation effect and establish more interactions with other residues causing their pK_a values to shift. There is a clear dependency between the solvent exposure of a residue, its ΔpK_a value and the prediction difficulty (Supplementary Figure S1). The excellent performance of pKAI is also demonstrated by the fact that most predictions (81.2%) exhibit an error below 0.5 pK units, which is a sufficient for most use cases.

The main advantage of DL models is the potential speedup they can provide. Since continuum electrostatics (CE) pK_a estimations need to sample thermodynamic equilibrium microstates, several iterative simulations have to be performed on each protonation state and on the environment of every residue. On the other hand, pKAI merely needs to apply its learned function over each residue and, as such, is remarkably faster (Figure 1C). Moreover, the convergence of the CE simulations is harder to achieve as the protein size increases. Consequently, in PypKa, as the protein size increases, so does the time required to estimate each pK_a value. In contrast, the run time of pKAI's DL model has a different dependence on the protein size. Since the bigger is the protein the larger is the amount of calculations that can be performed simultaneously, then the less significant becomes the model loading cost and the faster the average per-residue execution time. This results in sublinear scaling performance and in a speedup

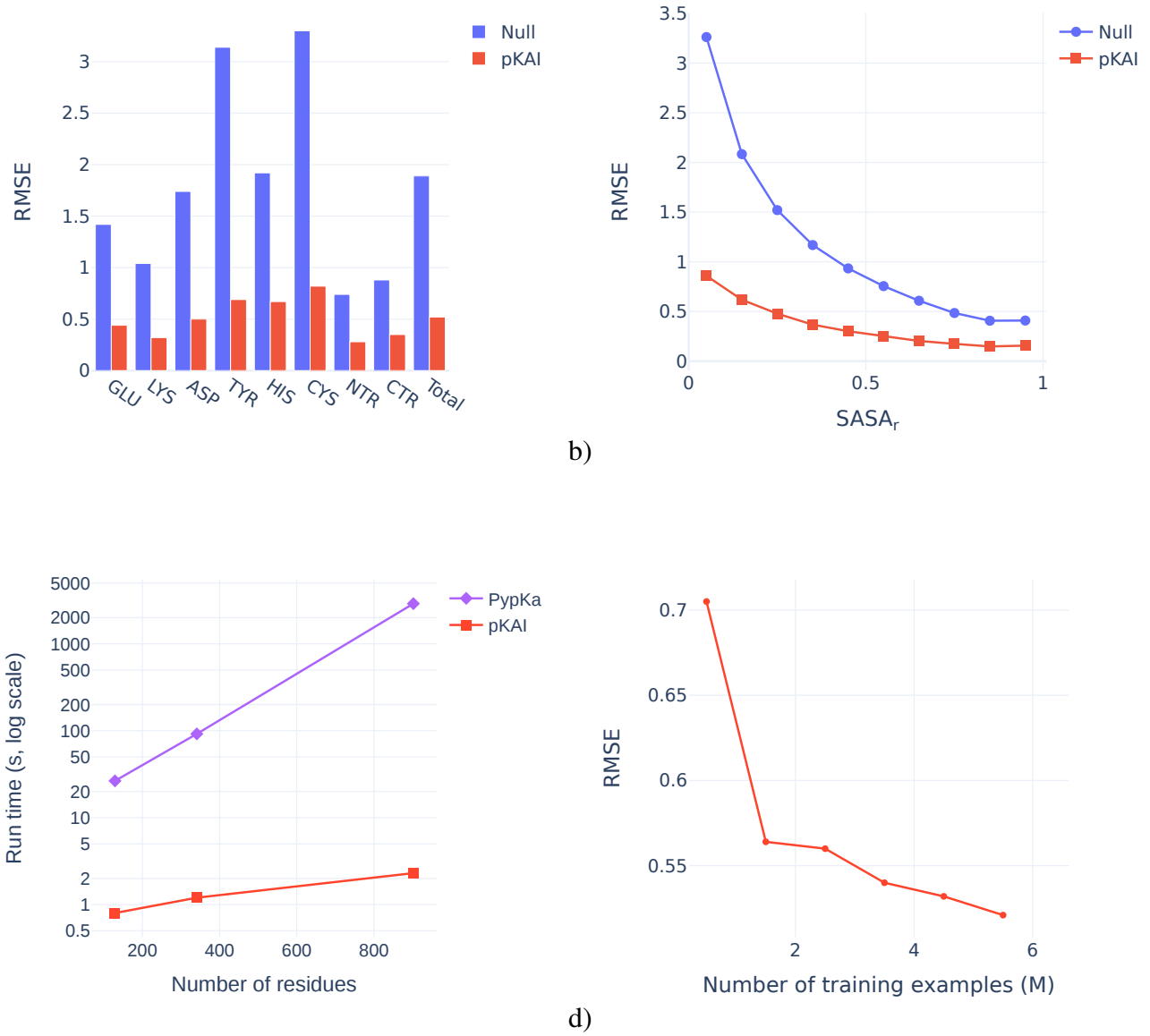


Figure 1. A) Comparison between Null model and pKAI RMSE values (values shown in Supplementary Table S1). The Null model is defined as the pK_a values of the residues in water taken from reference 29. B) Performance at predicting pK_{half} values dependency on the magnitude of solvent exposure (SASA). The calculations were performed for pKAI and Null model using the PypKa predictions as reference. C) Execution time comparison between PypKa and pKAI (values shown in Supplementary Table S2). This benchmark was executed on a machine with a single Intel Xeon E5-2620 processor. D) Effect of the size of the training set in the model performance on the validation set.

over its CE counterpart that can exceed over a thousand times. As such, pKAI is a particularly valuable tool for dealing with very large systems with thousands of residues where the only added computational cost stems from the prepossessing of the structure.

Another important factor contributing to the high accuracy obtained is the considerable size of the training set. Despite using the largest repository of experimental protein structures and the largest pK_a database available[28], we show that there is still a correlation between the number of examples in the training set and the accuracy of the model (Figure 1D).This indicates that our model can still be improved by providing further examples of pK_a values. To avoid limiting the scaling rate by the availability of new experimental protein structures, we can generate new and uncorrelated protein structures using conformational sampling methods, such as MD and MC. Another advantage of using computational methodologies is guiding the protein conformational sampling to achieve electrostatic environments that are underrepresented in the training set. We reserve this development for future work.

pKAI+: Predicting experimental pK_a values

The main goal of pK_a predictors, such as PypKa, is to estimate the macroscopic pK_a value for the titratable residues using structures (usually experimental ones). Since pKAI aims at reproducing the pK_{half} calculated with PypKa at a fraction of the computational cost, it is not expected to outperform the PB-based method in predicting experimental values. When using PB to predict experimental pK_as, a higher dielectric constant for the solute is often adopted to compensate for the lack of conformational flexibility in the method and the lack of representativity of the experimental input structure. A similar approach can be implemented in pKAI by introducing a regularization weight to the cost function (pKAI+). This regularization penalizes the magnitude of the ΔpK_a prediction. In practice, this procedure biases our estimates towards the pK_a values in water, similarly to what is done by the increased solute dielectric constant in PB-based approaches. It should be noted that pKAI+ has not been trained on experimental pK_a, but rather on the same training set as pKAI.

To evaluate the performance of our model, we have benchmarked it using a data set of 736 titratable residues in 97 proteins with experimentally determined pK_a values (Figure 2A). Remarkably, pKAI+ (RMSE of 0.98) is able to outperform both PypKa (RMSE of 1.07) and PROPKA (state-of-the art empirical pK_a predictor, RMSE of 1.11). Furthermore, the improvement over the other methods is significant for most residue types (Figure 2B), and can be quantified using metrics that are more (RMSE, 0.9 quantile) or less (MAE, error percentage under 0.5) sensitive to the presence of outliers (Supplementary Table S3). Cysteine residues are particularly difficult to predict because they naturally occur less frequently and are more buried than all other titratable residues. This leads to an under-representation of these residues in the

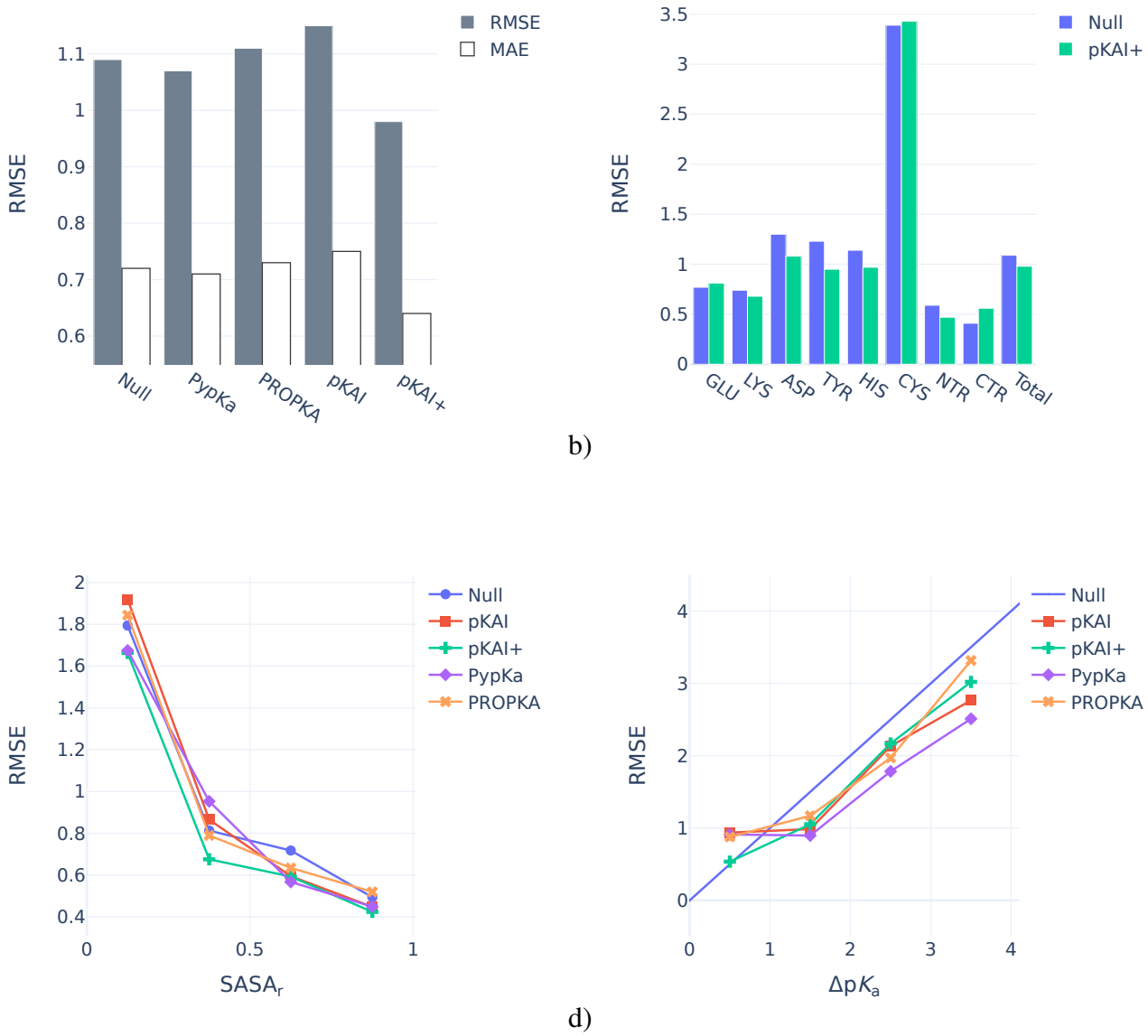


Figure 2. A) Experimental pK_a benchmark of several methods on a data set of 736 residues from 97 proteins (values shown in Supplementary Table S4). The null model values are the pK_a values of each amino acid substituted in an alanine pentapetide (Ace-AA-X-AA-NH_2)[29,30]. B) Comparison between Null model and pKAI+ performance by residue type. C) pKAI+ performance at predicting experimental pK_a values dependency on the magnitude of solvent exposure (SASA) of the residues. D) Prediction errors of the different models given the experimental pK_a shift (ΔpK_a)

training set while exhibiting the largest pK_a shifts. To illustrate the difficulty of this data set, note that some methodologies are not able to improve on the null model (RMSE of 1.09). The reported deviations are specific to this data set. Even though our benchmark is one of the largest ever used to validate a pK_a predictor, it is likely still insufficient to quantify the true accuracy of these methods. Furthermore, besides being limited, these test sets used for validating new pK_a predictors tend to always be different. This makes it very hard to compare methods without rerunnig them. In this benchmark, PypKa represents the PB-based methods like DelPhiPKa[7] or H++[3]. More computationally expensive methods such as MCCE[31] or constant-pH MD are not represented here. These methods are expected to outperform PB-based methods, which rely on a single structure, although the exact improvement on this test set is hard to predict. DeepKa is a recently published convolutional neural network trained on theoretical pK_a values from constant-pH MD (CpHMD) simulations[32]. As expected, the CpHMD implemented in the Amber suite[33] (RMSE of 1.02) outperformed PROPKA (RMSE of 1.12) in their test set which only includes the 4 residues (Asp, Glu, His and Lys) predicted by DeepKa (RMSE of 1.05).

Our test set can be divided by solvent exposure (SASA) of the titrating residue. pKAI+ shows comparable RMSE values to PypKa for both the most solvent exposed and buried residues (Figure 2C). Interestingly, it is also able to surpass the PB-based model for partially exposed residues. Notably, pKAI+ only improves the PypKa predictions for pK_a shifts smaller than 1 pK unit (Figure 2D). This indicates that pKAI+ corrects the pK_a values of partially exposed residues which are establishing non-representative interactions in the experimental structure. Since there is a large number of residues with these characteristics in the test set[28], the overall performance improvement is significant (Supplementary Table S4).

Model Explainability

The main driving force for pK_a shifts in proteins is electrostatic in nature. In our model, each atom of the environment represents the contribution of a chemical group or part of a residue. This individual contribution towards the final ΔpK_a prediction can be estimated (see XAI in the Methods section for further details) and it is shown in Figure 3A. Remarkably, although our model has been given no information about atomic charges, it assigns contributions that are in agreement with the expected overall charge of the atom class. Cationic amine groups (NZ_LYS; NH_ARG; NE_ARG; NE2_HIS) are clearly assigned positive scores (i.e. destabilize the protonation of the titratable residue) and are easily distinguishable from the anionic carbonyl groups (O_COOH from Glu, Asp and C-termini residues). These scores provide a general insight into the network's interpretation of each atom and should not be used for more quantitative analysis. Since the atom score is an averaged measure across the test set, an imbalance of closely interacting atoms of a specific class can dramatically skew its median contribution.

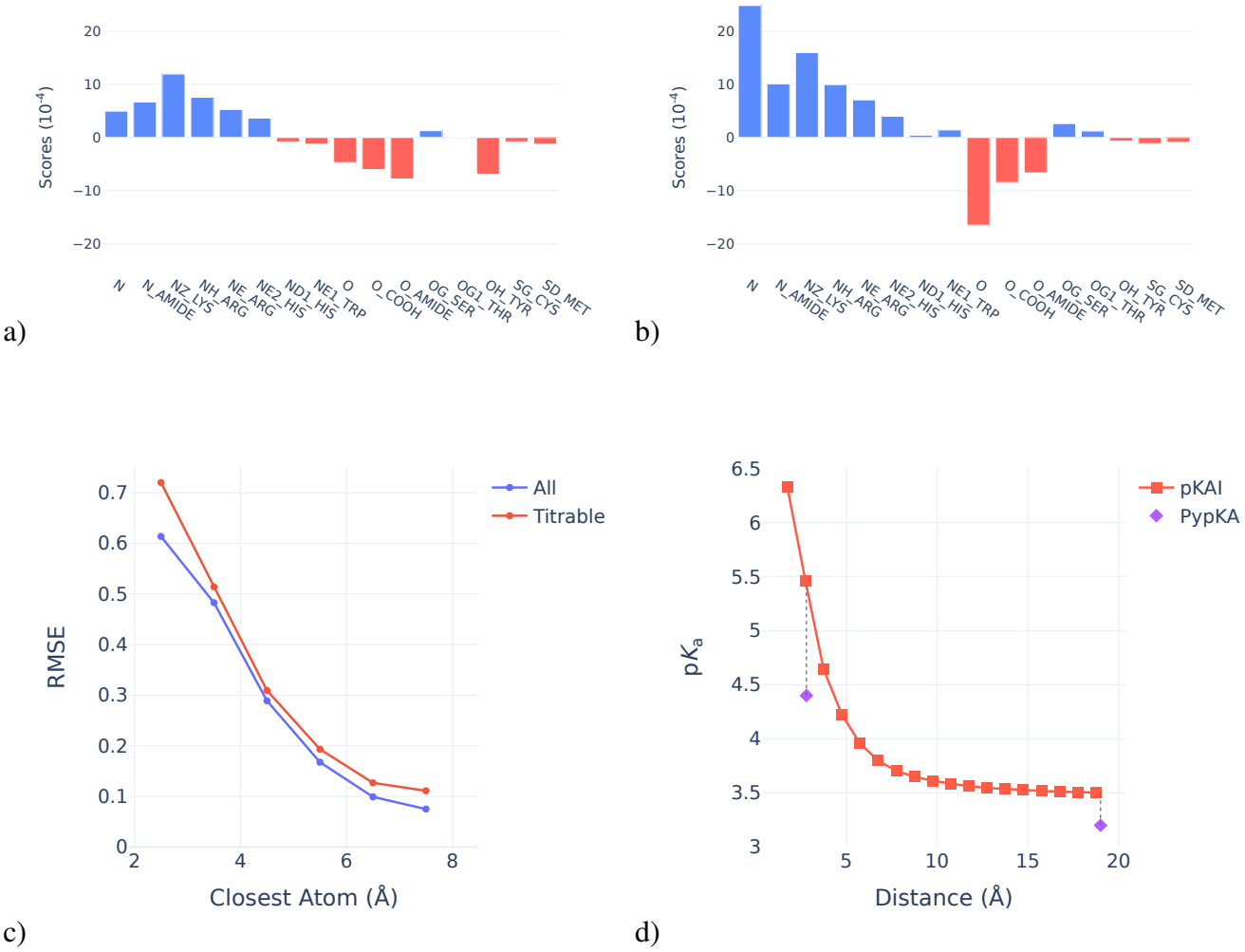
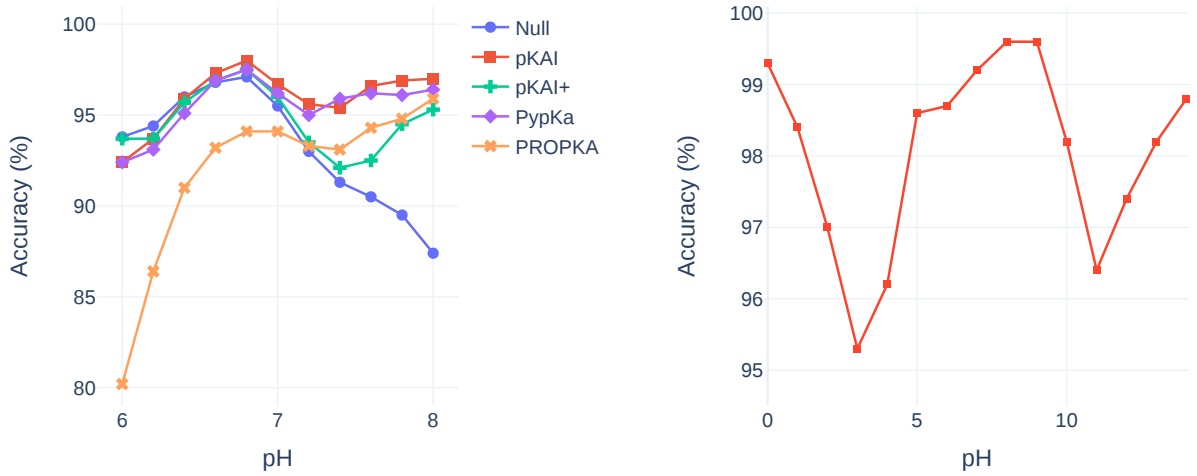


Figure 3. Charge scores attributed by pKAI to all considered input atoms classes (Supplementary Table S5) of all atoms (A) and atoms closer than 6 Å (B). C) Closest atom influence on pKAI performance. D) Impact of changing the distance of the closest atom on pKAI predictions of residue TYR-315 from structure 2BJU. For reference, we have included PypKa predictions of the same residue in the state presented in the experimental structure (closest distance 2.8 Å) and in a modified structure in which the closest atom is absent.

Hydrogen bonds are one of the strongest interactions found in proteins, and, as such, their proper description is crucial to obtain accurate pK_a predictions. By comparing Figures 3 A and B we can observe marked differences between the atom scores at close proximity and those farther away from the titrating residue. For example, the average score of the very abundant classes of primary amines (N; N_AMIDE) and carbonyl groups (O; O_AMIDE) is greatly diminished when compared to their short-range contributions, where these become hydrogen donors and acceptors, respectively. The anionic Tyr residue is perceived to have an overall negative contribution, except when it is close to another titratable residue; in this case there seems to be no preferred state as it can act both as a donor and as an acceptor – like any titratable residue. On the other hand, the contribution of neutral non-titrating alcohol groups (OG_SER; OG1_THR) is almost exclusively attributed to their potential to form hydrogen bonds at short range. Overall, the model is able to capture an astonishing amount of details concerning the physics underlying pK_a shifts.

Beyond the general understanding shown before, hydrogen bond contributions are hard to account for, compared to other interactions. As shown in Figure 3C the closer another residue (blue curve) is to the titrating one, the harder for the model is to correctly describe their interaction. The difficulty of the prediction increases dramatically at the typical distance of hydrogen bonds (2.5-3.2Å). This is even more marked if one considers interactions established between two titratable residues (red curve). In this case, the network has to solve for the pK_a of both residues simultaneously, and in many instances it is unable to do so. Hence, predicting the contribution of the remaining environment is easier than that of a single hydrogen bond. This is illustrated in Figure 3D, in which the agreement with the physics-based method is much higher when the closest atom is removed from the structure rather than when it is kept in its original position. Although many other profiles can be observed (Supplementary Figure S2), this trend is generally conserved. Considering that the model did not receive explicit information about hydrogen bonds, it is quite remarkable that it was able to correlate this type of interaction with larger pK_a shifts.

Solvent exposure is another property that is usually a key contributor to pK_a shifts. The models are trained without explicit knowledge of the 3D structure of the protein, and deprived of information regarding carbon atoms. Nevertheless, they seem to learn about the solvent exposure contribution. We compared the correlations (Pearson correlation coefficient r and Spearman's rank correlation coefficient ρ) between the calculated SASA and the pK_a shifts over the entire test data set. Using the known ΔpK_a, we obtained $r_{\Delta\text{pka}} = -0.68$; $\rho_{\Delta\text{pka}} = -0.60$, while using the predicted ΔpK_a, we got $r_{\text{pred}} = -0.66$; $\rho_{\text{pred}} = -0.62$. The similarity between these values indicates that the model has learned the correct correlation between SASA and the pK_a shift. Additionally, we trained a model where we provided SASA as an additional input and observed no performance improvement (data not shown).



a) b)

Figure 4. A) Accuracy of several methods at predicting a representative protonation states derived from experimental pK_a values. Residues at a pH within 1.5 units of the experimenal pK_a are considered not to have a representative protonation state. B) pKAI accuracy at predicting PypKa derived protonation states.

Finally, it is worth mentioning that XAI analysis was a driving factor in the development of pKAI. In fact, the importance that the model assigns to each atom class (similar to Figure 3) was pivotal to select the final set of atom classes aimed at describing the surrounding environment residues.

Discussion

We have introduced pKAI and pKAI+, two deep learning models to predict theoretical and experimental ΔpK_a values, respectively. pKAI offers unprecedented efficiency, exhibiting a remarkable trade-off between accuracy and computational speed, its performance rivaling that of CE-based methods, such as PypKa. pKAI could be used as a replacement for such methods, especially when dealing with large proteins or applications requiring multiple CE calculations, like constant-pH MD simulations[16–20]. Considering the latest advances in sequence to structure predictions[34], faster methods, such as pKAI, will likely be of use as exponentially more structures become available. Furthermore, when optimizing new structures for binding to specific targets (e.g. design of enzymes and/or antibodies), it is vital to have an accurate prediction of the protonation states.

While we strive for optimal accuracy, we are aware that many applications will only require a binary decision (hence a qualitative prediction of pK_a shifts would be sufficient). For example, when selecting the

most likely protonation state of a protein, one only needs to predict whether each pK_a is larger or smaller than the pH value of interest. The most abundant protonation states obtained from pKAI predictions are in high agreement with those derived from experiments (Supplementary Figure S3) and outperform those of PROPKA in a wide range of pH values. Interestingly, PROPKA and the Null model seem to perform particularly well at extreme pH values. Nevertheless, pKAI is the best model at assigning a fixed protonation state to a protein at biologically relevant pH values (Fig 4A), arguably the most common task pK_a predictors are used for. pKAI+ is biased to predict pK_a values between those of pKAI and the Null model. While this bias has granted the model an edge on experimental pK_a estimation, in tasks in which the Null model does not perform well, pKAI+'s ability is also affected. This can be seen in the biological range at the more basic pH values.

Several other applications only require an estimation of the proton binding affinity using a fixed conformation. This quantity, termed pK_{half}, renders a good prediction of the macroscopic pK_a when averaged over a representative ensemble of conformations. From pK_{half} values, the most abundant/representative protonation states for a particular conformation can be calculated, improving the realism of methods such as molecular dynamics[16–20] and molecular docking[35]. pKAI is nearly perfect at mimicking representative protonation states given by PypKa, being particularly effective at physiological pH, achieving an astounding accuracy of 99.4% (Fig 4B). In a conformational ensemble, there are always many representative protonation states which differ significantly from the one calculated using the macroscopic pK_a values. Therefore, coupling pK_{half} calculations with conformational sampling techniques is very appealing in theory but difficult in practice, due to their computational cost. By using pKAI instead of PypKa (or any other PB-based method), one would drastically decrease the computational overhead (up to 1000×).

pKAI does not handle all residues with the same performance. Difficult cases are caused by low representation in the training set, low solvent exposure, and/or close-by residues providing H bond interactions. These peculiar environments usually present a high ΔpK_a which is not handled very well by the method. One clear way to improve our models would therefore be to introduce more training examples. Furthermore, the inclusion of more training data with rare environments would definitely enhance performance. To better handle interactions with neighboring titratable groups, a change of environment encoding would be needed. One approach to be explored in future work would be to represent the whole protein as a graph, and use graph neural networks algorithms to learn the ΔpK_a values.

Although pKAI excels at predicting pK_{half} values, its performance is modest when estimating experimental pK_a values. Inspired by the observation that increasing the dielectric constant in PB-based methods improves their agreement with experimental results, we have introduced a regularization parameter into the cost function. Similar to the dielectric constant, this regularization weight biases all predictions towards

the residue's pK_a values in water. The new model, pKAI+, outperforms all methods tested in this work, including PypKa which was used to create the training set. However, this improvement, while significant for partially exposed residues which would otherwise exhibit overestimated pK_a shifts, penalizes the accuracy of more shifted residues.

With pKAI and pKAI+, we are introducing the first deep learning-based predictor of pK_a shifts in proteins trained on continuum electrostatics data. The unique combination of speed and accuracy afforded by our models represents a paradigm shift in pK_a predictions. pKAI paves the way for accurate estimations of macroscopic pK_a values from ensemble calculations of pK_{half} values, overcoming previous computational limits. By design, the models were trained using a very simplified view of the surroundings of the titratable group, accounting only for residues within a 15 Å cutoff, and ignoring all carbon and hydrogen atoms. This design choice allowed for the models to stay small and fast. Explainability methods confirmed that this input information was enough for the model to capture crucial features such as electrostatics, solvent exposure, and environment contributions. The models' initial success introduces several opportunities for further research, including problem encoding, accounting for conformational flexibility, interactions with other molecule types (i.e. small molecules, nucleic acids, lipids), and adding further target properties that could be of interest for other applications.

Online Methods

Data set

To train our DL models, we used a large publicly available data set of estimated pK values – the pKPDB database[28]. This data set of ∼3M pK_a values was created by running the PypKa tool with default parameters[6] over all the protein structures deposited on the Protein Data Bank. The target values to be fitted by our model are theoretical pK_{half} values estimated with a PB-based method. This implies that pKAI will inherit the assumptions and limitations of this class of predictors. Our approach contrasts with the one usually adopted for training empirical predictors, which entails using experimental values to fit the model's parameters. The main advantage of this novel approach is that we can train models with significantly more parameters, such as deep learning ones, since there is now a much larger abundance of training data. As a comparison, in PROPKA3 only 85 experimental values of aspartate and glutamate residues were used to fit 6 parameters[5]. Recently, traditional ML models have been trained on ∼1k experimental pK_a values[36,37]. However, testing the real world performance of such methods is difficult as there is a high degree of similarity among available experimental data. Our larger data set translates into more diversity in terms of protein and residue types and, more importantly, a wider variety of residue

environments. It also helps our models to steer away from the undesired overfitting. Furthermore, the relation between a structure and our target property is deterministic, contrary to that of experimental pK_a values, which suffers from the lack of entropic information.

The ultimate goal of these methods is to accurately predict experimental pK_a values and thus, we have assessed the model's performance with ∼750 experimental pK_a values taken from the largest compilation of experimentally determined pK_a values of protein residues reported in the literature – the PKAD database[38]. We compare our experimental results with a null model (attributing to each titratable group the corresponding pK_a value in water), PypKa (the method used to generate the training set) and PROPKA with default settings (the empirical method of reference).

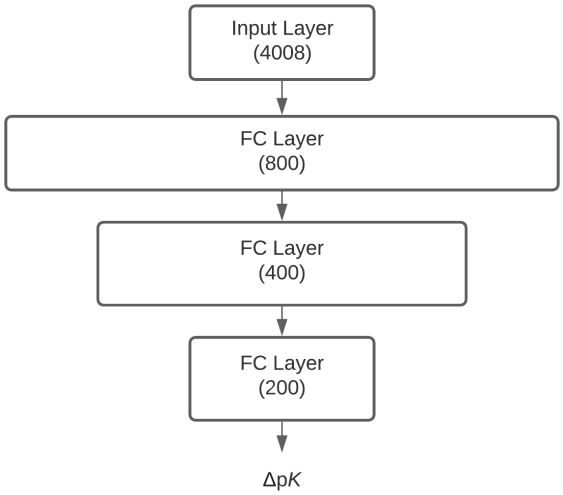
Before training our models on our data set, we applied a curated data split (Table 5A) to ensure that the training, validation, and test sets did not contain proteins with a high degree of similarity and prevent overfitting. First, we randomly selected 3k proteins from the full data set of ∼120k proteins as our holdout test set of theoretical pK_a values. The program mmseqs[39] was then used to exclude all proteins containing at least one chain similar to any of the chains found either in the experimental or in the theoretical test sets. Chains were considered to be similar if they presented sequence identity over 90%. From the remaining set of proteins, 3.000 more were randomly assigned to the validation set while the rest became the training set. Finally, we have excluded similar proteins to those of the validation set from the training set. In the experimental data set, we have excluded all duplicated proteins, non-exact pK_a values (e.g. >12.0), and residues for which PypKa or PROPKA failed to produce an estimate.

Model architecture and implementation

pKAI is implemented and trained using PyTorch v1.9.0[40] and PyTorch Lightning v1.2.10[41]. The model has a simple architecture comprised of 3 fully-connected hidden layers in a pyramidal configuration fitted to the pK_a shifts of titratable amino acids (Figure 5B). The simplicity of the architecture is intentional. pKAI is meant to serve as a proof-of-concept that deep learning models can capture the effect of complex electrostatic interactions. Describing such interactions would require at least 2 PB calculations per residue state for the physics-based counterpart (e.g. in PypKa each carboxylic acid has 5 states, hence 10 PB calculations are required for each Asp/Glu residue).

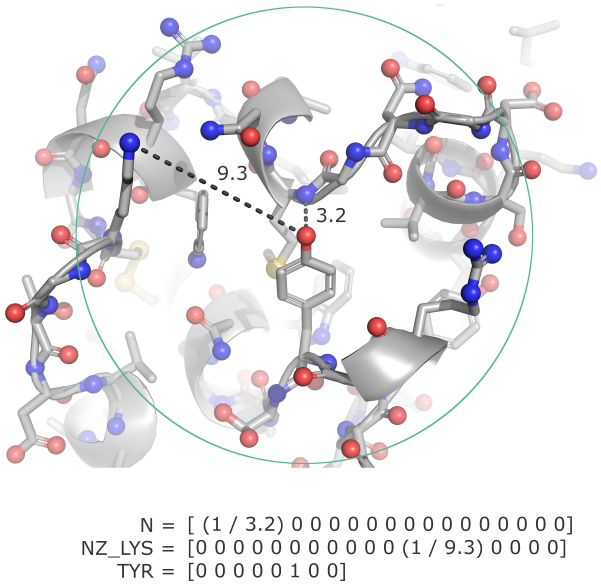
The encoding of the environment of each titratable residue has been simplified to retain only the most important electrostatic descriptors (Figure 5C). Considering the decay rate of the electrostatic potential, we decided to truncate the contributions to the environment of a residue by applying a cutoff of 15 Å around the labile atom(s) of the titratable residue. In practice, this cutoff is slightly smaller for some residue

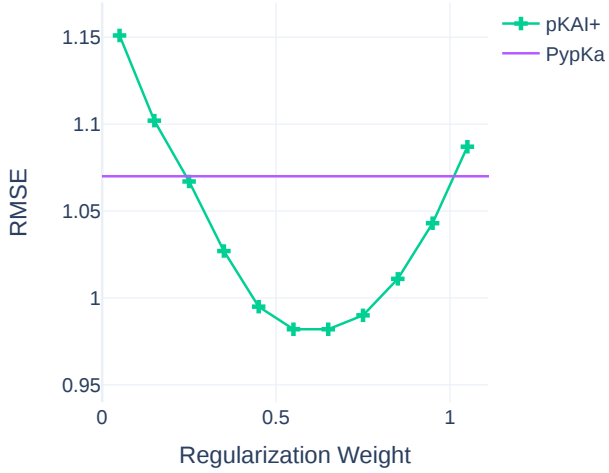
Split	Proteins	pK values
All Theor.	116.2k	12.6M
Train	56.8k	6.3M
Validation	3.0k	322.4k
Test Theor.	3.0k	325.3k
All Exp.	157	1350
Test Exp.	97	736



a)

b)





c)

d)

Figure 5. A) Overview of the data split and similarity exclusion performed on the pKPDB and PKAD databases[28,38]. B) pKAI model architecture. C) Illustration of the titratable amino acid environment encoding. Only Nitrogen, Oxygen and Sulfur atoms (shown as spheres) within a 15 Å cutoff (green circle) are included while all carbon (shown as sticks) and hydrogens (omitted) are ignored. The included atoms are represented by the inverse of their distance to the titratable residue in a OHE vector featuring 16 categories of atom classes (Supplementary Table S5). The titratable residue is represented by a OHE vector comprised of 8 classes. D) pKAI+ regularization weight performance test.

environments as the necessary input layer size normalization resulted in the truncation of the closest 250 atoms. A further approximation was made by considering only highly charged atoms and assuming they are the ones that contribute the most to pK_a shifts. This simplification can be slightly compensated by using atom classes instead of charges or element names as they implicitly provide information about adjacent atoms. The atoms were one-hot encoded (OHE) and, in order to reduce the input layer size, chemically similar atoms were assigned to the same category (Supplementary Table S5). While carboxylic oxygen atoms (C-termini OXT; aspartates OD1 and OD2; glutamates OE1 and OE2) and primary amine atoms (arginines NH1 and NH2) atoms have been merged, others with similar names but different chemical properties were separated (glutamines OE1 and NE2 from glutamates OE1 and histidines NE2, asparagines OD1 from aspartates OD1; main chain N from N-termini N).

The final 4008-sized input layer consisted of 250 atoms represented by 16 categories OHE classes concatenated to an 8-dimension OHE vector corresponding to the titrating amino acid. Each atom's OHE was multiplied by its reciprocal distance to the titrating residues so that this valuable information could be included without increasing the size of the input layer.

pKAI is freely available as a python module that can be installed via pip. The source code can be found at https://github.com/bayer-science-for-a-better-life/pKAI.

Training

Training was performed with mini-batches of 256 examples and the Adam optimizer[42] with a learning rate of $1e^{-6}$ and weight decay of $1e^{-4}$. Dropout regularization was applied to all fully-connected layers with the exception of the last one. Hyper-parameter optimization was performed with Optuna[43] using the performance in the validation set. Training these models takes approximately 10 minutes on an NVIDIA Tesla M40 24GB, using 16bit precision and an early stopping strategy on the minimization of the cost function with a delta of $1e^{-3}$ and patience of 5 steps.

The pKAI model was trained on an MSE cost function while for the pKAI+ we have added a regularization parameter α to penalize ΔpK_a predictions (y). Thus, the loss function of pKAI+ becomes

$$J(y_i, \hat{y}_i, \alpha) = (1-\alpha)(y_i - \hat{y}_i)^2 + \alpha \hat{y}_i^2 \quad (1)$$

where y_i is the true value and $\hat{y}_i$ the estimation. Different regularization weights were tested on the validation set to check for overfitting (Figure 5D). While we have selected an α of 50%, any value in the 40–70% range would lead to a similar improvement.

XAI Methods

For each input atom feature $\widehat{a} = (a, r_a)$, where a indicates the atom class and r_a the corresponding distance to the liable atom(s) of the titrating residue, we compute the corresponding attribution $I(\widehat{a})$ with the Integrated Gradients (IG) algorithm,[44] as implemented in the `shap` package[45]. $I(\widehat{a})$ measures the sensitivity of the network output with respect to changes in the input $\widehat{a}$. A large absolute value of $I(\widehat{a})$ indicates that the network assigns high importance to this feature, while the sign of $I(\widehat{a})$ indicates whether the feature contributes positively or negatively to the output. Given that the most important contributions to the $\Delta \mathrm{p}K_\mathrm{a}$ are of electrostatic nature, one can try to explain the model inferred charges for each atom class a by computing the distant-independent score C:

$$C(a) = \mathbb{E}\left[r_a^{-1} I_-(\widehat{a})\right] - \mathbb{E}\left[r_a^{-1} I_+(\widehat{a})\right], \tag{2}$$

where I_- and I_+ are negative and positive I values, respectively. The C score of an atom class is thus the difference between the distance weighted average of examples with negative and positive I values, over a large subset (10000 samples) of the test set. The sign of $C(a)$ in equation 2 resembles the charge that the network, on average, assigns to a given atom type. For example, if an atom class is being perceived by the model as contributing negatively to the $\Delta \mathrm{p}K_\mathrm{a}$ ($\mathbb{E}\left[r_a^{-1} I_-(\widehat{a})\right] > \mathbb{E}\left[r_a^{-1} I_+(\widehat{a})\right]$ hence $C(a) > 0$), this would mean that the network learned that this particular atom stabilizes the deprotonated state, which is characteristic of positively charged groups.

The solvent accessible surface area (SASA) values shown in Supplementary Table S1 and in the XAI subsection have been taken from pKPDB[28].

References

1. Warshel, A. & Åqvist, J. Electrostatic energy and macromolecular function. *Annu. Rev. Biophys. Biophys. Chem.* **20**, 267–298 (1991).
2. Kim, J., Mao, J. & Gunner, M. Are acidic and basic groups in buried proteins predicted to be ionized? *J. Mol. Biol.* **348**, 1283–1298 (2005).
3. Anandakrishnan, R., Aguilar, B. & Onufriev, A. V. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.* **40**, W537–W541 (2012).
4. Georgescu, R. E., Alexov, E. G. & Gunner, M. R. Combining conformational flexibility and continuum electrostatics for calculating pKas in proteins. *Biophys. J.* **83**, 1731 – 1748, DOI: https://doi.org/10.1016/S0006-3495(02)73940-4 (2002).

5. Olsson, M. H., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: consistent treatment of internal and surface residues in empirical pK_a predictions. *J. Chem. Theory Comput.* **7**, 525–537 (2011).

6. Reis, P. B. P. S., Vila-Viçosa, D., Rocchia, W. & Machuqueiro, M. PypKa: A flexible python module for poisson–boltzmann-based pK_a calculations. *J. Chem. Inf. Model.* **60**, 4442–4448, DOI: 10.1021/acs.jcim.0c00718 (2020). PMID: 32857502.

7. Wang, L., Zhang, M. & Alexov, E. DelPhiPKa web server: predicting pK_a of proteins, rnas and dnas. *Bioinformatics* **32**, 614–615 (2016).

8. Schutz, C. N. & Warshel, A. What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins: Struct. Funct. Genet.* **44**, 400–417 (2001).

9. Voges, D. & Karshikoff, A. A model of a local dielectric constant in proteins. *J. Chem. Phys.* **108**, 2219–2227, DOI: 10.1063/1.475602 (1998). https://doi.org/10.1063/1.475602.

10. Demchuk, E. & Wade, R. C. Improving the continuum dielectric approach to calculating pKas of ionizable groups in proteins. *J. Phys. Chem.* **100**, 17373–17387, DOI: 10.1021/jp960111d (1996). https://doi.org/10.1021/jp960111d.

11. Rocchia, W., Alexov, E. & Honig, B. Extending the applicability of the nonlinear Poisson–Boltzmann equation: multiple dielectric constants and multivalent ions. *J. Phys. Chem. B* **105**, 6507–6514 (2001).

12. Li, L., Li, C., Zhang, Z. & Alexov, E. On the dielectric "constant" of proteins: Smooth dielectric function for macromolecular modeling and its implementation in delphi. *J. Chem. Theory Comput.* **9**, 2126–2136, DOI: 10.1021/ct400065j (2013). PMID: 23585741, https://doi.org/10.1021/ct400065j.

13. Beroza, P. & Case, D. A. Including side chain flexibility in continuum electrostatic calculations of protein titration. *J. Phys. Chem.* **100**, 20156–20163 (1996).

14. Nielsen, J. E. & Vriend, G. Optimizing the hydrogen-bond network in Poisson–Boltzmann equation-based pK_a calculations. *Proteins Struct. Funct. Bioinf.* **43**, 403–412, DOI: 10.1002/prot.1053 (2001). https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.1053.

15. Baptista, A. M. & Soares, C. M. Some theoretical and computational aspects of the inclusion of proton isomerism in the protonation equilibrium of proteins. *J. Phys. Chem. B* **105**, 293–309 (2001).

16. Baptista, A. M., Teixeira, V. H. & Soares, C. M. Constant-pH molecular dynamics using stochastic titration. *J. Chem. Phys.* **117**, 4184–4200 (2002).

17. Mongan, J., Case, D. A. & McCammon, J. A. Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem.* **25**, 2038–2048 (2004).

18. Khandogin, J. & Brooks III, C. L. Toward the accurate first-principles prediction of ionization equilibria in proteins. *Biochemistry-US* **45**, 9363–9373 (2006).

19. Swails, J. M. & Roitberg, A. E. Enhancing conformation and protonation state sampling of hen egg white lysozyme using pH replica exchange molecular dynamics. *J. Chem. Theory Comput.* **8**, 4393–4404 (2012).

20. Vila-Viçosa, D., Reis, P. B. P. S., Baptista, A. M., Oostenbrink, C. & Machuqueiro, M. A pH replica exchange scheme in the stochastic titration constant-pH MD method. *J. Chem. Theory Comput.* **15**, 3108–3116, DOI: 10.1021/acs.jctc.9b00030 (2019).

21. Teixeira, V. H., Vila-Viçosa, D., Reis, P. B. & Machuqueiro, M. pK_a values of titrable amino acids at the water/membrane interface. *J. Chem. Theory Comput.* **12**, 930–934 (2016).

22. Vila-Viçosa, D., Campos, S. R. R., Baptista, A. M. & Machuqueiro, M. Reversibility of prion misfolding: insights from constant-pH molecular dynamics simulations. *J. Phys. Chem. B* **116**, 8812–8821 (2012).

23. Morrow, B. H., Koenig, P. H. & Shen, J. K. Atomistic simulations of pH-dependent self-assembly of micelle and bilayer from fatty acids. *J. Chem. Phys.* **137**, 194902–194902 (2012).

24. Swails, J. M. *et al.* pH-dependent mechanism of nitric oxide release in nitrophorins 2 and 4. *J. Phys. Chem. B* **113**, 1192–1201, DOI: 10.1021/jp806906x (2009).

25. Reis, P. B., Vila-Viçosa, D., Campos, S. R., Baptista, A. M. & Machuqueiro, M. Role of counterions in constant-pH molecular dynamics simulations of PAMAM dendrimers. *ACS Omega* **3**, 2001–2009 (2018).

26. Stanton, C. L. & Houk, K. N. Benchmarking pKa prediction methods for residues in proteins. *J. Chem. Theory Comput.* **4**, 951–966, DOI: 10.1021/ct8000014 (2008). PMID: 26621236.

27. Lee, A. C. & Crippen, G. M. Predicting pK_a. *J. Chem. Inf. Model.* **49**, 2013–2033, DOI: 10.1021/ci900209w (2009). PMID: 19702243.

28. Reis, P. B. P. S., Clevert, D.-A. & Machuqueiro, M. pKPDB: a protein data bank extension database of pKa and pI theoretical values. *Bioinformatics* DOI: 10.1093/bioinformatics/btab518 (2021).

29. Thurlkill, R. L., Grimsley, G. R., Scholtz, J. M. & Pace, C. N. pK values of the ionizable groups of proteins. *Protein Sci.* **15**, 1214–1218 (2006).

30. Grimsley, G. R., Scholtz, J. M. & Pace, C. N. A summary of the measured pK values of the ionizable groups in folded proteins. *Protein Sci.* **18**, 247–251 (2009).

31. Song, Y., Mao, J. & Gunner, M. Mcce2: improving protein pK_a calculations with extensive side chain rotamer sampling. *J. Comput. Chem.* **30**, 2231–2247 (2009).

32. Cai, Z., Luo, F., Wang, Y., Li, E. & Huang, Y. Protein pka prediction with machine learning. *ACS Omega* **6**, 34823–34831, DOI: 10.1021/acsomega.1c05440 (2021).

33. Huang, Y., Harris, R. C. & Shen, J. Generalized born based continuous constant ph molecular dynamics in amber: Implementation, benchmarking and analysis. *J. Chem. Inf. Model.* **58**, 1372–1383, DOI: 10.1021/acs.jcim.8b00227 (2018). PMID: 29949356.

34. Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589, DOI: 10.1038/s41586-021-03819-2 (2021).

35. Onufriev, A. V. & Alexov, E. Protonation and pK changes in protein–ligand binding. *Q. Rev. Biophys.* **46**, 181–209, DOI: 10.1017/S0033583513000024 (2013).

36. Gokcan, H. & Isayev, O. Prediction of protein pka with representation learning. *Chem. Sci.* **13**, 2462–2474, DOI: 10.1039/D1SC05610G (2022).

37. Chen, A. Y., Lee, J., Damjanovic, A. & Brooks, B. R. Protein pka prediction by tree-based machine learning. *J. Chem. Theory Comput.* **0**, null, DOI: 10.1021/acs.jctc.1c01257 (0).

38. Pahari, S., Sun, L. & Alexov, E. PKAD: a database of experimentally measured pK_a values of ionizable groups in proteins. *Database* **2019** (2019).

39. Mirdita, M., Steinegger, M. & Söding, J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* **35**, 2856–2858, DOI: 10.1093/bioinformatics/bty1057 (2019).

40. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In Wallach, H. *et al.* (eds.) *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019).

41. Falcon, W. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning* **3** (2019).

42. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic gradient descent. In *ICLR: International Conference on Learning Representations*, 1–15 (2015).

43. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2019).

44. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks (2017). 1703.01365.

45. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems 30*, 4765–4774 (Curran Associates, Inc., 2017).

Acknowledgements

We would like to thank Paulo Martel and Diogo Vila-Viçosa for the fruitful discussions as well as the attendees of the Protein Electrostatics meeting (www.proteinelectrostatics.org). We thank Artemi Bendandi for proofreading the manuscript.

PR and MM acknowledge financial support from FCT: SFRH/BD/136226/2018, CEECIND/02300/2017, UIDB/04046/2020, and UIDP/04046/2020. This work benefited from services and resources provided by the EGI-ACE project (receiving funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101017567), with the dedicated support from the CESGA and IN2P3-IRES resource providers. MB and FM acknowledge funding from the Bayer AG Life Science Collaboration ("Explainable AI").

Author Contributions

PR, FM and MB planned the experiments. PR performed the implementation of pKAI and analysis. MB performed the explainability experiments. PR, FM and MM conceived and designed pKAI. PR, MM, FM and MB wrote the paper with the help of DC and WR.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- si.pdf