Accurate and Rapid Prediction of Protein pK_a: Protein Language Models Reveal the Sequence-pK_a Relationship

Shijie Xu[†] and Akira Onoda[*,†,‡]

[†]*Graduate School of Environmental Science, Hokkaido University, Sapporo, Japan*

[‡]*Faculty of Environmental Earth Science, Hokkaido University, Sapporo, Japan*

E-mail: akira.onoda@ees.hokudai.ac.jp

Abstract

Protein pK_a prediction is a key challenge in computational biology. In this study, we present pKALM, a novel deep learning-based method for high-throughput protein pK_a prediction. pKALM uses a protein language model (PLM) to capture the complex sequence-structure relationship of proteins. While traditionally considered a structure-based problem, our results show that a PLM pre-trained on large-scale protein sequence databases can effectively learn this relationship and achieve state-of-the-art performance. pKALM accurately predicts the pK_a values of six residues (Asp, Glu, His, Lys, Cys, and Tyr) and two termini with high precision and efficiency. It performs well at predicting both exposed and buried residues, which often deviate from standard pK_a values measured in solvent. We demonstrate a novel finding that predicted protein isoelectric points (pI) can be used to improve the accuracy of pK_a prediction. High-throughput pK_a prediction of the human proteome using pKALM achieves a speed of 4,965 pK_a predictions per second, which is several orders of magnitude faster than existing state-of-the-art methods. The case studies illustrate the effi-

cacy of pKALM in estimating pK_a values and the constraints of the method. pKALM will thus be a valuable tool for researchers in the fields of biochemistry, biophysics, and drug design.

1 Introduction

Protein pK_a is crucial for understanding various biological processes, including enzyme catalysis,[1–3] protein folding,[4–6] and ligand binding.[7,8] Chemical groups in the side chain and terminals can be protonated or deprotonated depending on the pH of the environment, affecting protein structure, stability, and function. Currently, the most reliable pK_a values of proteins are only obtained through laborious, costly, and challenging experiments such as NMR titration.[9] Computational methods for the prediction of protein pK_a have been developed to provide relatively fast and convenient pK_a estimations while maintaining reasonable accuracy. One widely used method is the empirical approach, PropKa,[10,11] which employs a set of human-designed rules to predict pK_a values based on a single protein structure. PropKa is considered one of the fastest in protein pK_a prediction due to the efficient computation, though its accuracy is limited by the simplicity of its rules. In contrast, constant pH molecular dynamics (CpHMD) simulation methods,[12–21] based on the molecular dynamics (MD) simulations, are considered to be the most accurate computational methods for pK_a prediction. Recent advancements in CpHMD methods have been seamlessly incorporated into simulation platforms such as GROMACS,[22] CHARMM,[23] and AMBER,[24] facilitating both accurate predictions and user-friendly application.[25] However, they require a long-time sampling of different protein conformations and protonation states, making them computationally intensive and unsuitable for high-throughput pK_a prediction. Furthermore, Poisson-Boltzmann (PB) methods,[26–28] based on continuum electrostatics theory,[29] provide the pK_a values for ionizable residues in proteins. PB methods reduce computational costs compared to CpHMD simulations, though at the expense of accuracy. This accuracy decreases due to theoretical assumptions on the basis of modeling the protein-solvent interface

as a continuum medium[14] and single protein conformations in the calculations.

Protein pK_a prediction is typically approached as a structure-based problem, where the protein structure serves as input features of predictors. However, several issues arise with this approach, especially for high-throughout prediction: 1) Protein structures are not always available for all proteins, limiting the applicability of structure-based methods. Although structure predictors such as AlphaFold2[30] and RoseTTAFold[31] can provide highly accurate atomic-level structures, their high computational cost makes them unsuitable for high-throughput pK_a prediction. Furthermore, studies indicate that the pK_a prediction based on structures predicted by AlphaFold2 tend to underestimate of the pK_a shifts,[32] rending predictions for significantly shifted pK_a values less reliable. 2) Protein structures are inherently flexible and dynamic, and specific regions can be ambiguous,[33] affecting prediction accuracy. Some pK_a predictors use clustering of different conformations to enhance prediction accuracy.[34] However, this clustering process is time-consuming and does not fully address the aforementioned structural discrepancy issue. These discrepancies between structures[35] pose significant challenges to the prediction accuracy of the existing methods. These challenges highlight the need for improved methods in protein pK_a prediction that can account for structural discrepancies, flexibility, and the availability of protein structures.

Machine learning (ML) methods have recently emerged as a promising approach for protein pK_a prediction. ML models such as random forests,[36] graph neural networks,[37] and multilayer perceptron,[38] have been applied to predict pK_a values based on protein structures.[34,39–41] These ML methods are built with numerous parameters and trained on data sets consisting experimentally determined pK_a values[42,43] or pK_a values derived from other computational methods.[32,44] Due to their flexible architectures and scalability, these methods effectively model the complex relationship between the protein structures and pK_a values of protein residues. ML methods achieve both high efficiency and high accuracy in pK_a prediction compared to traditional methods.

In recent years, large language models (LLMs) have achieved significant success in vari-

ous modeling tasks, including natural language processing (NLP) and computer vision (CV). Protein language models (PLMs) have been also developed using architectures such as transformers[45,46] and the recurrent neural network (RNN).[47] In a PLM, each residue in a sequence is represented as a high-dimensional vector encapsulating the essential information of the sequences and its context. For instance, residues in similar protein environments are likely represented by proximate vectors in the high-dimensional space. PLMs have demonstrated their effectiveness in various protein property prediction tasks, including secondary structure prediction,[48] contact prediction,[45] atomic-level structure prediction,[46] and intrinsic disorder prediction,[49] in terms of prediction accuracy and computational efficiency.

This study introduces pKALM, a novel protein pK_{a} prediction method based on a protein language model. pKALM is a deep learning-based approach capable of predicting pK_{a} values solely from protein sequences. It has been trained to predict the pK_{a} values of the side chains of Asp, Glu, His, Lys, Cys, Tyr, as well as the N-terminus and C-terminus of proteins. The performance of pKALM was evaluated across a range of residue types, pK_{a} shifts and burial depths. The results demonstrated that pKALM achieved comparable performance to state-of-the-art methods. Two isoelectric point (pI) predictors were also developed, achieving state-of-the-art performance, and integrated into the pKALM framework to improve pK_{a} prediction accuracy. Furthermore, we demonstrate the high-throughput prediction of pK_{a} values for all proteins in the human proteome, emphasizing the significantly faster prediction capabilities of pKALM. A case study is also provided to elucidate the reasons why sequence-based pKALM can effectively predict pK_{a} values, with the interpretability of the visualized attention maps of the PLM.

2 Method

2.1 Data sets

We utilized the recently released PKAD-2 data set[43] to train and test various protein pK_a prediction methods. PKAD-2 compiles experimental pK_a data from multiple references, providing 1,456 pK_a values for wild-type proteins and 269 pK_a values for mutant proteins. We meticulously revised the original data set to correct missing or erroneous information, excluding unverifiable or non-numerical pK_a values. The revised data set now includes 1,450 pK_a values for 165 wild-type proteins and 262 pK_a values for 47 mutant proteins. The revised PKAD-2 is available in the associated GitHub repository `https://github.com/xu-shi-jie/revPKAD2`.

The PKAD-2 data set exhibits considerable redundancy, with homologous proteins having very similar pK_a values for the same residues. This redundancy presents two critical issues: 1) it causes overlap between training and test sets, compromising the fairness of ML model evaluations, and 2) it hampers the generalization of ML models to novel, non-homologous proteins. To address these issues, we utilized the cd-hit program[50] to cluster the revised PKAD-2 data set at 30% sequence similarity, a critical threshold (compared to 30% in the work of Cai *et al.*[40] and 90% in the work of Reis *et al.*[41]). The resulting non-redundant data set was randomly divided into training and test sets in a 5:2 ratio, with 376 pK_a values in the training set and 152 pK_a values in test set. Notably, the size of test set closely aligns with those used in related studies.[40]

The literature bias results in an extremely unbalanced distribution of pK_a values among different residues. Predictions for less abundant residues (Cys, Tyr, N-terminus, and C-terminus) are more challenging compared to abundant residues (Asp, Glu, His, and Lys). Table 1 presents the pK_a values for both abundant and less abundant residues. The detailed data set construction process is described SI Section 1, and the pK_a value distribution are shown in SI Figure S1. To compare pKALM with CpHMD methods, we leveraged a test set

"EXP67S" established in prior research.[51] It contains 167 pK_a values and we also removed redundancies with more than 30% sequence similarity to our training set via the mmseqs2 program.[52] The resulting reduced test set, termed "rEXP67S", comprises 96 pK_a values, with their detailed distribution depicted in SI Figure S2.

Table 1: The numbers of pK_a values of different residues in the data set.

	Residue	Training set	Test set
Abundant	Asp	108	40
	Glu	121	44
	His	64	25
	Lys	38	15
Less abundant	Cys	11	4
	Tyr	14	6
	N-terminus	8	4
	C-terminus	12	5
	Total	376	143

We utilized isoelectric point (pI) data from Kozlowski's work,[53] which includes both peptide and protein pI data sets. These data sets, derived from references containing experimental pI values, comprise 119,092 peptide pI values and 2,324 protein pI values. Redundancies in both data sets were removed, and they were split into training and test sets in a 3:1 ratio. SI Figure S3 illustrates the detailed distribution of pI values. Notably, compared to the peptide pI data, the protein pI data are significantly fewer and more concentrated, posing a greater challenge for accurate prediction.

2.2 Protein language model

Protein language models are large-scale neural networks pre-trained on extensive unlabeled protein sequences. Architectures such as to BERT[54] or T5[55] are commonly employed for PLMs, characterized by parameters counts ranging from millions to billions, each exhibiting distinct properties. In this study, we evaluated nine PLMs of varied architectures and sizes, including the ESM series[45,46] and ProtTrans series.[48] Detailed information are provided in SI Table S1.

The scaling law[56] of large language models demonstrates that their performance consistently improves with larger model and training data sizes. This finding aligns with previous research,[49] indicating that performance increases with PLM model size but eventually plateaus due to the limited fine-tuning data. Consequently, selecting appropriate PLMs is critical for tasks involving protein property prediction, contingent upon the scale of available fine-tuning data.

2.3 Architecture of pKALM

The architecture of pKALM predictor is depicted in Figure 1a. The input comprises the protein sequence, encoded through a protein language model, a protein pI model, and a peptide pI model to derive feature vectors. These models are held frozen during pKALM training. The concatenated feature vectors are then fed into a bi-directional long short-term memory (BiLSTM)[57] network to capture sequential patterns in the protein sequences. To distinguish the residue types (i.e., Asp, Glu, His, Lys, Cys, Tyr, N-terminus, and C-terminus), they are encoded via an embedding layer and added to the BiLSTM network outputs. The combined vectors are subsequently processed by a simple fully connected layer to predict the pK_a shift. The final predicted pK_a is derived from the sum of the pK_a shift and the standard pK_a, consistent with prior work,[11] as detailed in SI Table S2.

Furthermore, both pI models share identical architectures, as illustrated in Figure 1b. The input of each pI model includes either a protein or peptide sequence, encoded into vector sequences via a residue embedding layer. These vectors are then fed into a BiLSTM network to learn sequential patterns. From the BiLSTM outputs, only the hidden states corresponding to protonatable residues (i.e., Arg, Asp, Glu, His, Lys, Cys, and Tyr) and the termini are extracted and averaged to form the final feature vector. A fully connected layer converts this vector into the pI value. Each pI model is trained on distinct data sets: the peptide pI data set and the protein pI data set, respectively.

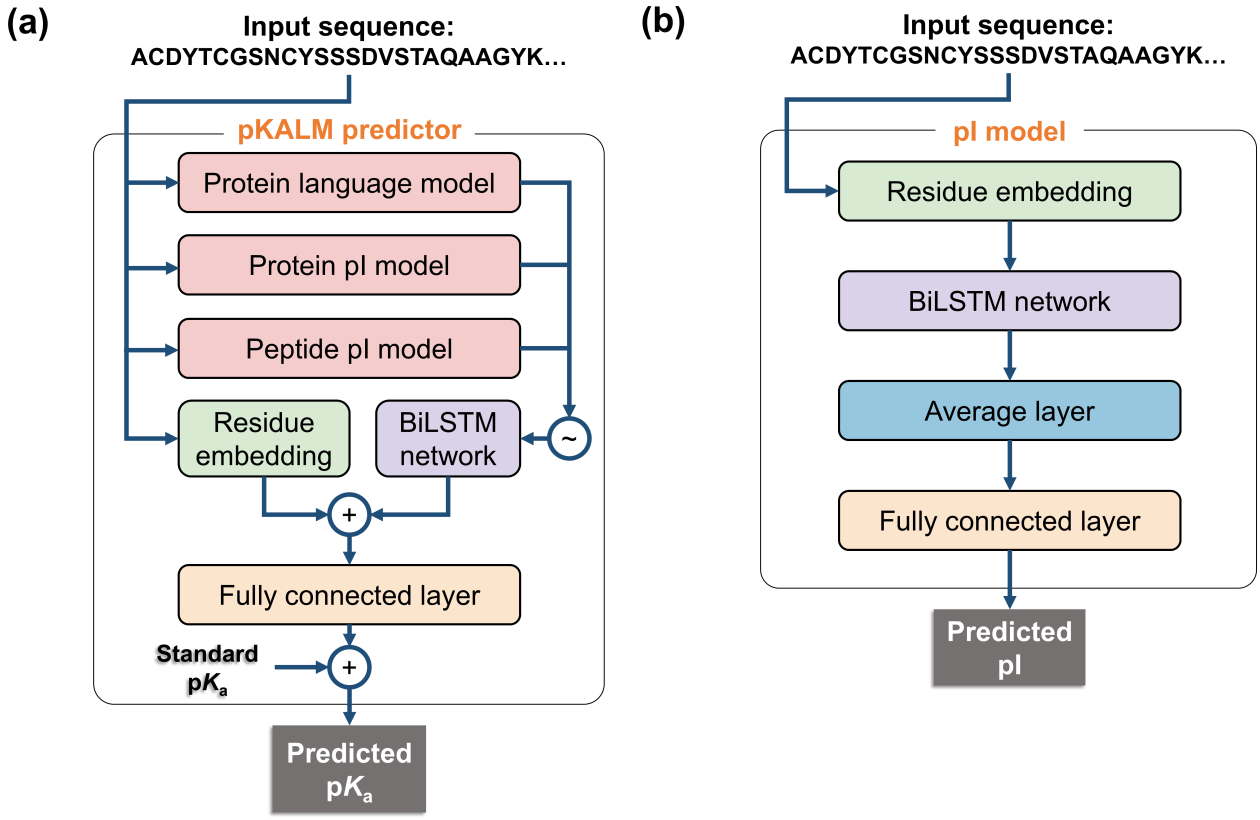

Figure 1: Overview of pKALM. (a) The architecture of pKALM. Arrows indicate data flow, while colored blocks represent different neural networks. The symbol "∼" denotes concatenation of feature vectors and "+" denotes vector addition. Flow bifurcation indicates data duplication. (b) Architecture of the pI predictors. The pI model is shared by the peptide pI model and protein pI model.

2.4 Compared methods and data preparation

We compare pKALM with five existing pK_a prediction methods including PropKa,[11] PKAI,[41] PypKa,[28] DeepKa,[40] and PKAI+.[41] For DeepKa predictions, we utilized the online service available at `http://www.computbiophys.com/DeepKa/main`, while for the other methods, we installed their respective software packages locally and assessed their performance. To test the performance of reported methods based on structures, protein structures within test sets were prepared as below. We retrieved PDB files from the Protein Data Bank.[58] For proteins lacking PDB entries (UniProt IDs: P03230 and P53378), we employed ColabFold[59] to predict their structures. Repairing the protein structures involved using openMM[60] to handle non-canonical residues (e.g. selenomethionine is replaced to methionine), mutation

of wild-type residues, and removal of HETATM records and water molecules.

As both pK_a and pI prediction involve regression tasks, we employed root-mean-square error (RMSE) and mean absolute error (MAE) to evaluate the performance of our method and others. RMSE, preferred for its sensitivity to data set outliers, served as the primary metric. Furthermore, we utilized RMSE, MAE, and the coefficient of determination (R^2) to comprehensively compare the accuracy of different pI prediction methods.

3 Results and discussion

3.1 Hyperparameter tuning and optimization

We utilize a grid search methodology alongside the `wandb` package[61] to optimize the hyperparameters of models in the experiments. Specifically, we adjust the dimensions of the hidden states and the layers within the BiLSTM networks for both the pKALM and two pI models. We explore hidden state dimensions and the number of layers in different ranges (see SI Table S3). The optimal hyperparameters are selected based on the performance evaluated through 5-fold cross-validation on the training data set.

We employ the AdamW optimizer[62] throughout our experiments, with a fixed learning rate of 0.001 and a weight decay of 0.001 to mitigate overfitting. Remaining optimizer hyperparameters are set to their default values. Furthermore, we utilize a cosine annealing learning rate scheduler. The batch sizes are set to 32, with model training spanning 100 epochs for protein pK_a prediction, 20 epochs for peptide pI prediction, and 10 epochs for protein pI prediction. Loss functions employed are mean-absolute-error (MAE) for pK_a prediction and mean-squared-error (MSE) for pI prediction.

All experiments are conducted on a Linux workstation equipped with an AMD Ryzen 9 7950X3D 16-core processor, 64 GB memory, and dual NVIDIA RTX 3090Ti GPUs. Implementation utilizes `Python` 3.12.2 and `PyTorch` 2.2.1 with CUDA 11.8. For the PLM inference, we utilize Hugging Face's `transformers` package (ver 4.39.3)[63] and `accelerate`

package (ver 0.28.0).[64] Except the native automatic mixed-precision feature of PyTorch,[65] no additional optimization techniques are employed to accelerate the model training and inference.

During the training step of a mini-batch, the protein sequences varied in length were pad and truncated into a fixed length for efficient batch processing. We segment longer sequences into smaller segments with a maximum length of 960 tokens (i.e., residues), due to the constraint that some PLMs, such as ESM1B_650M, only accept sequences shorter than 1024 tokens. To retain contextual information, an additional 31 tokens at the end of the segments are appended to each segment, facilitating overlap between adjacent segments, which will be discarded. Final sequence encodings are derived from the concatenated non-overlapping of segment encodings.

3.2 Prediction of peptide and protein isoelectric point

The isoelectric point (pI) of a protein is the pH at which the protein exhibits no net charge, while the pK_a values denote the pH at which protein residues are half-protonated. The relationship between the pK_a values of a small molecule and its isoelectric point (pI) can be elucidated through the Henderson-Hasselbalch equation for weak acids:[66]

$$\mathrm{pH} = \mathrm{p}K_\mathrm{a} + \log_{10}\left(\frac{[\mathrm{A}^-]}{[\mathrm{HA}]}\right)$$

The average charge of an ionizable residue at a specific pH can be calculated by $[\mathrm{A}^-]/[\mathrm{HA}] = 10^{\mathrm{pH}-\mathrm{p}K_\mathrm{a}}$. The total charge of the small molecule can be determined by summing the charges of all ionizable groups. The isoelectric point (pI) is thus the pH at which the total charge of the protein is zero. For instance, the pI of a small molecule containing both an anionic and a cationic group, with p$K_\mathrm{a}{}^1$ and p$K_\mathrm{a}{}^2$, can be calculated as $\mathrm{pI} = (\mathrm{p}K_\mathrm{a}{}^1 + \mathrm{p}K_\mathrm{a}{}^2)/2$.

The Henderson-Hasselbalch (HH) equation is not directly applicable to proteins, as they encompass numerous ionizable residues capable of mutual interaction. These interactions

further complicate their behavior, often giving rise to cooperativity. Consequently, the titration curve of interacting residues cannot be presumed to adhere strictly to the HH equation. Nevertheless, the HH equation offers a qualitative insight into the relationship between pK_a values and pI. This relationship, we posit, can be harnessed by machine learning models to enhance pK_a prediction accuracy, given the models' ability to discern and capture intricate and non-linear relationships.

This study initially trains peptide and protein pI predictors using certain data sets and integrates them into the pKALM framework. Performance comparisons against state-of-the-art methods for peptide pI prediction are summarized in Table 2. Our peptide pI predictor exhibits superior performance in terms of RMSE, MAE, and R^2 compared to IPC2, the previous leading method trained on identical data sets. 5-fold cross-validation results for peptide pI prediction are illustrated in SI Figure S4, with the optimal hyperparameters selected.

Table 2: Performance comparison of different methods for peptide isoelectric point prediction. Bold font indicates our method.

Method	RMSE	MAE	R^2
pKALM, this work	**0.2204**	**0.1071**	**0.9765**
IPC2[53]	0.2216	0.1216	0.9761
Bjellqvist[67]	0.4051	0.2836	0.9204
Nozaki[68]	0.4083	0.2673	0.9191
DTASelect[69]	0.4235	0.2796	0.9130
Thurlkill[70]	0.4466	0.2535	0.9033
Sillero[71]	0.4747	0.2696	0.8907
Dawson[72]	0.4910	0.2642	0.8831
Wikipedia [a]	0.5178	0.2974	0.8700
Grimsley[73]	0.5264	0.3796	0.8656

[a]The Wikipedia method refers to the pK_a values as presented in the Wikipedia page in 2005.

We develop a protein pI predictor and evaluate its performance against existing benchmarks. Performance comparisons against state-of-the-art methods for protein pI prediction are summarized in Table 3. Our method achieves the third-best RMSE and R^2 scores, and the fourth-best MAE among the methods evaluated. Notably, our approach employs deep

learning, which typically demands larger data sets for excellent performance. The limited size of available protein pI data sets (see Section "Data set") constrains the efficacy of our predictor despite employ a comparable architecture. 5-fold cross-validation results for protein pI prediction are depicted in SI Figure S5, with the optimal hyperparameters selected.

Table 3: Performance comparison of different methods for protein isoelectric point prediction. Bold font indicates our method.

Method	RMSE	MAE	R^2
IPC2[53]	0.8479	0.5906	0.5934
IPC[74]	0.8677	0.6109	0.5760
pKALM, this work	**0.9069**	**0.6468**	**0.5589**
ProMoST[75]	0.9113	0.6444	0.5183
Toseland[76]	0.9278	0.6537	0.5095
Dawson[72]	0.9365	0.6586	0.4977
Bjellqvist[67]	0.9369	0.6536	0.5005
Wikipedia[a]	0.9484	0.6795	0.4860
Rodwell[77]	0.9579	0.6762	0.4706
Grimsley[73]	0.9588	0.6953	0.4779

[a]The Wikipedia method refers to the pK_a values as presented in the Wikipedia page in 2005.

3.3 Performance of different protein language models

We assess the performance of multiple protein language models for predicting pK_a values, as depicted in the blue and green curves in Figure 2. The evaluation includes validation and test RMSE for nine PLMs: eight PLM-based configurations (with "ESM" or "PT" prefixes) and a compared configuration replacing PLM with an embedding layer ("No PLM"). The points in the curves representing the best configurations selected from validated RMSEs, which consistently decreases with an increase in model parameters from 8 to 650 millions, and eventually achieve the lowest for three largest PLMs: ESM2_650M, ESM2_3B, and ESM2_15B. Other PLMs (ESM1B_650M, PT_BFD, PT_UR), and the non-PLM configuration exhibit varying performances due to their distinct architectures and training data sets from ESM2 series, and their details can be found in the original literature.[45,48]

The test RMSEs, however, diverge notably from the validation RMSEs. We define the

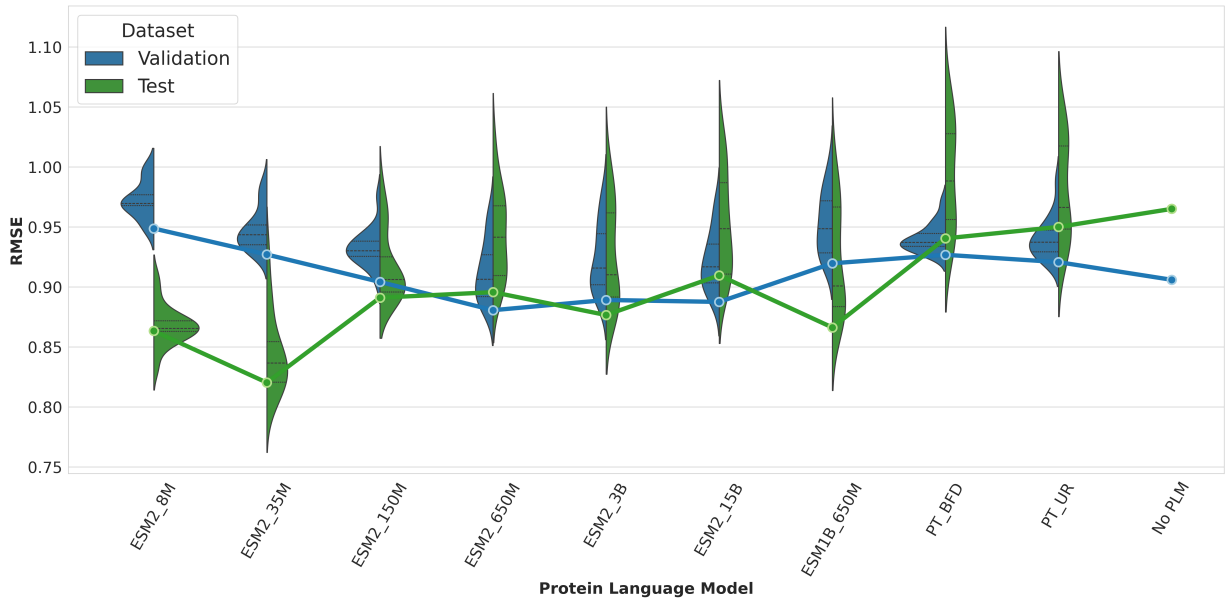

Figure 2: Comparison of the performance of different protein language models on the validation and test sets for all models and the best model selection. The violin graph illustrates the distribution of (Validation/Test) RMSEs for different PLMs with different hyperparameters (hidden states and layers). The curves represent the best (Validation/Test) RMSEs selected based on the validation set. Note that for the non-PLM configuration, the distribution of RMSEs across different configurations is very concentrated and we only show the best RMSE.

pK_a shift as the difference between model-predicted and experimentally determined pK_a, categorizing shifts into five ranges split by 0.5 pH units: 0-0.5, 0.5-1.0, 1.0-1.5, 1.5-2.0, and >2.0. SI Figure S6 illustrates the test RMSE of different PLMs across these ranges. Notably, larger PLMs (ESM1B_650M, ESM2_3B, and ESM2_15B) exhibit superior performance when pK_a shifts are small (<1.0 pH unit). In contrast, for shifts larger than 1.5, the configurations with smaller PLMs (ESM2_8M and ESM2_35M), demonstrate better performance. These findings suggest that larger PLMs may suffer from overfitting when predicting pK_a values with significant shifts while smaller PLMs may not. Thus, selecting an appropriate PLM is crucial to balance the prediction accuracy of pK_a values that shift significantly and those close to the standard pK_a values.

The distribution of RMSEs for small-size PLMs is more concentrated, while the gaps between the validation and test RMSEs are larger. In contrast, large-size PLMs exhibit less

concentrated RMSEs with smaller gaps between validation and test RMSEs, as shown in the violin graph in Figure 2. Overall, the variance in performance across different PLMs and hyperparameters underscores the importance of selecting an appropriate PLM for protein pK_a prediction and highlights the necessity of hyperparameter tuning for optimal performance. Notably, our choice of PLM is considered *a priori* knowledge rather than a hyperparameter. For subsequent experiments, we prioritize the RMSE of four abundant residues (Asp, Glu, His, and Lys) as the primary metric for selecting the best PLM. ESM2_35M achieves the best performance in terms of test RMSE among the evaluated models.

3.4 pKALM achieves state-of-the-art performance

We evaluate the performance of our pKALM method against reported methods (Figure 3a and 3b). The results demonstrate that all methods can predict pK_a values for Asp, Glu, His, and Lys. Importantly, pKALM achieved the lowest test RMSE on three abundant residues (Asp 0.6891, His 1.0343, Lys 0.8653) and third-best on Glu (0.7171). PKAI+, another deep learning-based method, showed the second-best performance, followed by DeepKa, PypKa, and PKAI. The empirical method PropKa performs the worst, due to its simplistic empirical rules and limited model complexity. The predictions of Glu residues exhibited relatively small test RMSE among all predictors due to the concentrated pK_a values of glutamic acid. In contrast, the predictions of His residues had the highest RMSE, due to the limited and diverse distribution of pK_a values in the data set. Asp residues, with the second largest data (108/40 in training/test set) and a broader pK_a distribution, provided a fair and challenging to benchmark. Performance comparisons on different residues in terms of RMSE and MAE can be found with details in SI Tables S4 and S5. Overall, pKALM achieved the state-of-the-art performance compared to existing structure-based methods.

The pK_a values of less abundant residues Cys, Tyr, N-terminus, and C-terminus were predicted in superior performance in terms of RMSEs (Cys 2.173, Tyr 0.6599, N-terminus 0.7820, C-terminus 0.8539) by pKALM relative to existing methods, as illustrated in Figure

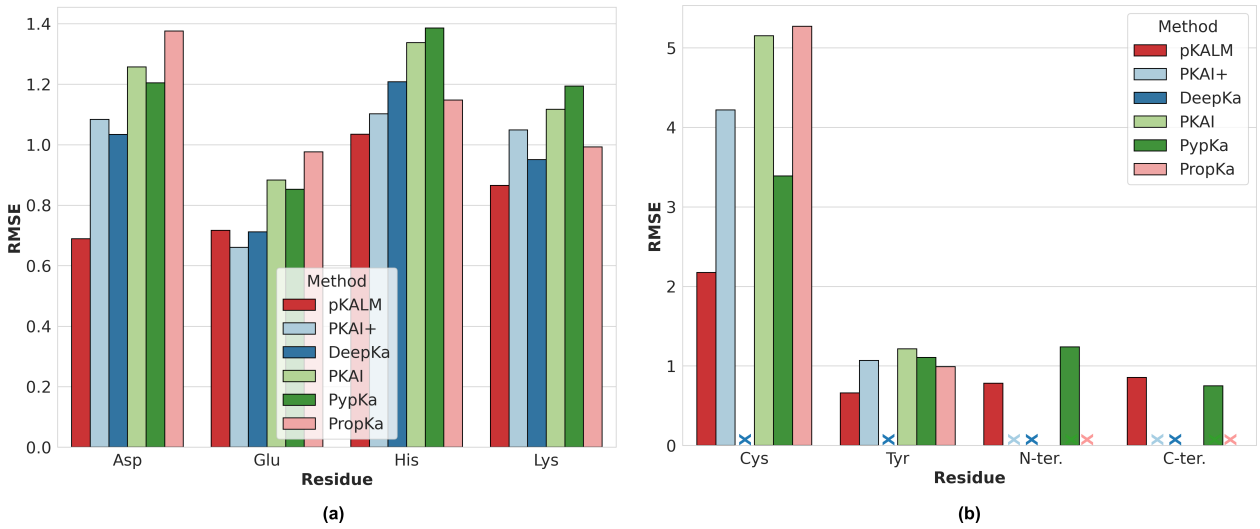

Figure 3: Performance comparison of pKALM with existing methods for (a) abundant residues: Asp, Glu, His, and Lys and (b) less abundant residues: Cys, Tyr, N-terminus (N-ter.), and C-terminus (C-ter.). Some methods are not available for certain residues, as indicated by "X".

3b. Predictions for Cys were particularly challenging due to the their limited data and very diverse pK_a values.

We compared pKALM with the available methods, marking unavailable ones with an "X". pKALM is unique in predicting pK_a values for Cys with an RMSE of approximately 2 pH units, whereas other methods showed significantly large RMSEs (PKAI+>4, PypKa>3, PKAI and PropKa>5), rending their predictions for Cys unreliable. Despite a limited test set, pKALM achieved the lowest RMSE for Tyr (14/6 in training/test set) and N-terminus (8/4 in training/test set). Detailed performance comparisons in terms of RMSE and MAE can be found in SI Tables S6 and S7. pKALM and PypKa provided predictions for the C-terminus with comparable accuracy (pKALM 0.8539 vs PypKa 0.7496). This indicates that traditional methods that do not rely on machine learning remain dispensable for residues lacking sufficient data to train machine learning models.

The Pearson correlation coefficients are also computed as shown in SI Figure S7 and S8. On Asp, His, and Lys, pKALM achieves the highest correlation coefficients among all methods. On Glu, pKALM is only slightly lower compared to others. This is consistent to

rank on RMSE and MAE scores. For less abundant residues, all methods performs poor correlation coefficients. This is because the most of the pK_a prediction methods are trained and optimized with L_1 or L_2 loss functions, which are not fit the correlation coefficients. In overall, we also compute the correlation coefficients on all residues and pKALM achieves the highest correlation coefficient among all methods, as shown in SI Figure S8.

Notably, the compared methods rely heavily on structural data and are highly sensitive to protein conformations. To fairly compare pKALM with these methods, we extend the comparison to CpHMD methods, which are not dependent on conformational sensitivity. As illustrated in Table 4, pKALM is more accurate than the CpHMD method on the test set. Our approach demonstrates a lower RMSE and MAE, and a higher Pearson correlation coefficient (R) than CpHMD. The RMSE of 0.8876 for pKALM is consistent with the performance of our test set, and the RMSE of 1.0139 for CpHMD is consistent with the previous study.[40] The scatter plot in SI Figure S9 shows the predicted pK_a shifts of pKALM and CpHMD against experimental pK_a shifts. While CpHMD exhibits several pronounced outliers that deviate significantly from experimental values, predictions from pKALM align more closely with the diagonal line, reflecting better performance on rEXP67S test set. It is worth noting the distribution of pK_a dataset, rEXP67S, is also balanced as shown in SI Figure S2.

Table 4: Comparison of pKALM and CpHMD methods on the rEXP67S test set.

Method	RMSE	MAE	R
pKALM, this work	0.8876	0.5372	0.8114
CpHMD	1.0139	0.7304	0.7700

3.5 Performance on pK_a shifts

We evaluate the performance of pK_a predictors on different pK_a shifts using PKAD-2 data set. This data set consist of pK_a values mostly within 1 pH unit of their standard values and ones significantly shifted (e.g., >2 pH units). Predicting small pK_a shifts is relatively

straightforward, while predicting large shifts is more challenging, necessitating evaluation across different pK_a shift ranges. We thus split the pK_a shifts into five ranges with 0.5 pH unit interval: 0-0.5, 0.5-1.0, 1.0-1.5, 1.5-2.0 and >2.0 and assessed the performance of pKALM and existing methods on these ranges (Figure 4a) according to the previous study.[40] The number of pK_a values in different pK_a shift ranges is shown in Table 5.

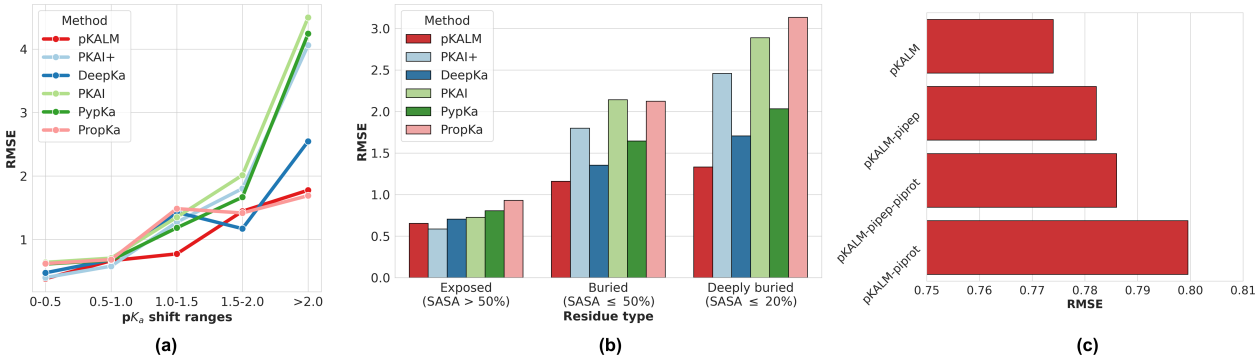

Figure 4: (a) Performance comparison of pKALM with reported methods for largely shifted pK_a values. (b) Performance comparison of pKALM with reported methods for exposed, buried, and deeply buried residues. (c) Ablation study of pKALM by removing the pI models.

Table 5: The distribution of pK_a values across various pK_a shift ranges in the test set utilized in this study.

Range	0-0.5	0.5-1.0	1.0-1.5	1.5-2.0	>2.0
Count	70	37	20	9	7

pKALM achieves the lowest RMSE in ranges ≤1.5 while the third best in the range of 1.5-2.0 and the second-best in the range of >2.0. It is worth noting that, another two deep learning-based methods, PKAI and PKAI+, also lag behind the best performance for larger ranges. Deep learning methods are trained for minimizing the overall loss function and may not always focus on learning the largely shifted pK_a values. Considering the large part of small pK_a shifts in the data set, it is reasonable that the deep learning methods are not the best for the largely shifted pK_a values. On the other hand, DeepKa is optimized for relatively balanced distribution of pK_a shifts and thus performs better than PKAI and PKAI+ for the largely shifted pK_a values. PropKa is the best one for the pK_a shifts larger than 2.0 pH units while the data point of this range is very limited. Furthermore, its performance on

small pK_a shifts is relatively low, suggesting such a method not reliable for the general pK_a prediction since we do not know which range the pK_a value belongs to before the prediction. In contrast, our method pKALM achieves balanced performance across different pK_a shift ranges.

Considering the less abundance of largely shifted pK_a values, we utilized Matthews correlation coefficient (MCC) as an indicator to evaluate the performance of predicting such pK_a values. The positive is defined as a large shift ($>$ 2 pH units) and vice versa. MCC is defined by the following equation:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively. Note that the MCC ranges from -1 to 1, where 1 indicates a perfect prediction, 0 indicates a random prediction, and -1 indicates a completely wrong prediction. The results are shown in SI Figure S10. pKALM gives high MCC scores on Asp and Cys, while zero for others, and no negative scores. Other methods have negative scores on Asp, His, Cys. PropKa has the only positive MCC on Asp. This suggests that pKALM is capable of predicting large shifted pK_a values with high accuracy.

3.6 Performance on the exposed and buried residues

We evaluate the performance of reported methods for exposed and buried residues separately. The definitions of exposed, buried, and deeply buried residues are based on the relative solvent accessibility (RSA) of the residues, which is already calculated in the PKAD-2 data set. We choose 50% as the threshold, same as the previous study.[39] To critically evaluate the performance of pKALM on deeply buried residues, we further introduce a new category of residues with RSA $<$ 20%. The results are shown in Figure 4b. It is clear that pKALM achieves the lowest RMSE for challenging buried and deeply buried residues. For exposed

residues, pKALM also achieves the second-lowest RMSE (0.6281), following PKAI+ (0.5862), while all predictors have relatively similar performance. This is reasonable since the exposed residues are more accessible to the solvent and thus their pK_a values are closer to the standard pK_a values, suggesting they are much easier to predict. It is worth noting that PKAI+ is the best for the exposed residues since it tends to attenuate the larger pK_a shift. Consequently, its performance on buried and deeply buried residues is relatively worse than pKALM, DeepKa, and PypKa.

3.7 Ablation study and comparison with existing methods

The necessity of adopting the pI models in pK_a prediction was confirmed by an ablation study of different modules (Figure 4c). To investigate the importance of modules inside a deep learning model, we remove the pI models one by one, retrain them with the same data sets, and evaluate them on the same test sets. Four versions of configurations are tested: 1) pKALM, 2) pKALM without peptide pI model "pKALM-pipep", 3) pKALM without protein pI model "pKALM-piprot", and 4) pKALM without both pI models "pKALM-pipep-piprot". We observe the changes of the test RMSE of abundant residues and the results show that the pI models are crucial for the pK_a prediction, though the version without any pI models are good. By removing the peptide pI model, the RMSE increases; removing the protein pI model, the RMSE also increases. When we remove only the protein pI model, the RMSE increases the most. This suggest that protein pI model is the most crucial for this architecture.

However, the impact of the pI models on largely shifted pK_a values and buried residues is less pronounced than on more abundant and exposed residues. As illustrated in SI Tables S6 and S7, the removal of the pI models does not lead to a consistent decrease or increase in the RMSEs and MAEs. In certain ranges (1.5-2.0, >2.0), both the RMSEs and MAEs rise upon the removal of the pI models. In contrast, in other ranges (0.5-1.0, buried), the RMSEs and MAEs decrease. We attribute this to the incomplete training of the pI models,

owing to the limited availability of pI data and the constraints of the training and test pK_a data. The biases within these datasets may hinder a robust analysis of the ablation study.

3.8 Performance on the test set with balanced distribution of pK_a shifts

pKALM is trained on a limited set of experimental data and tested with imbalanced pK_a values, which may introduce bias when evaluating largely shifted pK_a values. Here, we present pKALMs, a variant of pKALM, trained on a balanced training and test set using the same model architecture. Due to the scarcity of experimental data, constructing a balanced experimental data set of adequate size to train pKALMs is unfeasible. We adopt the data sets used in the series of works of DeepKa,[32,40,51] where they utilized calculated pK_a values from CpHMD simulations for training (named PHMD549), and experimental data for testing (named EXP67S). The distribution of pK_a shifts in the PHMD549 dataset is unbalanced, which contrasts with the test set and may effect the performance of the model on the balanced test set. To address this, we balanced the distribution of pK_a shifts in the training set by randomly sampling 1,000 pK_a values from each shift range of $[0, 0.5), [0.5, 1), [1, 1.5), [1.5, 2), [2, \infty)$, and constructed a new training set, PHMD549S. Detailed statistics of the PHMD549S dataset are shown in SI Figure S11. We retrained pKALM on the PHMD549S dataset and evaluated its performance on the EXP67S test set.

We compared the performance of pKALMs with different protein language models, as shown in SI Figure S12. In contrast to pKALM, pKALMs achieved the lowest RMSE on the larger PLM, ESM2_650M, while pKALM performed best on ESM2_35M. This is logical, as PHMD549S is considerably larger than the revPKAD2 dataset, necessitating more parameters to store the learned knowledge of pK_a prediction. It is important to note that model parameters increase with the input dimension of PLMs. Moreover, the test RMSE first decreases and then increases as the PLM size grows. This behavior reflects the scaling law of deep learning models, which aligns with prior studies.

The comparative performance of pKALM, DeepKa, and other methods is summarized in Table 6. It is important to note that, since our input structures may be processed through different pipelines than those used in the original work, their performance may differ slightly from the literature.[32] The results demonstrate that the performance of pKALMs is slightly better than that of DeepKa (0.970 vs 0.973), followed by PypKa, PropKa, PKAI, and PKAI+. Performance also varies across different residues, suggesting fundamental disparities in the methodologies of the predictors. For example, Poisson-Boltzmann-based method PypKa yields the best accuracy for Asp, while pKALMs excel at Glu and Lys, and DeepKa outperforms others on His. The poorer performance of pKALMs on His can be attributed to the absence of structural information, as pKALMs rely solely on sequence data.

Table 6: RMSEs of the pKALMs model, DeepKa, and CpHMD on the EXP67S test set. Here "Asp", "Glu", "His", and "Lys" are the RMSEs of the corresponding residues, while "Total" is the total RMSE. "MAE" is the mean absolute error, and "Max" is the maximum error. "R" and "m" are the correlation coefficient and slope of the linear regression, respectively.

Method	Asp	Glu	His	Lys	Total	MAE	Max	R	m
pKALMs	1.084	**0.962**	1.000	**0.717**	**0.970**	**0.673**	3.542	**0.946**	**0.965**
DeepKa	1.068	1.082	**0.768**	0.776	0.973	0.736	**2.799**	0.944	0.869
PypKa	**1.020**	1.278	0.864	0.821	1.066	0.790	4.278	0.940	0.926
PropKa	1.041	1.146	1.516	0.929	1.168	0.888	3.900	0.914	0.813
PKAI	1.260	1.490	0.976	0.806	1.236	0.850	5.770	0.918	0.924
PKAI+	1.324	1.401	1.095	0.856	1.244	0.916	5.850	0.908	0.849

3.9 High-throughput prediction

To evaluate the speeds of pKALM, the pK_a values of all proteins in the human proteome are predicted. The human proteome data set is downloaded from the UniProt website (`https://www.uniprot.org/proteomes/UP000005640`) which contains a total of 20,598 proteins and 3,554,768 protonatable residues. The whole prediction process took 11 minutes 56 seconds on our machine, which produced an average speed of 4,965 pK_a per second. The distribution histograms of the predicted pK_a values of different residues (SI Figure S13 for pKALM and SI Figure S14 for pKALMs) indicates the predictions have a wide range of

pK_a value distributions for all residues. This suggests that pKALM is capable of predicting the significantly shifted pK_a values in the proteins. The predicted pK_a of terminal groups have relatively concentrated distributions, which is consistent with the fact that N-terminal and C-terminal are more likely exposed to the solvent and thus have more concentrated pK_a to its standard pK_a values. Notably, the distribution produced by our method is more concentrated than those of existing methods, such as Chen *et al.*,[39] which may be attributed to the simplified input features and the absence of structural information in our approach.

We evaluate the speed of different predictors that compared in this study. We cite the reported speeds of DeepKa in the published work.[51] We have tested the speed on human proteome data set for our pKALM. Other methods are installed with their official programs on our machine. Furthermore, we randomly sample 1,000 structures from the PDB database and predict the pK_a values of all residues in the structures by PKAI, PKAI+, PropKa. Since PypKa is very slow, we sampled only 100 structures and predicted their pK_a values by PypKa. The speed of these predictors are shown in Table 7.

It is clear that pKALM is the fastest method among the compared methods, which is more than 37 times faster than PKAI/PKAI+, 130 times faster than PropKa, 694 times faster than DeepKa, and 5773 times faster than PypKa. The speed of pKALM is mainly due to its sequence-based design and the parallel computation of the deep learning models on the GPUs. PLMs are considered as the bottleneck of the deep learning models. However, the pK_a prediction of pKALM is performed for all protonatable residues in the same protein simultaneously. The price of PLM inference is paid only once for each protein, instead of each residue, which is more efficient than the residue-by-residue prediction. Another deep learning methods, PKAI/PKAI+ also provide the second and the third best speed. Surprisingly, the empirical method PropKa predicts more slowly than the deep learning methods, possibly due to its implementation on the CPU and the sequential prediction of the residues' pK_a values.

Table 7: Speed comparison of different methods for pK_a prediction.

Method	Total pK_a	Total time	pK_a per sec
pKALM, this work	**3,554,768**	**11 min 56 s**	**4,964.76**
PKAI	187,312	23 min 31 s	132.75
PKAI+	187,312	23 min 38 s	132.10
PropKa	219,198	96 min 14 s	37.96
DeepKa	1,537	3 min 35 s	7.15
PypKa	11,565	3 h 43 min 49 s	0.86

a Note that compared methods are tested on different data sets.

3.10 Case study and discussion

3.10.1 Prediction for ribonuclease H

We use pKALM to predict the pK_a values of ribonuclease H (PDB ID: 2RN2) to have a case study. Ribonuclease H is a small protein consisting of 155 residues (Figure 5a). It has been reported to have pK_a values as shown in Table 9.[78] The predicted pK_a values from pKALM are very close to the experimental values. This example demonstrates that pKALM accurately predicts the pK_a values of the residues in the protein.

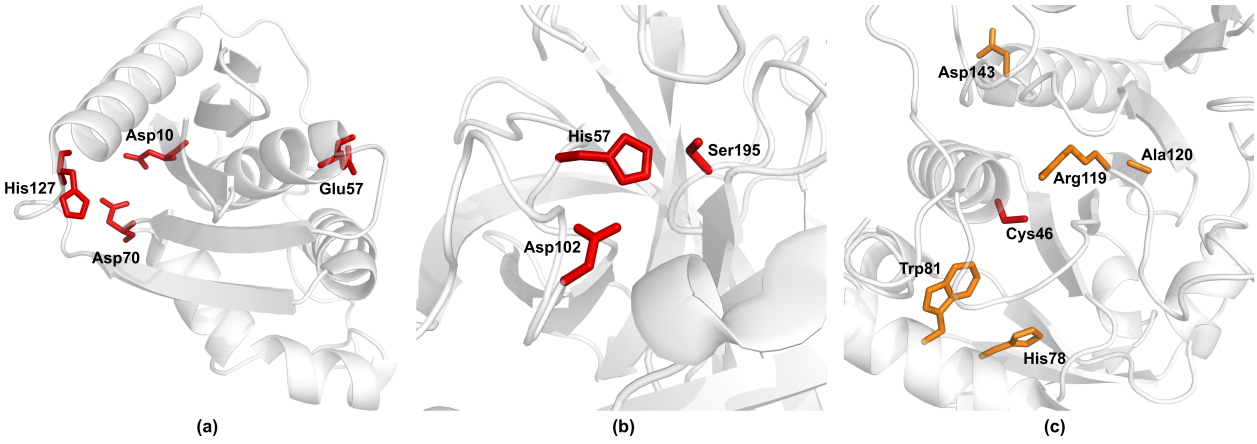

Figure 5: Exemplary predictions. (a) The five titrated residues of ribonuclease H. (b) The catalytic triad of chymotrypsin. (c) The bacterial peroxiredoxin AhpC. The catalytic Cys46 is shown in red. Residues with high attention maps to Cys46 are shown in orange.

Table 8: The experimental (Expr.), predicted (Pred.), and standard (Std.) pK_a values of selected residues in ribonuclease H, chymotrypsin, and AhpC. The solvent-accessible surface area (SASA) and relative solvent accessibility (RSA) of the residues are also shown.

Protein	Residue	Expr.	Pred.	Std.	SASA (Å^2)	RSA (%)
Ribonuclease H (2RN2)	Asp10	6.1	6.38	3.8	18.62	9.65
	Glu57	3.2	4.41	4.5	81.35	36.48
	Asp70	2.6	2.89	3.8	62.96	32.62
	His127	3.96	4.07	4.5	68.97	30.79
Chymotrypsin (4CHA)	His57	7.5	7.65	6.5	2.42	1.08
AhpC (4MA9)	Cys46	5.94	6.67	9.0	0.00	0.00

3.10.2 Prediction for chymotrypsin

pKALM is employed to predict the catalytic triad motif of chymotrypsin, a serine protease. This motif is comprised of Asp, His, and Ser (Figure 5b and Table 9). The input for pKALM is the B chain of the apoenzyme structure (PDB ID: 4CHA). Given the strong interaction between these three residues, the pK_a values of His57 are significantly altered from their standard pK_a value of 6.5. The predicted pK_a value of His57 from pKALM is 7.65, which is in close agreement with the experimental value of 7.5.[2] This indicates that pKALM is effective in capturing the intricate interactions between the residues and accurately predicting the pK_a values of the residues in the catalytic triad. It is worth noting that Ser in the catalytic triad is not on the input chains but the prediction is still accurate, possibly due to the strong interactions between the residues in the catalytic triad, which are captured by the protein language model.

3.10.3 Prediction on HEWL, SNase, CatD, BACE1, and BACE2

We evaluated the prediction of pKALMs for five enzymes: HEWL, SNase, CatD, BACE1, and BACE2, which are commonly employed to gauge the efficacy of pK_a predictors.[32,79] Residues with lower pK_a values function as proton donors, whereas those with higher pK_a values do as proton acceptors. These residues exhibit moderate pK_a shifts and have been predicted with high or acceptable accuracy (Table 9).

Table 9: Prediction of pK_a values for HEWL, SNase, CatD, BACE1, and BACE2 using pKALMs.

Enzyme	Residue	Expr.	Pred.	Std.	SASA (Å^2)	RSA (%)
HEWL (2LZT)	Glu35	6.1	5.88	4.5	6.43	2.88
	Asp52	3.6	3.49	3.8	0	0
SNase (3BDC)	Asp21	6.5	5.26	3.8	0	0
	Asp19	2.2	2.45	3.8	1.20	0.62
CatD (1LYA)	Asp33	5.8	6.31	3.8	4.42	2.27
	Asp231	4.1	3.76	3.8	3.21	1.65
BACE1 (1SGZ)	Asp32	5.2	4.92	3.8	1.07	0.55
	Asp228	3.5	4.78	3.8	24.39	12.5
BACE2 (3ZKQ)	Asp48	/	6.16	3.8	0	0
	Asp241	/	5.69	3.8	0	0

We also evaluated the pK_a predictions for three SNase mutants: Asp104 of 3P75, Glu92 of 1TQO, and Lys66 of 2SNM. However, the predicted values closely align with the standard pK_a values: 3.52 (Expr. 9.7), 3.78 (Expr. 9.0), and 11.15 (Expr. 5.6). These residues exhibit exceptionally large deviations from the standard pK_a values, exceeding 5 pH units. Such extreme pK_a shifts are exceedingly rare and may be inadequately represented in the training dataset. The PLM-based model captures only subtle signals of spatial interactions among residues, and shifts of this magnitude likely involve substantial structural alterations, which appear to exceed the model's current capabilities.

3.10.4 Why PLM-based design works?

We conduct a case study to demonstrate why the PLM-based design of pKALM works very well. We use pKALM to predict the pK_a values of Cys46 in the bacterial peroxiredoxin AhpC (PDB ID: 4MA9), a large protein consisting of five chains and hundreds of residues. The key residues for analysis are shown in Figure 5c and Table 9. The pK_a value of Cys46 is 5.94,[80] which is largely shifted from the standard pK_a value of 9.0. The predicted pK_a value from pKALM is 6.67, which is close to the experimental value, while other methods predict the values including PropKa (9.0), PKAI (9.64), and PKAI+ (9.08), close to the standard pK_a value of 9.0. This suggests that pKALM is more reliable for the prediction of the pK_a

values of Cys residues in proteins.

To visualize the effectiveness of pKALM, the attention maps are considered as the useful tool to interpret the deep learning models.[81] We thus visualize the attention maps of all 20 heads of the PLM ESM2_35M when the input is the sequence of AhpC (SI Figures S15 and S16). Most heads only capture the short-distance interactions between the residues, while some heads capture the long-distance interactions. For example, the 14th head in the attention map show the most significant long-distance interactions between the catalytic residue Cys46 and the other residues including His78, Trp81, Arg119, Ala120, and Asp143. These residues are distant from Cys46 in sequence and close in the 3D structure. The attention map imply that pKALM may successfully build the sequence-pK_{a} relationship by capturing the long-distance interactions and possibly more essential information of the residues in the protein sequences.

The PLM-based framework underlying our method exhibited exceptional performance in protein pK_{a} prediction in both precision and computational efficiency, though certain limitations remain. The current implementation of pKALM is constrained to utilizing experimental data exclusively as input, a limitation that is particularly pronounced for less prevalent residues such as Cys, Tyr, N-terminal, and C-terminal groups. As a result, the evaluation for these residues is less exhaustive compared to their more common counterparts. Furthermore, the relatively low occurrences of significantly shifted pK_{a} values hinder a comprehensive assessment of such cases. Traditional methodologies including CpHMD and Poisson-Boltzmann calculations can determine pK_{a} values independently of training data and may provide valuable synergies with deep learning-based approaches. These methods could mitigate the reliance on extensive experimental datasets and facilitate the expansion of experimental data for validation purposes.

Protein language models have proven their remarkable efficacy across diverse protein prediction and engineering tasks. However, their application to pK_{a} value prediction remains underexplored. This study introduces, for the first time, a framework that predicts pro-

tein pK_a values exclusively from the high-dimensional representations generated by protein language models. The findings reveal that these models capture the intricate relationship between sequence and pK_a, offering a powerful and innovative tool for pK_a prediction. We anticipate that integrating protein language models with structural information will significantly enhance the accuracy and robustness of pK_a predictions.

Additionally, this study introduces two accurate and fast pI models, which do not employ PLMs and are based on a sequence-centric design. These models underscore that PLMs are not universally effective for all protein prediction tasks, particularly when fine-tuning data, such as pI datasets, is insufficient. The primary contribution here lies in bridging the gap between pK_a and pI prediction problems through transfer learning, a conclusion further reinforced by ablation studies. However, the connection between pI and pK_a values appears complex, reflecting the intricate nature of protein environments. By leveraging deep learning methods, we aim to unveil these complexities and model the intricate relationships between these parameters. These pI models, characterized by their speed and precision, hold significant potential for large-scale protein pI value predictions.

pKALM was designed to operate on single-chain inputs and leverages simplicity and efficiency as key advantages while inevitably losing structural information. It cannot account for conformational changes not encoded in the sequence, such as ligand binding or post-translational modifications. Another concern arises from the quality of the experimental dataset employed in this study, PKAD-2, which suffers from inaccuracies in the annotation of pK_a values. Although we attempted to clean the dataset by verifying consistency, errors may still persist, potentially affecting the model's performance.

4 Conclusions

The understanding of protein functions in biological processes requires the knowledge of protein pK_a. It is a significant challenge to accurately measure the pK_a of specific residues

in a protein. However, computational methods provide an alternative approach to predict the pK_a values. In this study, we propose a deep learning-based method, pKALM, which employs a protein language model to achieve highly accurate predictions for Asp, Glu, His, Lys, Cys, Tyr, N-terminus, and C-terminus. Furthermore, two pI models were trained and integrated into the pKALM architecture. The pI models are of particular importance for the prediction of pK_a, with the protein pI model being especially crucial in this regard. The performance of pKALM was evaluated on the PKAD-2 dataset, and it was compared with existing methods. The results demonstrate that pKALM achieved state-of-the-art performance in terms of root mean square error (RMSE) and mean absolute error (MAE). Furthermore, the performance of pKALM was evaluated on data sets with significantly shifted pK_a values, exposed and buried residues, and the human proteome.

Supporting Information

The Supporting Information is available free of charge at https://xxx.

Details of building the training and test set in this work, histograms of pK_a distribution in revPKAD2 training and test set, histograms of distribution of pK_a shifts in the rEXP67S test set, histograms of protein and peptide pI distribution, heatmap of validated RMSE of peptide pI prediction with different hyperparameters, heatmap of validated RMSE of protein pI prediction with different hyperparameters, performance of different protein language models on different pK_a shift ranges, Pearson correlation coefficients between predicted and experimental pK_a values on each residue, Pearson correlation coefficients between predicted and experimental pK_a shifts on all residues, scatter plot of the predicted pK_a shifts to experimental pK_a shifts of pKALM and CpHMD method, comparison of Matthews correlation coefficient of different methods, histograms of pK_a distribution of PHMD549S, predicted pK_a distribution of human proteome from pKALM, predicted pK_a distribution of human proteome from pKALMs, visualized attention maps for the protein AhpC (head 1-10), vi-

sualized attention maps for the protein AhpC (head 11-20), parameters of different protein language models used in this study, standard pK_{a} values, hyperparameters used for tuning pKALM and pI models, comparison of test RMSE of different models for different residues, comparison of test MAE of different models for different residues, comparison of test RMSE of different models for different shift ranges, buried/exposed, and major/total residues, and comparison of test MAE of different models for different shift ranges, buried/exposed, and major/total residues

Data Availability Statement

All data sets used in this study are publicly available at PKAD-2 official website and the revised data set is available at `https://github.com/xu-shi-jie/revPKAD2`. The corresponding reproducible source code is publicly available at `https://github.com/xu-shi-jie/pKALM`. A web server for pKALM is available at `https://onodalab.ees.hokudai.ac.jp/pkalm`.

Author Contributions

All authors conceived and designed the study. S.X. developed the method, performed data analysis, and wrote the paper. A.O. supervised the research project and critically reviewed and revised the manuscript. All authors approved the final manuscript.

Conflicts of Interest

The authors declare no competing financial interest.

Acknowledgements

This work was supported by Hokkaido University DX Doctoral Fellowship (JST SPRING, Grant Number JPMJSP2119) to S.X. and SATREPS Project "Recovering High-Value Bioproducts for Sustainable Fisheries in Chile (ReBiS)" funded by JST/JICA (Grant Number JPMJSA2206) and JSPS KAKENHI Grant Number JP24H02213 in Transformative Research Areas (A) JP24A202 Integrated Science of Synthesis by Chemical Structure Reprogramming (SReP) and JP24K01533 to A.O.

References

(1) Knowles, J. R.; Jencks, W. P. The intrinsic pK_a-values of functional groups in enzymes: Improper deductions from the pH-dependence of steady-state parameter. *CRC Crit. Rev. Biochem.* **1976**, *4*, 165–173.

(2) Harris, T. K.; Turner, G. J. Structural basis of perturbed pK_a values of catalytic groups in enzyme active sites. *IUBMB Life* **2002**, *53*, 85–98.

(3) Kisgeropoulos, E. C.; Bharadwaj, V. S.; Mulder, D. W.; King, P. W. The contribution of proton-donor pK_a on reactivity profiles of [FeFe]-hydrogenases. *Front. Microbiol.* **2022**, *13*, 903951.

(4) Yang, A.-S.; Honig, B. On the pH dependence of protein stability. *J. Mol. Biol.* **1993**, *231*, 459–474.

(5) Tan, Y.-J.; Oliveberg, M.; Davis, B.; Fersht, A. R. Perturbed pK_a-values in the denatured states of proteins. *J. Mol. Biol.* **1995**, *254*, 980–992.

(6) Bradley, J.; O'Meara, F.; Farrell, D.; Nielsen, J. E. Highly perturbed pK_a values in the unfolded state of hen egg white lysozyme. *Biophys. J.* **2012**, *102*, 1636–1645.

(7) Onufriev, A. V.; Alexov, E. Protonation and pK changes in protein–ligand binding. *Q. Rev. Biophys.* **2013**, *46*, 181–209.

(8) Jones, R. L.; Wilson, W. D. Effect of ionic strength on the pK_{a} of ligands bound to DNA. *Biopolymers: Original Research on Biomolecules* **1981**, *20*, 141–154.

(9) Haslak, Z. P.; Zareb, S.; Dogan, I.; Aviyente, V.; Monard, G. Using Atomic Charges to Describe the pK_{a} of Carboxylic Acids. *J. Chem. Inf. Model.* **2021**, *61*, 2733–2743.

(10) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very fast prediction and rationalization of pK_{a} values for protein–ligand complexes. *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 765–783.

(11) 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.* **2011**, *7*, 525–537.

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

(13) Lee, M. S.; Salsbury Jr, F. R.; Brooks III, C. L. Constant-pH molecular dynamics using continuous titration coordinates. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 738–752.

(14) Wallace, J. A.; Shen, J. K. Predicting pK_{a} values with continuous constant pH molecular dynamics. *Methods Enzymol.* **2009**, *466*, 455–475.

(15) Williams, S. L.; de Oliveira, C. A. F.; McCammon, J. A. Coupling constant pH molecular dynamics with accelerated molecular dynamics. *J. Chem. Theory Comput.* **2010**, *6*, 560–568.

(16) Itoh, S. G.; Damjanović, A.; Brooks, B. R. pH replica-exchange method based on discrete protonation states. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3420–3436.

(17) Chen, W.; Morrow, B. H.; Shi, C.; Shen, J. K. Recent development and application of constant pH molecular dynamics. *Mol. Simul.* **2014**, *40*, 830–838.

(18) Lee, J.; Miller, B. T.; Damjanovic, A.; Brooks, B. R. Constant pH molecular dynamics in explicit solvent with enveloping distribution sampling and Hamiltonian exchange. *J. Chem. Theory Comput.* **2014**, *10*, 2738–2750.

(19) Swails, J. M.; York, D. M.; Roitberg, A. E. Constant pH replica exchange molecular dynamics in explicit solvent using discrete protonation states: implementation, testing, and validation. *J. Chem. Theory Comput.* **2014**, *10*, 1341–1352.

(20) Radak, B. K.; Chipot, C.; Suh, D.; Jo, S.; Jiang, W.; Phillips, J. C.; Schulten, K.; Roux, B. Constant-pH molecular dynamics simulations for large biomolecular systems. *J. Chem. Theory Comput.* **2017**, *13*, 5933–5944.

(21) Harris, R. C.; Liu, R.; Shen, J. Predicting reactive cysteines with implicit-solvent-based continuous constant pH molecular dynamics in amber. *J. Chem. Theory Comput.* **2020**, *16*, 3689–3698.

(22) Aho, N.; Buslaev, P.; Jansen, A.; Bauer, P.; Groenhof, G.; Hess, B. Scalable constant pH molecular dynamics in GROMACS. *J. Chem. Theory Comput.* **2022**, *18*, 6148–6160.

(23) Sequeira, J. G.; Rodrigues, F. E.; Silva, T. G.; Reis, P. B.; Machuqueiro, M. Extending the stochastic titration CpHMD to CHARMM36m. *J. Phys. Chem. B* **2022**, *126*, 7870–7882.

(24) Huang, Y.; Harris, R. C.; Shen, J. Generalized Born based continuous constant pH molecular dynamics in Amber: Implementation, benchmarking and analysis. *J. Chem. Inf. Model.* **2018**, *58*, 1372–1383.

(25) Buslaev, P.; Aho, N.; Jansen, A.; Bauer, P.; Hess, B.; Groenhof, G. Best practices in

constant pH MD simulations: accuracy and sampling. *J. Chem. Theory Comput.* **2022**, *18*, 6134–6147.

(26) Nielsen, J. E.; Vriend, G. Optimizing the hydrogen-bond network in Poisson–Boltzmann equation-based pK_a calculations. *Proteins: Struct., Funct., Bioinf.* **2001**, *43*, 403–412.

(27) Warwicker, J. pK_a predictions with a coupled finite difference Poisson–Boltzmann and Debye–Hückel method. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3374–3380.

(28) Reis, P. B.; Vila-Vicosa, D.; Rocchia, W.; Machuqueiro, M. PypKa: A Flexible Python Module for Poisson–Boltzmann-Based pK_a Calculations. *J. Chem. Inf. Model.* **2020**, *60*, 4442–4448.

(29) Chen, J.; Hu, J.; Xu, Y.; Krasny, R.; Geng, W. Computing protein pK_as using the tabi Poisson–Boltzmann solver. *J. Comput. Biophys. Chem.* **2021**, *20*, 175–187.

(30) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A., *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

(31) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D., *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876.

(32) Cai, Z.; Liu, T.; Lin, Q.; He, J.; Lei, X.; Luo, F.; Huang, Y. Basis for Accurate Protein pK_a Prediction with Machine Learning. *J. Chem. Inf. Model.* **2023**, *63*, 2936–2947.

(33) Habchi, J.; Tompa, P.; Longhi, S.; Uversky, V. N. Introducing protein intrinsic disorder. *Chem. Rev.* **2014**, *114*, 6561–6588.

(34) Gokcan, H.; Isayev, O. Prediction of Protein pK_a with Representation Learning. *Chem. Sci.* **2022**, *13*, 2462–2474.

(35) Garbuzynskiy, S. O.; Melnik, B. S.; Lobanov, M. Y.; Finkelstein, A. V.; Galzitskaya, O. V. Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray-and NMR-resolved protein structures? *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 139–147.

(36) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.

(37) Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE Trans. Neural Networks* **2008**, *20*, 61–80.

(38) Haykin, S. *Neural networks: a comprehensive foundation*; Prentice Hall PTR, 1998.

(39) Chen, A. Y.; Lee, J.; Damjanovic, A.; Brooks, B. R. Protein pK_a Prediction by Tree-Based Machine Learning. *J. Chem. Theory Comput.* **2022**, *18*, 2673–2686.

(40) Cai, Z.; Luo, F.; Wang, Y.; Li, E.; Huang, Y. Protein pK_a Prediction with Machine Learning. *ACS Omega* **2021**, *6*, 34823–34831.

(41) Reis, P. B.; Bertolini, M.; Montanari, F.; Rocchia, W.; Machuqueiro, M.; Clevert, D.-A. A Fast and Interpretable Deep Learning Approach for Accurate Electrostatics-Driven pK_a Predictions in Proteins. *J. Chem. Theory Comput.* **2022**, *18*, 5068–5078.

(42) Pahari, S.; Sun, L.; Alexov, E. PKAD: a database of experimentally measured pK_a values of ionizable groups in proteins. *Database* **2019**, *2019*, baz024.

(43) Ancona, N.; Bastola, A.; Alexov, E. PKAD-2: New entries and expansion of functionalities of the database of experimentally measured pK_a's of proteins. *J. Comput. Biophys. Chem.* **2023**, *22*, 515–524.

(44) Reis, P. B.; Clevert, D.-A.; Machuqueiro, M. pKPDB: a protein data bank extension database of pK_a and pI theoretical values. *Bioinformatics* **2022**, *38*, 297–298.

(45) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J., *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* **2021**, *118*, e2016239118.

(46) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y., *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123–1130.

(47) Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **2019**, *16*, 1315–1322.

(48) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M., *et al.* Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7112–7127.

(49) Xu, S.; Onoda, A. Accurate and fast prediction of intrinsically disordered protein by multiple protein language models and ensemble learning. *J. Chem. Inf. Model.* **2023**, *64*, 2901–2911.

(50) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152.

(51) Cai, Z.; Peng, H.; Sun, S.; He, J.; Luo, F.; Huang, Y. DeepKa Web Server: High-Throughput Protein pK_{a} Prediction. *J. Chem. Inf. Model.* **2024**, *64*, 2933–2940.

(52) Steinegger, M.; Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **2017**, *35*, 1026–1028.

(53) Kozlowski, L. P. IPC 2.0: prediction of isoelectric point and pK_a dissociation constants. *Nucleic Acids Res.* **2021**, *49*, W285–W292.

(54) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**,

(55) Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.

(56) Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* **2020**,

(57) Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.

(58) Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chao, H.; Chen, L.; Craig, P. A.; Crichlow, G. V.; Dalenberg, K.; Duarte, J. M., *et al.* RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res.* **2023**, *51*, D488–D508.

(59) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: making protein folding accessible to all. *Nat. Methods* **2022**, *19*, 679–682.

(60) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D., *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **2017**, *13*, e1005659.

(61) Biewald, L. Experiment Tracking with Weights and Biases. 2020; `https://www.wandb.com/`, Software available from wandb.com.

(62) Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* **2017**,

(63) Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M., *et al.* Transformers: State-of-the-art natural language processing. Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. 2020; pp 38–45.

(64) Gugger, S.; Debut, L.; Wolf, T.; Schmid, P.; Mueller, Z.; Mangrulkar, S.; Sun, M.; Bossan, B. Accelerate: Training and inference at scale made simple, efficient and adaptable. `https://github.com/huggingface/accelerate`, 2022.

(65) Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; Garcia, D.; Ginsburg, B.; Houston, M.; Kuchaiev, O.; Venkatesh, G., *et al.* Mixed precision training. *arXiv preprint arXiv:1710.03740* **2017**,

(66) Po, H. N.; Senozan, N. The Henderson-Hasselbalch equation: its history and limitations. *J. Chem. Educ.* **2001**, *78*, 1499.

(67) Bjellqvist, B.; Basse, B.; Olsen, E.; Celis, J. E. Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis* **1994**, *15*, 529–539.

(68) Nozaki, Y.; Tanford, C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions: establishment of a hydrophobicity scale. *J. Biol. Chem.* **1971**, *246*, 2211–2217.

(69) Tabb, D. L.; McDonald, W. H.; Yates, J. R. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **2002**, *1*, 21–26.

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

(71) Sillero, A.; Ribeiro, J. M. Isoelectric points of proteins: theoretical determination. *Anal. Biochem.* **1989**, *179*, 319–325.

(72) Dawson, R. M. C.; Elliott, D. C.; Elliott, W. H.; Jones, K. M. Data for biochemical research. **1959**,

(73) 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.* **2009**, *18*, 247–251.

(74) Kozlowski, L. P. IPC–isoelectric point calculator. *Biol. Direct* **2016**, *11*, 1–16.

(75) Halligan, B. D. *ProMoST: A Tool for Calculating the pI and Molecular Mass of Phosphorylated and Modified Proteins on Two-Dimensional Gels*; Springer, 2009; pp 283–298.

(76) Toseland, C. P.; McSparron, H.; Davies, M. N.; Flower, D. R. PPD v1. 0—an integrated, web-accessible database of experimentally determined protein pK_{a} values. *Nucleic Acids Res.* **2006**, *34*, D199–D203.

(77) Rodwell, J. D. Heterogeneity of component bands in isoelectric focusing patterns. *Anal. Biochem.* **1982**, *119*, 440–449.

(78) Oda, Y.; Yamazaki, T.; Nagayama, K.; Kanaya, S.; Kuroda, Y.; Nakamura, H. Individual ionization constants of all the carboxyl groups in ribonuclease HI from Escherichia coli determined by NMR. *Biochemistry* **1994**, *33*, 5275–5284.

(79) Huang, Y.; Yue, Z.; Tsai, C.-C.; Henderson, J. A.; Shen, J. Predicting catalytic proton donors and nucleophiles in enzymes: How adding dynamics helps elucidate the

structure–function relationships. *The journal of physical chemistry letters* **2018**, *9*, 1179–1184.

(80) Nelson, K. J.; Parsonage, D.; Hall, A.; Karplus, P. A.; Poole, L. B. Cysteine pK_{a} values for the bacterial peroxiredoxin AhpC. *Biochemistry* **2008**, *47*, 12860–12868.

(81) Vig, J.; Madani, A.; Varshney, L. R.; Xiong, C.; Socher, R.; Rajani, N. F. Bertology meets biology: Interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222* **2020**,