Protein Electrostatic Properties are Finetuned Through Evolution

Mingzhe Shen,[†] Guy W. Dayhoff II,[†,‡] Daniel Kortzak,[†] and Jana Shen[*,†]

† *Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy, Baltimore, MD 21201, U.S.A.*
‡ *Joint first author*

E-mail: jana.shen@rx.umaryland.edu

Abstract

Ionization states play crucial roles in protein functions. Yet despite decades of research, predicting protein pK_a values remains a formidable challenge. Here, we present KaML-ESMs, neural networks built on the ESM protein language models and trained on a synthetically augmented experimental dataset. The KaML-ESMs significantly outperform structure-based approaches for protein pK_a prediction, achieving accuracies approaching experimental resolution across six ionizable amino acids in native proteins. On the most challenging test set of engineered OBTRUDEs (iOnizable suBsTitutions foR bUrieD rEsidues), KaML-ESM2 emerges as the top model, achieving RMSE of 1.36. We applied KaML-ESM2 to the entire human proteome and demonstrated that the predicted pK_a's can identify functional sites and infer catalytic mechanisms. Our work challenges the structure-based paradigm and suggests a view that protein electrostatic properties are encoded in sequence, potentially co-optimized with structure and function through evolution. We offer KaML, a sequence-based end-to-end platform to support applications ranging from biological exploration to drug design, protein engineering, and biomolecular simulations. GAINES, a latent space sampling approach introduced in this work, provides a general framework for addressing data scarcity bottlenecks in machine learning applications.

Introduction

Ionization states play crucial roles in protein structure and function. In the past, prediction and interpretation of ionization states relied on structure-based approaches, including physics-based[1–3] and empirical calculations.[4] Recently, structure-based ML models have been developed and trained on the Poisson-Boltzmann (PB) calculations,[5] constant pH molecular dynamics (CpHMD) simulations[6,7] and experimental data.[8–10] Given that protein functions (e.g., catalysis) are often governed by ionization states, here we hypothesized that the latter can be directly predicted from the protein sequence.

Through masked learning of protein sequences evolved over hundreds of millions of years, evolutionary scale models (ESMs), a class of protein large language models (pLLMs), demonstrated powerful capacities in predicting structures and functions.[11–13] This is because representations emerging within the pLLM reflect the biological structure and function and improve with scale, e.g., ESM3 is trained with 2.78 billion protein sequences.[12] We posited that residue-level representations learned by ESMs encode information about residue-specific ionization states and pK_a shifts that occur when residues transition from solution to protein environment, which are often large for functional sites.

To test the sequence-electrostatics hypothesis, we developed sequence-based ML models, where per-token (i.e., residue-specific) embeddings from the ESM2[11] or ESMC[13] model were used as inputs to a feedforward neural network for predicting residue-specific pK_a values. Training was conducted on a revised experimental pK_a dataset (PKAD-3[8]) augmented by synthetic data harvested using a novel approach called GAINES (auGment dAta wIth lateNt spacE Sampling). The new models are named KaML-ESMs, as the training protocol follows our recent development of the structure-based KaML (p$\boldsymbol{K_a}$ **M**achine **L**earning) models, including KaML-CBT which achieved top performance on random holdouts of PKAD-3 dataset for predicting pK_a values of all six types of titratable amino acids (DECYHK).

Across four benchmark test sets, including the most challenging pK_a dataset, composed of the engineered OBTRUDEs (iOnizable suBsTitutions foR bUrieD rEsidues) of staphylococcal nuclease (SNase) variants,[14,15] KaML-ESM2 outperforms alternative ap-

proaches. Our data suggest that protein sequence encodes not only structure and function but also electrostatic properties, which may have been co-optimized through evolution. We developed an end-to-end KaML platform and applied KaML-ESM2 to predict the pK_a's of the entire human proteome, illustrating the utility for functional annotation and mechanistic interpretation. We expect KaML-ESMs to enable a wide range of applications, from understanding biology to supporting drug discovery, protein engineering and MD simulations.

Results and Discussion

The experimental protein pK_a datasets. We recently curated an experimental pK_a database PKAD-3,[8] which significantly expanded upon PKAD-2 developed by the Alexov group.[16,17] Part of the expansion includes the pK_a values of SNase OBTRUDEs,[14,15] which are extremely difficult to predict due to abnormal pK_a shifts in the hydrophobic environment. The SNase OBTRUDEs served as a blind pK_a prediction challenge and inspired the development of physics-based and empirical pK_a calculation methods.[1] While most OBTRUDEs show very large (over 2 pH units) pK_a shifts relative to the model peptide values (called model pK_a's), a few lysines show zero pK_a shifts, suggesting partial unfolding which allows bulk solvent access as evidenced by simulations and NMR experiment.[18,19] In this work, we expanded the OBTRUDE dataset to 89 pK_a's of SNase variants, arriving at a revised dataset PKAD-3r (see Table S1 for details).

ESM representations distinguish between titratable amino acids. Before deploying ESM as a foundation for downstream learning and predicting residue-specific protein pK_a's, we used the t-distributed stochastic neighbor embedding (t-SNE) analysis to address two questions. 1) Can ESM per token embeddings (residue representations) differentiate between different amino acids? 2) Can ESM representations capture the degrees of ionization of amino acid residues to some degree? We extracted the representations of the titratable residues in PKAD-3r from ESM2 (trained with ∼65 million unique sequences and 650 M parameters)[11] and the latest ESM Cambrian or ESMC (trained with 2.78 billion sequences and 6 B parameters),[12,13] which has been demonstrated to predict emergent structures with higher precision than ESM2. The t-SNE algorithm[20,21] was then applied to visualize pairwise residue representation similarities on two dimensions.

With six titratable amino acids colored distinctly, the t-SNE maps (Fig. 1a) demonstrate that both ESM2 and ESMC can distinguish between different amino acids as evident from the formed clusters. In particular, basic (HK) and acidic (DECY) residues are clearly separated, whereas Asp and Glu residues are somewhat mixed due to their similar sequence and (implicitly) structure environments. Notable exceptions are model peptide residues (AAXAA, X=titratable) and OBTRUDEs, which overlap with each other regardless of the residue type. This demonstrates that the ESM representations encode sequence and structural contexts rather than amino acid identities. It is also note worthy that most OBTRUDEs cluster together despite different sequence positions, suggesting that their shared hydrophobic, buried environment is distinct from those of native ionizable residues (see later discussion).

ESM representations capture weak protein pK_a signals. Next, we examined whether the ESM representation similarity is correlated with similarity in pK_a by coloring data on the t-SNE map according to the experimental pK_a shifts. While most data points with positive and negative pK_a shifts are mixed, a few clusters with pK_a shifts greater than 1 are visible for both acids and bases (Fig. 1b and Fig. S1a for ESMC analysis). Comparison of the maps with and without OBTRUDEs reveals that the two largest acid clusters of upshifted pK_a's and the largest base cluster of downshifted pK_a's belong to OBTRUDEs (Fig. 1b and Fig. S2), which reinforces the notion that OBTRUDEs and native titratable residues have dissimilar environments. Collectively, the t-SNE analysis demonstrates that while the ESM representations capture some weak pK_a signals, they cannot be directly used for pK_a predictions.

Developing KaML-ESMs for protein pK_a predictions. Having assessed the suitability of ESM2 or ESMC as foundation model, we proceeded to developing downstream ML models for protein pK_a predictions. We built a 4-layer multilayer perceptron (MLP) task head, whereby ESM representations of the input sequence serves as input and model training is conducted on experimental pK_a shifts. We compared the performance of the MLP and a finetuned ESM2 model, both trained on pK_a shifts from PKAD-3. In one holdout test, the MLP yielded an overall RMSE of 0.76, significantly outperforming the finetuned ESM2 (RMSE of 1.16). Therefore, we focused on developing MLP-based task heads for ESM models, termed KaML-ESMs, extending our KaML framework for protein pK_a predictions.

Following the development of structure-based KaML models,[8] separate MLPs were trained for acid and base residues. To avoid potential overfitting, acid and base models were pretrained on a synthetic dataset comprised of pK_a shifts of 29,457 residues in 9,945 proteins from the PDB predicted by our recent structure-based KaML-CBT model[8] (Supplemental Methods and Fig. S3). The pretrained models were then finetuned on the experimental PKAD-3 dataset

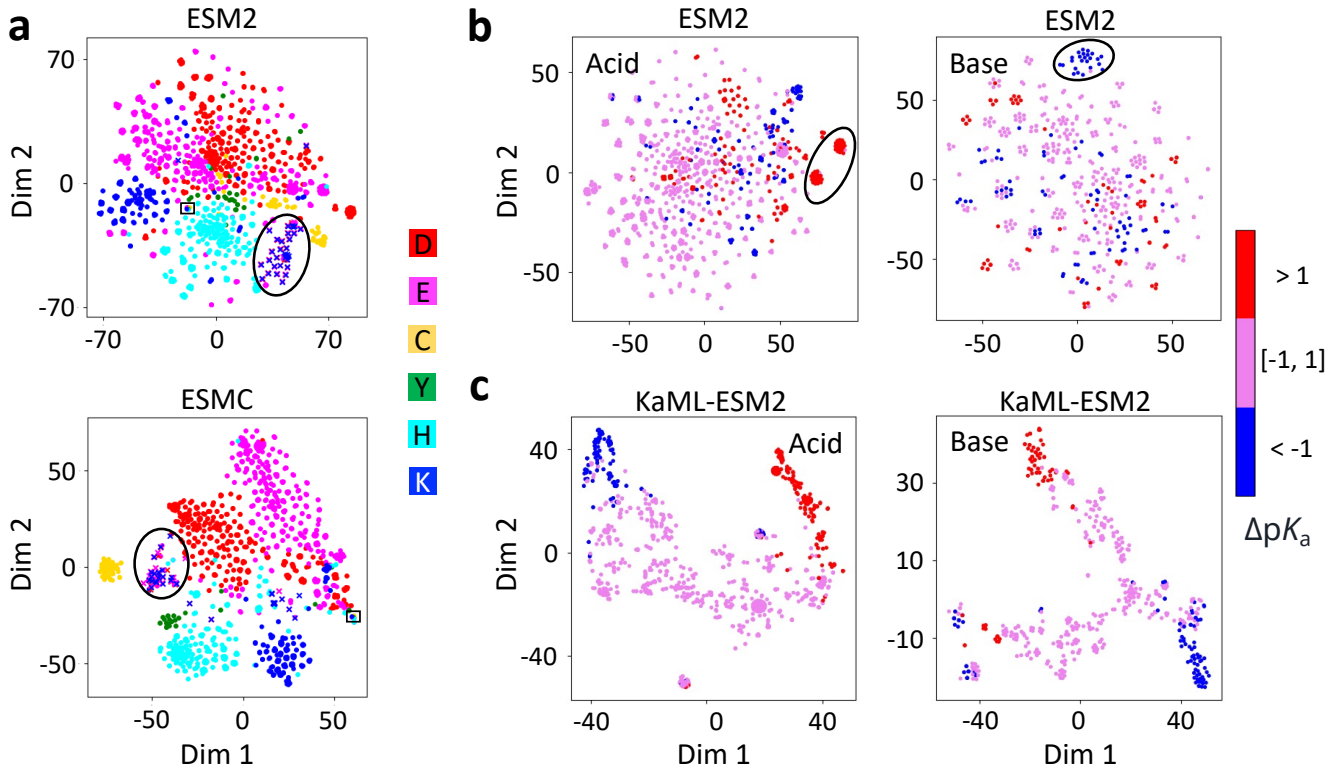

Figure 1: **ESM representations can distinguish between titratable amino acids and capture weak signals of residue-specific pK_a shifts. a.** t-SNE visualization of the pairwise ESM2 (top) and ESMC (bottom) representation similarities between titratable residues in PKAD-3r dataset. Data points are colored according to amino acid. Circle indicates the OBTRUDEs (in crosses); box indicates model peptide residues. **b, c.** t-SNE visualization of the pairwise representation (ESM2 in b, KaML-ESM2 in c) similarities between acids (DECY) and between bases (HK). Data points are colored according to the experimental pK_a shifts relative to the model values. Embeddings were extracted from layer 31 of ESM2_650M (1280-dimensional), layer 80 of ESMC (2560-dimensional), and the last hidden layer of KaML-ESM2 (32-dimensional) after training on PKAD-3r. See related Fig. S1 and S2.

(Fig. S4). Both pretraining and separating acid and base models decreased the pK_a errors (Appendix and Table S2).

Since protein pK_a shifts are primarily determined by local structure environment[8,22] and previous work found that local contacts are learned across different transformer layers in the pLLM,[23] we analyzed different ESM layers by training models using each layer's embeddings. The second-to-last layer of ESM2 and the last layer of ESMC gave the lowest overall test RMSEs, these layers were selected for further development. We also examined the potential impact of increasing ESM's parameter scale by comparing models trained with ESM2 used so far (ESM2_650M) with a larger ESM2 (ESM2_15B). Surprisingly, KaML-ESM2_15B gave a larger overall RMSE and was therefore excluded from further discussion. Details are given in Appendix, Table S3-S5, and Fig. S5.

KaML-ESMs outperform the structure-based KaML-CBT but RMSEs for Cys and Tyr remain above 1 unit. We examined the amino acid based RMS pK_a errors from KaML-ESM2 and KaML-ESMC predictions across 20 holdout sets generated from random splits of PKAD-3 (Table S6). Note, the RMSE converges after 20 random splits and holdout tests (Fig. S6). Overall, the RMSEs for Asp, Glu, His, and Lys residues are 0.50–0.69 units, while the RMSEs for Cys and Tyr residues are much larger, 1.1–1.54 units. This discrepancy is expected, because the training data size for Cys (60) and Tyr (39) residues is extremely small. It is also noteworthy that ESMC gives somewhat smaller pK_a and CER for bases, while the performance for acids is similar. Encouragingly, the errors for all residue types are decreased compared to our recent structure-based KaML-CBT models, which were trained using the same protocols and dataset.

Developing GAINES to overcome data scarcity. The extremely limited pK_a data for cysteines and tyrosines exemplify the data scarcity challenge common in scientific ML: limited experimental data for rare events. To tackle this challenge, we propose a data augmentation strategy called GAINES (auGment dAta wIth lateNt spacE Sampling). Inspired by the attention mechanism in transformers,[24] GAINES uses training data as queries to retrieve data from an unrelated database based on embedding similarity. The retrieved data are then assigned the labels of their corresponding training queries. We implemented a GAINES protocol for augmenting experiment pK_a dataset (Fig. 2a). For each query residue (e.g., a cysteine with an experimental pK_a), its ESM2 or ESMC representation serves as the query embedding, which is compared against the embeddings of all residues (keys) in a protein database (proteins overlapping the training set excluded). If the cosine similarity between query and key embeddings exceeds a predefined threshold (e.g., 0.8 used in this work), the value residue is assigned the query residue's pK_a and added to the synthetic data pool. For convenience in testing structure-based KaML-CBT models, we used RCSB PDB as the protein database in this work. The GAINES protocol yielded a large pool of synthetic pK_a data; however, to preserve the pK_a distributions similar to the experimental PKAD-3 dataset, we randomly down-sampled to a maximum of 10 value residues (and proteins) per query residue. The resulting augmented dataset comprises pK_a data of 4,229 unique residues (1,064 Asp, 1,209 Glu, 760 His, 675 Cys, 417 Lys, and 104 Tyr) from 2,165 proteins, approximately ten times the experimental dataset (Fig. S7 and Table S1).

A potential concern is that residue embedding similarity may simply reflect sequence similarity rather than meaningful structural or physicochemical relationships, which could hinder model generalization. This is however not the case, as the majority of the query and value protein pairs share a sequence identity below 40% (Fig. S8). Taking Cys12 in E. Coli glutaredoxin 3 (experimental pK_a of 4.1[25]) as an example, the global and local sequence (within 10 amino acids from the cysteine of interest) identities between P0AC62 and most (9 out of 10) value proteins (from different species) fall below 40% (Fig. 2b and 2c). The ESM3-predicted structures demonstrate that while global structures may differ, the local environment around the cysteine of interest is highly similar (Fig. 2d). Protein family analysis with Pfam[26] reveals that all these proteins are related as glutathione S-transferase-dependent enzymes, suggesting that embedding similarity captures functional conservation across evolutionarily distant species.

Significant error reduction for both sequence- and structure-based models using GAINES. Using the GAINES augmented training set, we retrained KaML-ESMs and tested on 50 holdout sets (only experimental data) (Supplemental Methods). Strikingly, the RMSEs for all residue types are decreased for both KaML-ESM2 and KaML-ESMC models (filled/striped color vs. filled/striped gray bars in Fig. 3 and Table S7). As expected, the most significant performance gain is for Cys and Tyr pK_a's, with the RMSE of KaML-ESM2 decreased to 0.50 and 0.33, respectively. Note, the latter needs to be taken with a grain of salt due to the extremely small test set. Similarly, the RMSEs of KaML-ESMC for Cys and Tyr pK_a's are decreased to 0.57 and 0.89, respectively. Importantly, the critical error rate, defined as the fraction of misclassified protonation states at pH 7, is reduced by 8 or 5 times to 4 or 5 out of 178 for Cys for KaML-ESM2 or KaML-ESMC, respectively (Table S7). We also tested the models with and without GAINES training on EXP67S,[6] a challenging holdout set comprised of 167 pK_a's of Asp, Glu, His, and Lys residues in 52 proteins. The overall RMSEs of KaML-

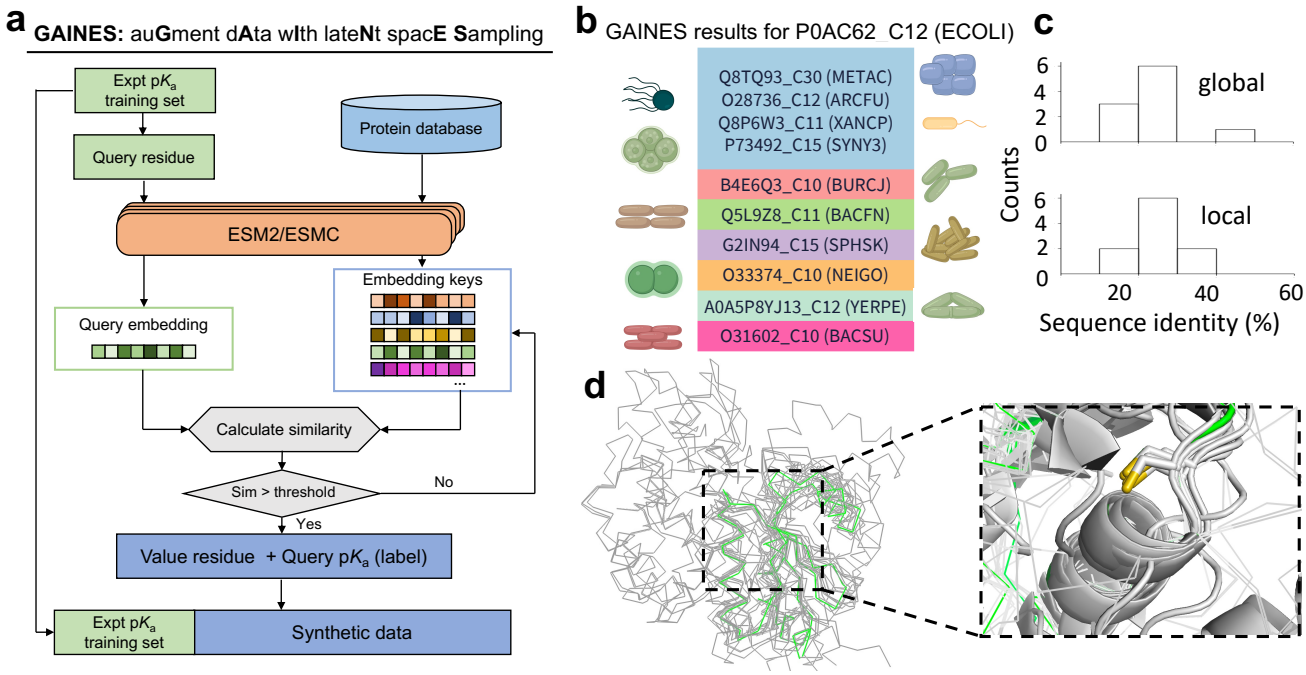

Figure 2: **GAINES protocol and its application to identify synthetic data related to a downshifted cysteine pK_a in glutaredoxin 3. a.** The GAINES flowchart. The query, key, and value concept is used to identify synthetic data from an external database. **b.** GAINES identified 10 value proteins across 10 species (in parenthesis) that contain a cysteine similar to query C12 in E. Coli glutaredoxin 3 (UniProt ID: P0AC62). **c.** A histogram of the global and local sequence identities between query (P0AC62) and value proteins. **d.** Overlay of the ESM3-predicted structures of P0AC62 (green) and value proteins (gray). A zoomed-in view shows a similar local environment (within 5 Å) for the query and value cysteines (yellow and green sticks).

ESM2 and KaML-ESMC are reduced by 0.35 and 0.23, respectively (Supplemental Data Sheet 5).

To test if GAINES training also boosts performance of structure-based models, we retrained KaML-CBT using a GAINES augmented training set, in which protein structures corresponding to value residues and sequences were added.[8] As with KaML-ESM models, GAINES introduces a significant performance gain for KaML-CBT predictions of Asp, His, Cys and Tyr pK_a's, where the average RMSEs are reduced to 0.66, 0.63, 1.01, and 0.52, respectively (empty color vs. empty gray bars in Fig. 3 and Table S7).

Random holdout and external tests establish KaML-ESMs as a new SOTA for protein pK_a predictions.

We evaluated the performance of KaML-ESM models using 50 random holdout sets, 1 external evaluation set, and 2 challenging datasets (EXP67S and OBTRUDE), comparing them against recent high-performing ML models as well as established empirical and physics-based methods (Table 1, Table S7, Supplemental Data Sheets 3–6). Across all test sets, KaML-ESMs significantly outperform alternative models.

Tested on 50 random holdout sets, KaML-ESM2 achieves an overall RMSE of 0.46±0.02 and amino acid-specific RMSEs of 0.32–0.51 with standard errors of 0.02–0.05. KaML-ESMC gives a similar performance, achieving an overall RMSE of 0.51±0.02 and amino acid-specific RMSEs of 0.37–0.57 with standard errors of 0.03–0.08 (except for 0.14 for Tyr). Given that the precision of NMR pK_a estimates is roughly 0.5 units based on a survey of NMR titration studies of hen egg lysozyme,[27] the average errors of KaML-ESM models approach experimental uncertainty.

In the random holdout tests, the best performing non-KaML model is DeepKa, a structure-based CNN model[6,7] trained on the implicit-solvent continuous constant pH molecular dynamics (GBNeck2-CpHMD)[28] data. DeepKa gives an overall RMSE of 0.96 for DEHK residues (DeepKa cannot predict Cys and Tyr pK_a's), compared to 0.46 and 0.49 by KaML-ESM2 and KaML-ESMC, respectively (Table 1 and Supplemental Data Sheet 3). In comparison, the graph neural network (GNN) based model aLCnet gives slightly larger RMSE of 1.03 for DEHK residues. The errors of Poisson-Boltzmann solver PypKa[2] are even larger, with the DEHK RMSE of about 1.86. Importantly, KaML-ESMs have substantially reduced fraction of misclassified protonation states, by more than 5, 7, and 13 times compared to DeepKa, aLCnet, and PypKa, respectively (Table 1). This performance leap establishes KaML-ESMs as a practical tool for applications requiring accurate protonation states, including MD simulations and

Table 1: Comparison of sequence-based KaML-ESM2 against structure-based ML and physics-based methods for pK_a predictions[a]

	KaML-ESM2		KaML-CBT2[b]		DeepKa[7]		aLCnet[10]		PypKa[2]	
	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC
Asp	$\mathbf{0.51 \pm 0.02}$	0.95 ± 0.01	0.66 ± 0.03	0.89 ± 0.01	1.14 ± 0.03	0.64 ± 0.02	0.92 ± 0.02	0.81 ± 0.01	1.95 ± 0.07	0.19 ± 0.14
Glu	$\mathbf{0.41 \pm 0.02}$	0.94 ± 0.01	0.58 ± 0.01	0.86 ± 0.01	0.75 ± 0.01	0.75 ± 0.01	0.80 ± 0.02	0.69 ± 0.02	1.54 ± 0.06	0.43 ± 0.03
Cys	$\mathbf{0.50 \pm 0.05}$	0.95 ± 0.02	1.01 ± 0.09	0.73 ± 0.08	n/d[c]	n/d[c]	1.71 ± 0.09	0.33 ± 0.09	3.52 ± 0.16	0.22 ± 0.09
Tyr	$\mathbf{0.32 \pm 0.03}$	0.34 ± 0.11	0.52 ± 0.05	0.12 ± 0.12	n/d[c]	n/d[c]	1.80 ± 0.12	-0.12 ± 0.13	1.26 ± 0.15	0.16 ± 0.14
His	$\mathbf{0.39 \pm 0.02}$	0.92 ± 0.01	0.63 ± 0.01	0.75 ± 0.02	0.79 ± 0.03	0.61 ± 0.03	1.11 ± 0.02	0.63 ± 0.02	1.68 ± 0.09	0.18 ± 0.05
Lys	$\mathbf{0.50 \pm 0.03}$	0.89 ± 0.01	0.60 ± 0.03	0.87 ± 0.02	0.85 ± 0.05	0.77 ± 0.03	0.76 ± 0.04	0.74 ± 0.03	2.03 ± 0.19	0.30 ± 0.05
All	$\mathbf{0.46 \pm 0.02}$	0.98 ± 0.01	0.64 ± 0.01	0.96 ± 0.01	n/d[c]	n/d[c]	1.09 ± 0.01	0.91 ± 0.01	1.92 ± 0.05	0.72 ± 0.01
DEHK[c]	$\mathbf{0.46 \pm 0.01}$	0.98 ± 0.01	0.63 ± 0.01	0.96 ± 0.01	0.92 ± 0.02	0.92 ± 0.01	1.03 ± 0.01	0.89 ± 0.01	1.86 ± 0.05	0.72 ± 0.01
MAXE[d]	2.29 ± 0.09		2.73 ± 0.07		4.49 ± 0.10		4.37 ± 0.08		7.88 ± 0.13	
CER[e]	$51/6741$		$74/6387$		$277/6774$		$370/6847$		$689/6739$	
External	**0.43**	0.98	0.88	0.96	0.64[c]	0.97[c]	0.70	0.96	1.3	0.77
EXP67S[f]	0.52	0.92	0.85	0.80	0.97	0.74	1.08	0.62	2.01	0.30
OBTRUDE[g]	1.36	0.56	2.36	0.32	1.82	0.56	2.11	0.56	2.21	0.36

[a]RMSEs (mean with standard error) and Pearson correlation coefficients (PCCs) for 50 random (PKAD-3r) and two challenging holdout sets are listed. Random holdout sets likely have partial overlap with the training set of aLCnet.[10] [b]KaML-CBT model[8] retrained with GAINES. [c]DeepKa cannot predict Cys and Tyr pK_a's. [d]Metrics for only DEHK residues. [e]Maximal absolute error. [f]Critical error rate (CER) refers to the fraction of misclassified protonation states at pH 7.[8] [g]EXP67S refers to a set of highly shifted DEH pK_a's compiled in Ref[6] for testing models. KaML models were retrained by removing overlapping data. DeepKa and aLCnet metrics are taken from Ref.[10] PCCs for EXP67S are calculated from pK_a shifts. [h]The most challenging test set for pK_a prediction methods[1] (see main text). DeepKa predictions were made with the web server,[7] while aLCnet[10] and PypKa[2] predictions used command line programs. Metrics of KaML-ESMC and additional ML models are given in Supplemental Data Sheets 3–6.

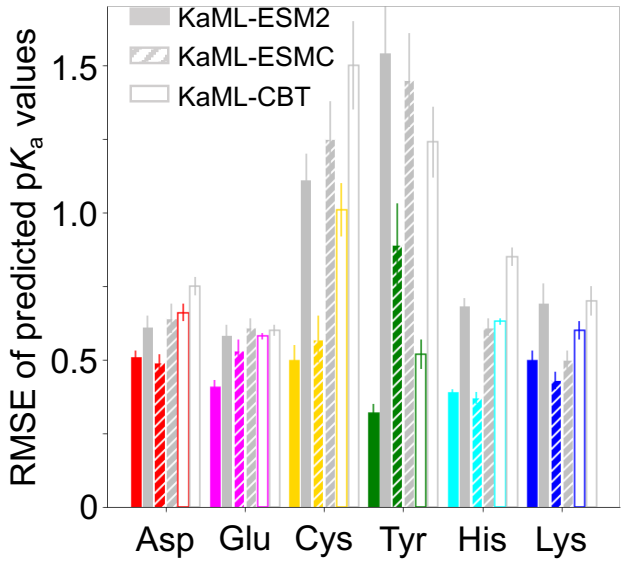

Figure 3: **GAINES improves sequence- and structure-based KaML models.** Amino acid-specific RMSEs of pK_a predictions by KaML-ESM2 (solid color bars), KaML-ESMC (striped color bars), and KaML-CBT (open color bars). The RMSEs of models trained without GAINES are shown in gray bars. The mean and standard errors from 50 holdout tests are shown. Also see Table S6.

mechanistic studies of pH-dependent biological processes.

We further evaluated KaML-ESMs and alternative models on non-redundant experimental pK_a data published after April 2023 (time stamp for PKAD-3[8]). Testing on this external test set, which comprises 54 pK_a values (39 His, 1 Cys, 1 Tyr, and 13 Lys) from 16 proteins , KaML-ESM2 and KaML-ESMC achieve overall RMSEs of 0.43 and 0.39, respectively, which are lower than all other models (Supplemental Data Sheet 4).

KaML-ESM performance extends to the EXP67S benchmark. We asked if the performance of KaML-ESMs extends to the more balanced EXP67S test set,[6] which includes large pK_a shifts and has been used as a benchmark by others.[6,7,10,29] KaML-ESM2 and KaML-ESMC retrained by excluding EXP67S achieve overall RMSEs of 0.52 and 0.56, respectively, comparable to their performance on the random holdout sets, which confirms the predictive power (Table 1). In comparison, the RMSEs of DeepKa, aLCnet, and PypKa are 0.97, 1.08, and 2.01, respectively (Table 1 and Supplemental Data Sheet 5).

The OBTRUDE test demonstrates KaML-ESM2's capacity for extrapolation. The pK_a's of 89 engineered, deeply buried OBTRUDEs in SNase variants are extremely difficult to predict using physics-based and empirical pK_a methods,[1] due to ionization-induced conformational rearrangement or partial unfolding not captured by the static protein

structure.[18,19,30] The OBTRUDEs are also expected to challenge pLLM-based models due to their lack of coevolutionary information. Indeed, our GAINES protocol did not identify any related residue in the PDB, suggesting that the engineered mutations are unprecedented in nature.

The KaML-ESM2_no_OB and KaML-ESMC_no_OB models (trained without OBTRUDEs), achieve RMSEs of 1.36 and 1.76, respectively, outperforming DeepKa (1.82), aLCnet (2.11), and PypKa (2.21), which rely on structures but not evolutionary information. Importantly, KaML-ESM2 gives the lowest RMSE for all three residues types (Asp, Glu, and Lys), establishing it as the top performer. It is worth noting that the empirical method PROPKA3,[4] which was "adjusted" for the buried OBTRUDEs, achieves RMSE of 1.73 (Supplemental Data Sheet 6), while the temperature replica-exchange GBSW-CpHMD[31,32] titration simulations which explicitly account for ionization-induced conformational changes gave RMSE of 1.69 for a similar OBTRUDE set.[30] The OBTRUDE test indicates that KaML-ESM2 possesses capacity for extrapolation, although its performance is somewhat degraded due to the lack of direct evolutionary information. A more detailed comparison, including additional ML models, is given in Supplemental Data Sheet 6.

Engineered mutations dominate critical errors and learning reduces them. We analyzed the critical errors (misclassification of protonation states) in the 50 holdout tests. Unsurprisingly, the majority of them are associated with OBTRUDEs or engineered mutations, supporting the notion that the KaML-ESM models rely primarily on the evolutionary information. Focusing on the 8 OBTRUDEs, we see that the direction of pK_a shift is correctly predicted, but the magnitude is underestimated with only one exception (Fig. S9, Fig. S10, and Supplemental Data Sheet 8). We then turned to the predictions made by the KaML-ESM2_no_OB model. Interestingly, the above 8 OBTRUDEs also display critical errors and most errors are larger compared to those in the random holdouts (Fig. S11 and Supplemental Data Sheet 8). The improved prediction of OBTRUDEs demonstrates that the KaML task head is able to learn from the representations of some OBTRUDEs and extrapolate to others.

Comparison to other pLLM-based models. Most recently, another pLLM-based model pKAML[29] was published. pKAML differs from KaML-ESMs in many critical aspects. Architecturally, pKAML differs from KaML-ESMs by concatenating embeddings from a pLLM (e.g., ESM2_650M) with those from a protein isoelectric point predictor before passing them to a single fully connected layer. Unlike KaML-ESMs, pKAML uses a single model for both acid and base predictions and is trained directly on experimental pK_a shifts (no pretraining) from a subset of the PKAD-2 database,[17] which is significantly smaller than PKAD-3.[8] Furthermore, no synthetic data augmentation strategy was employed. Although it is not available for independent testing, pKAML achieved an overall RMSE of 0.9 in a single hold-out test and RMSE of 0.97 for EXP67S,[29] both significantly larger than the RMSEs of the KaML-ESM2 or KaML-ESMC models.

Developing the KaML platform for pK_a predictions and visualization. The capacity of the KaML MLPs for learning pK_a values is evident from the t-SNE map of KaML-ESM representations in comparison to the "zero-shot" ESM (Fig. 1c and Fig. S1). To provide an accessible tool for the scientific community, we developed a KaML platform, utilizing KaML-ESM models and the most recent ESM3[12] to generate protein structures for visualization and optional structure-based pK_a predictions by KaML-CBT2 (Fig. 4 left, Supplemental Methods). Both a command-line program and a browser-based GUI (https://kaml.computchem.org) are provided, which support input via protein sequence, UniProt ID, PDB ID, or user-provided PDB file (Supplemental Methods and Fig. S12). As an example, KaML-ESM2 was applied to predict pK_a's of UCHL1 (Fig. 4 right), which belongs to the ubiquitin hydrolase family characterized by a conserved Cys-His-Asp catalytic triad. KaML-ESM2 predicted pK_a of 4.67±0.03 for Cys90, 6.95±0.02 for His161, and 2.37±0.02 for Asp176, allowing functional and mechanistic interpretations (see later discussion).

Application: pK_a prediction for the entire human proteome. Knowledge of protein pK_a values is valuable for functional and mechanistic studies. We applied KaML-ESM2 to the human proteome, which resulted in 1,877,704 pK_a values of Asp, Glu, His, Cys, Tyr, and Lys residues in 18,192 proteins with unique genes (Fig. 5a). This application is not feasible with structure-based models, as most targets lack experimental structures and AlphaFold predictions[35] contain large low-confidence regions. Note, 1,882 proteins were excluded as their sequence lengths exceed 1022 (limit for ESM2). Interestingly, the median predicted pK_a's for all residue types are within 0.3–0.4 units from the corresponding solution model pK_a's (Fig. 5a), even though they were not used as model constraints. This consistency also suggests the absence of systematic errors, at least superficially.

Application: functional annotations and mechanistic interpretations. To test if the pK_a predictions by KaML-ESM2 can be used to understand biology, we mapped 130 cysteines with a predicted pK_a below 5 (i.e., deprotonated at physiological pH) to protein domain families using Pfam[26] (Fig. 5b). Remarkably, except for proteins in the thioredoxin family (many of which have two nucleophilic cysteines), there is a one-to-one mapping between cysteine and

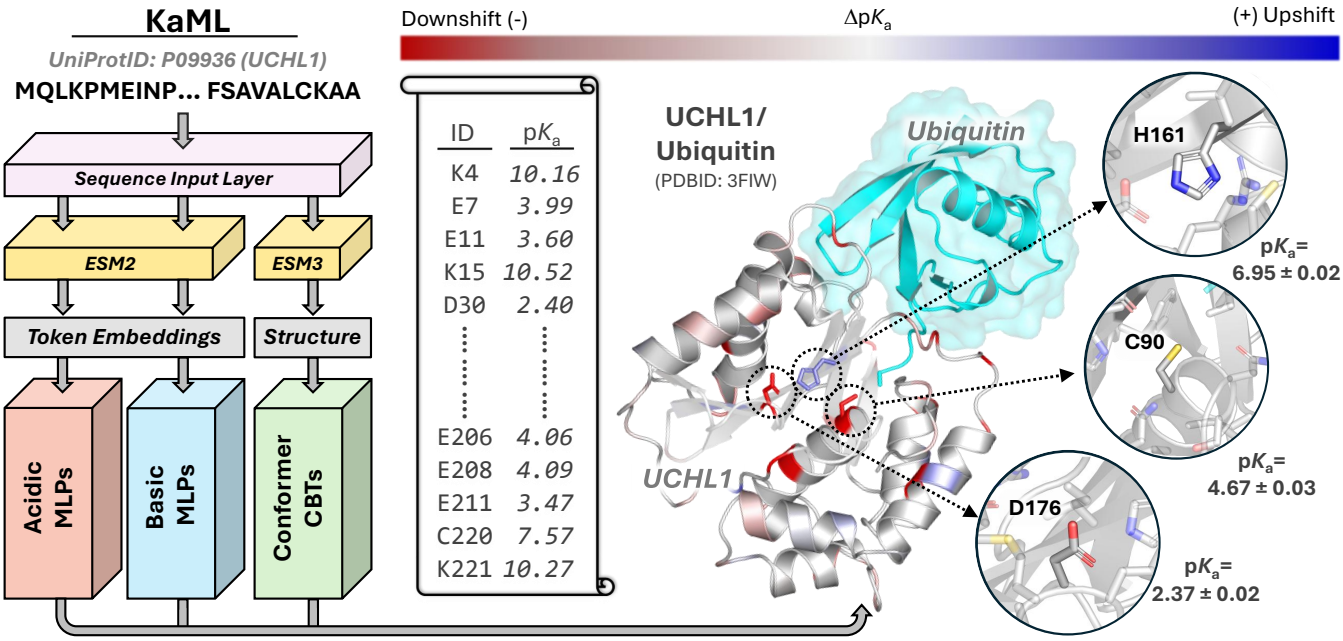

Figure 4: **Architecture of the KaML platform and illustration of pK_a predictions.** KaML accepts a user-provided protein sequence through the sequence input layer. ESM2 (or ESMC) is used to generate token embeddings, while ESM3 predicts a three-dimensional structure if not provided. Embeddings are fed to the acid and base task heads (ensemble of 500 MLPs), respectively. Optionally, the ESM3-derived or user-provided structure is processed for conformer-dependent predictions by KaMLs-CBT2. The predicted acid and base pK_a's, standard errors, and pK_a shifts are given in the output. A vertical scroll illustrates the output for a human deubiquitinase UCHL1. A cartoon representation of UCHL1 is given, with substrate ubiquitin (not included in the prediction) shown in cyan. Residues with up- and down-shifted pK_a's are colored in blue and red, respectively. The catalytic triad and their KaML-ESM2 predicted pK_a's are highlighted and discussed in main text. The inferred catalytic mechanism is displayed in Fig. 5c.

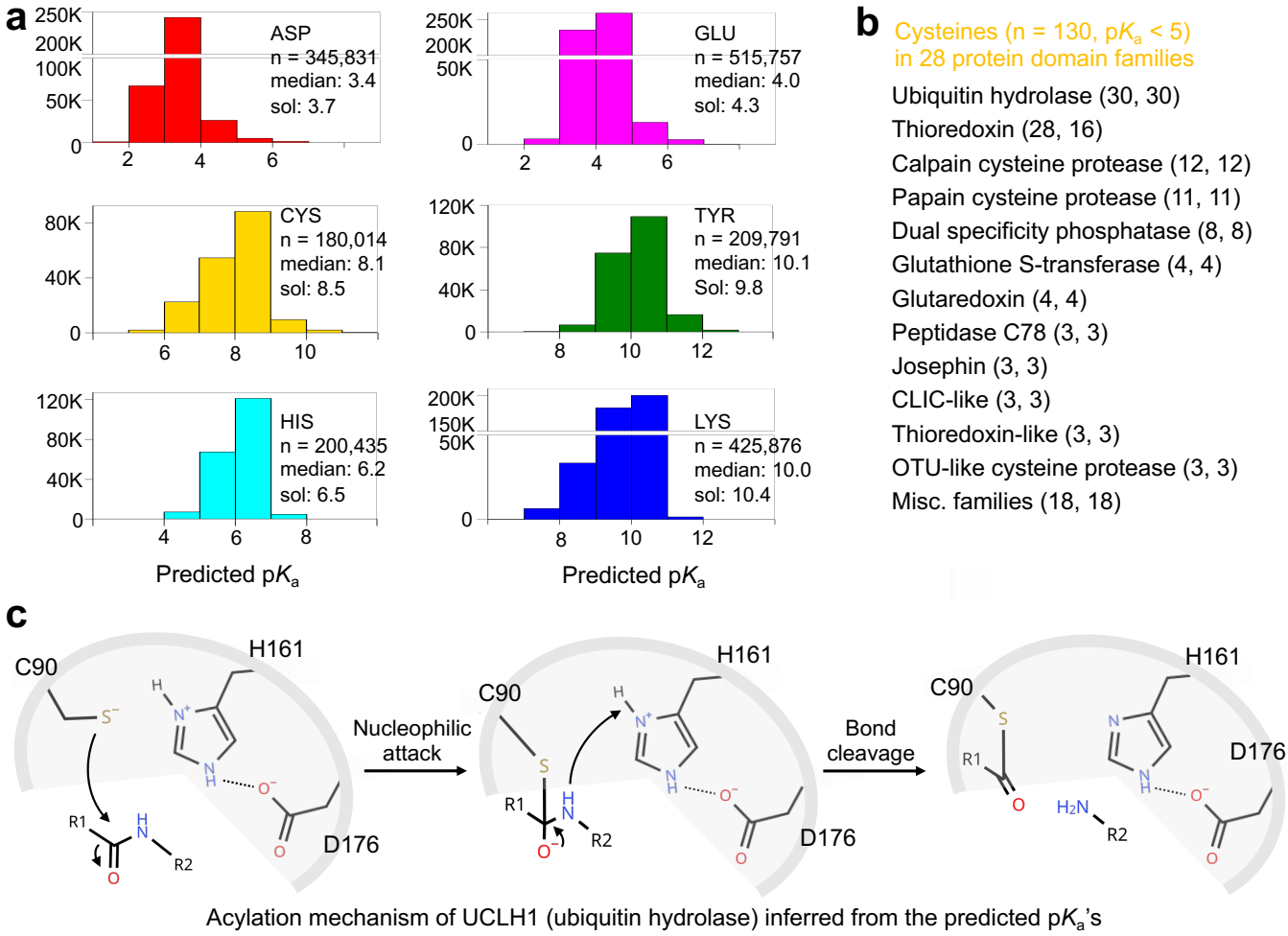

Figure 5: **Human proteome pK_a predictions reveal functional residues and infer catalytic mechanisms. a.** Histograms of the KaML-ESM2 predicted pK_a's of Asp, Glu, His, Cys, Tyr, and Lys residues in the human proteome. Sol: Experimental pK_a's of model tripeptides (GXG, X= titratable).[33] n: total number of residues. A total of 1,877,704 pK_a's of 18,192 unique human proteins (from UniProt[34]) are shown. 1,882 proteins with sequence length $\geq$ 1022 are excluded. **b.** Pfam[26] analysis of cysteines (n=130) with a predicted pK_a <5. Domain families names are followed by the number of low-pK_a cysteines and the number of proteins. Details are given in Supplemental Data Sheet 9. **c.** Predicted pK_a's (Fig. 4) of the catalytic triad in UCLH1 support a widely accepted mechanism for cysteine proteases.

protein, which led us to test a hypothesis that these cysteines with extremely low pK_a's are functional.

We examined the low-pK_a cysteines in 30 ubiquitin hydrolase proteins, which belong to the largest domain family here. According to UniProt[34] records, 29 of them are indeed catalytic nucleophiles. Among the remaining three cysteines, one (U17L4_C89) is inactive ubiquitin hydrolase proteins. To further investigate these cysteines, we zoomed-in on UCHL1_C90, which has a predicted pK_a of 4.67 (Fig. 4). A combined biochemical and crystallography study identified Cys90 as the nucleophile in a catalytic triad composed of Cys90-His161-Asp176.[36] Notably, the predicted pK_a's of His161 and Asp176 are 6.95 and 2.37, respectively. Together, the three pK_a values infer the following catalytic mechanism of UCHL1 (Fig. 5c), which is consistent with the structural data[36] and knowledge of cysteine proteases. Deprotonated Cys90 performs a nucleophilic attack on the carbonyl of the scissile bond, forming a tetrahedral intermediate. Subsequent proton transfer from His161 results in bond cleavage and release of the N-terminal product. This prospective case study illustrates the capability of KaML-ESM for functional annotation and mechanistic interpretation of proteins.

Conclusion

This work challenges the structure-based paradigm by demonstrating that the sequence-based KaML-ESM models (particularly KaML-ESM2) achieve pK_a prediction accuracies well surpassing the structure-based approaches. The extensive data support the notion that the primary sequence encodes not only structure and function but also precise electrostatic properties, which may have been co-optimized through evolution.

KaML-ESM2's extrapolation performance on the OBTRUDE set is striking, outperforming both KaML-ESMC which has a larger representation space and DeepKa,[6,7] which is trained on GBNeck2-CpHMD titration simulations that explicitly account for ionization-induced conformational changes common for OBTRUDEs[28] Error analysis further demonstrates that the KaML task head can learn from the ESM2 representations of some OBTRUDEs and extrapolate to others, suggesting that the KaML-ESM2 performance could be further improved through more extensive training on OBTRUDEs.

While eliminating the requirement for structural information is a major strength of KaML-ESMs, it also represents a limitation. Some proteins function by switching between two macroscopically distinct conformational states, which cannot be captured by the current KaML-ESM framework. While KaML-CBT2 is implemented in the KaML platform to predict conformational state-dependent pK_a's, incorporating sequence information and exploring alternative architectures may improve predictive performance. Incorporating structural information may also enable the prediction of proton tautomers, which currently can be obtained only through constant pH MD simulations.[31]

The KaML platform enables proteomewide electrostatic characterization and a variety of other applications, including identification of functional residues and interpretation of catalytic mechanisms. Furthermore, we envisage the integration of further improve KaML with constant pH MD simulation[3] to model dynamic interplays between protonation state changes and conformational transitions critical to biological functions. Such an integrated approach would further advance our understanding of how electrostatic remodeling drives protein functions, e.g., in proton-coupled gating of ion channels and activation of membrane transporters. Finally, our work demonstrated the power of the GAINES augmentation approach, which provides a general framework to overcome data scarcity challenges in many machine learning applications.

Appendix

Ablation study to examine the contributions from model pretraining (PT) and separation of acid/base models (AB). To evaluate the contributions from PT and AB, we tested four KaML-ESM2 models: (1) no PT and no AB (baseline model); (2) PT only; (3) AB only; and (4) PT and AB. Compared to the baseline model, applying PT decreased the hold-out test root-mean-square error (RMSE) from 0.93 to 0.89, and applying AB decreased the test RMSE from 0.93 to 0.76 (Supplemental Table S1). Moreover, the combination of PT and AB demonstrates a synergistic effect, reducing RMSE from 0.93 to 0.73 (Supplemental Table S1).

Analysis of ESM layers for optimal representation of protein electrostatics. One earlier study found that different aspects of protein evolutionary features such as structures and functions are learned across different layers of the pLLM transformer, with deeper layers attending to residue contact relationships.[23] Considering that protein pK_a's are determined by local environment,[8] we asked if there is a particular ESM2 layer that offers the most accurate representation of protein ionization states. To test this, we trained dedicated KaML-ESM models using embeddings from different layers of ESM2_650M, ESM2_15B, or ESMC_6B. Interestingly, the overall RMSEs of ESM2_650M and ESM2_15B models exhibit multiple local minima, while the RMSE of ESMC decreases monotonically with layer depth (Fig. S4). This difference reflects learning saturation in ESM2 models, where representation learning plateaus for several amino acids prior to the final layer (Tables S3-S4), while ESMC maintains continuous improvement

across all amino acids through the final layers (Table S5). It is noteworthy despite its smaller architecture (33 layers and 640 M parameters), ESM2_650 M achieves comparable performance in learning protein pK_a values.

Acknowledgment

Financial support by the National Institutes of Health (R35GM148261 and R01CA256557) is acknowledged. We thank Marius Wiggert (EvolutionaryScale, PBC, New York, NY, USA.) for advice and facilitating the usage of ESM3 and ESMC. We also thank Ronald Kasl for hardware support. This work was supported by an EvolutionaryScale compute grant.

Supplemental Materials

Supplemental Materials contains Supplemental Methods, Supplemental Tables, Supplemental Figures, and Supplemental Data Sheets.

Supplemental Methods: pretraining dataset; training and testing datasets; the GAINES protocol, model architecture and parameters; model training protocol; implementation of the KaML platform; data for external evaluation; and the pK_a predictions for the entire human proteome.

Supplemental Tables: summary of the pK_a datasets (Table S1); ablation study of pretraining and acid-base model separation (Table S2); transformer layer analysis (Table S3–S5); effect of GAINES training on model performance (Table S6).

Supplemental Figures: t-SNE analysis of ESMC and KaML-ESMC representations (Figure S1); t-SNE analysis of ESM2 and ESMC representations without the OBTRUDEs (Figure S2); Distributions of amino acid-specific pK_a values in the pretraining and PKAD-3 datasets (Figure S3 and S4); Analysis of ESM2 and ESMC layers (Figure S5); RMSE convergence with respect to the number of random data splits (Figure S6); distribution of the GAINES augmented dataset (Figure S7); distribution of the pairwise sequence identity between query and value proteins (Figure S8); error analysis of the KaML-ESM2 and KaML-ESMC predictions for the 50 holdout sets and the OBTRUDE set (Figure S9–S11); and Screenshots of the browser-based KaML platform (Figure S12).

Supplemental Data Sheets: residues without GAINES augmentation (1_GAINES_no_aug). (2_KaML-ESM_holdouts); performance metrics from all 50 holdout tests (KaML-ESM_holdouts); performance metrics from random, EXP67S, OBTRUDE holdout tests as well as external evaluation of KaML and alternative models (3_Test_holdouts; 4_Test_external; 5_Test_EXP67S; 6_Test_OBTRUDEs); OBTRUDE dataset (7_SNase_OBTRUDEs); critical error analysis for SNase OBTRUDEs (8_Error_OBTRUDEs); cysteines from the human proteome with a predicted pK_a below 5 (9_Cys_lowpka_Pfam).

Data Availability

The PKAD-3_v2 database is freely searchable and downloadable at https://database.computchem.org/pkad-3. All training and testing datasets are freely downloadable at https://github.com/JanaShenLab/KaML-ESM. Model weights for KaML-ESM2_650M are accessible at https://doi.org/10.5281/zenodo.17943825. Model weights for KaML-ESMC_6B are accessible at https://doi.org/10.5281/zenodo.17943447. Model weights for KaML-CBT2 are accessible at: https://doi.org/10.5281/zenodo.17943947.

Code Availability

All KaML models are open-source and freely available for non-commercial use. The source code, pretrained model weights, documentation, and instructions are available at https://github.com/JanaShenLab/KaML-ESM.

References

(1) Alexov, E.; Mehler, E. L.; Baker, N.; M. Baptista, A.; Huang, Y.; Milletti, F.; Erik Nielsen, J.; Farrell, D.; Carstensen, T.; Olsson, M. H. M.; Shen, J. K.; Warwicker, J.; Williams, S.; Word, J. M. Progress in the Prediction of pK_a Values in Proteins. *Proteins* **2011**, *79*, 3260–3275.

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

(3) Martins de Oliveira, V.; Liu, R.; Shen, J. Constant pH Molecular Dynamics Simulations: Current Status and Recent Applications. *Curr. Opin. Struct. Biol.* **2022**, *77*, 102498.

(4) Olsson, M. H. M.; 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.

(5) 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 p K_a Predictions

in Proteins. *J. Chem. Theory Comput.* **2022**, *18*, 5068–5078.

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

(7) 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.

(8) Shen, M.; Kortzak, D.; Ambrozak, S.; Bhatnagar, S.; Buchanan, I.; Liu, R.; Shen, J. KaMLs for Predicting Protein p K_a Values and Ionization States: Are Trees All You Need? *J. Chem. Theory Comput.* **2025**, *21*, 1446–1458.

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

(10) Wozniak, S.; Janson, G.; Feig, M. Accurate Predictions of Molecular Properties of Proteins via Graph Neural Networks and Transfer Learning. *Journal of Chemical Theory and Computation* **2025**, *21*, 4830–4845.

(11) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379*, 1123–1130.

(12) Hayes, T.; Rao, R.; Akin, H.; Sofroniew, N. J.; Oktay, D.; Lin, Z.; Verkuil, R.; Tran, V. Q.; Deaton, J.; Wiggert, M.; Badkundri, R.; Shafkat, I.; Gong, J.; Derry, A.; Molina, R. S.; Thomas, N.; Khan, Y. A.; Mishra, C.; Kim, C.; Bartie, L. J.; Nemeth, M.; Hsu, P. D.; Sercu, T.; Candido, S.; Rives, A. Simulating 500 Million Years of Evolution with a Language Model. *Science* **2025**, *387*, 850–858.

(13) ESM Team ESM Cambrian: Revealing the Mysteries of Proteins with Unsupervised Learning. https://www.evolutionaryscale.ai/blog/esm-cambrian.

(14) Isom, D. G.; Castañeda, C. A.; Cannon, B. R.; Velu, P. D.; García-Moreno E., B. Charges in the Hydrophobic Interior of Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 16096–16100.

(15) Isom, D. G.; Castaneda, C. A.; Cannon, B. R.; Garcia-Moreno E., B. Large Shifts in pKa Values of Lysine Residues Buried inside a Protein. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 5260–5265.

(16) Pahari, S.; Sun, L.; Alexov, E. PKAD: A Database of Experimentally Measured pK_a Values of Ionizable Groups in Proteins. *DATABASE* **2019**, *2019*.

(17) Ancona, N.; Bastola, A.; Alexov, E. PKAD-2: New Entries and Expansion of Functionalities of the Database of Experimentally Measured pKa's of Proteins. *J. Comput. Biophys. Chem.* **2023**, 1–10.

(18) Shi, C.; Wallace, J. A.; Shen, J. K. Thermodynamic Coupling of Protonation and Conformational Equilibria in Proteins: Theory and Simulation. *Biophys. J.* **2012**, *102*, 1590–1597.

(19) Kougentakis, C. M.; Grasso, E. M.; Robinson, A. C.; Caro, J. A.; Schlessman, J. L.; Majumdar, A.; García-Moreno E., B. Anomalous Properties of Lys Residues Buried in the Hydrophobic Interior of a Protein Revealed with ^{15} N-Detect NMR Spectroscopy. *J. Phys. Chem. Lett.* **2018**, *9*, 383–387.

(20) Hinton, G. E.; Roweis, S. Stochastic Neighbor Embedding. Adv Neur Inf. Process Sys. 2002.

(21) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Machine Learn. Res.* **2011**, *12*, 2825–2830.

(22) Li, H.; Robertson, A. D.; Jensen, J. H. Very Fast Empirical Prediction and Rationalization of Protein pKa Values. *Proteins* **2005**, *61*, 704–721.

(23) Vig, J.; Madani, A.; Varshney, L. R.; Xiong, C.; Socher, R.; Rajani, N. F. BERTology Meets Biology: Interpreting Attention in Protein Language Models. ICLR 2021. 2021.

(24) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; ukasz Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. NeurIPS Proc. 2017.

(25) Marino, S. M. Protein Flexibility and Cysteine Reactivity: Influence of Mobility on the H-Bond Network and Effects on pKa Prediction. *The Protein Journal* **2014**, *33*, 323–336.

(26) Paysan-Lafosse, T.; Andreeva, A.; Blum, M.; Chuguransky, S. R.; Grego, T.; Pinto, B. L.; Salazar, G. A.; Bileschi, M. L.; Llinares-López, F.; Meng-Papaxanthos, L.; Colwell, L. J.; Grishin, N. V.; Schaeffer, R. D.; Clementel, D.; Tosatto, S. C. E.; Sonnhammer, E.; Wood, V.; Bateman, A. The Pfam Protein Families Database: Embracing AI/ML. *Nucleic Acids Research* **2025**, *53*, D523–D534.

(27) Webb, H.; Tynan-Connolly, B. M.; Lee, G. M.; Farrell, D.; O'Meara, F.; Søndergaard, C. R.; Teilum, K.; Hewage, C.; McIntosh, L. P.;

Nielsen, J. E. Remeasuring HEWL pKa Values by NMR Spectroscopy: Methods, Analysis, Accuracy, and Implications for Theoretical pKa Calculations. *Proteins* **2011**, *79*, 685–702.

(28) Harris, R. C.; Shen, J. GPU-Accelerated Implementation of Continuous Constant pH Molecular Dynamics in Amber: pKa Predictions with Single-pH Simulations. *J. Chem. Inf. Model.* **2019**, *59*, 4821–4832.

(29) Xu, S.; Onoda, A. Accurate and Rapid Prediction of Protein pKa: Protein Language Models Reveal the Sequence-pKa Relationship. *J. Chem. Theory Comput.* **2025**, *xx*, xx.

(30) Wallace, J. A.; Wang, Y.; Shi, C.; Pastoor, K. J.; Nguyen, B.-L.; Xia, K.; Shen, J. K. Toward Accurate Prediction of pKa Values for Internal Protein Residues: The Importance of Conformational Relaxation and Desolvation Energy. *Proteins* **2011**, *79*, 3364–3373.

(31) Khandogin, J.; Brooks, C. L., III Constant pH Molecular Dynamics with Proton Tautomerism. *Biophys. J.* **2005**, *89*, 141–157.

(32) Khandogin, J.; Brooks, C. L., III Toward the Accurate First-Principles Prediction of Ionization Equilibria in Proteins †. *Biochemistry* **2006**, *45*, 9363–9373.

(33) 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.

(34) The UniProt Consortium; Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Adesina, A.; Ahmad, S.; Bowler-Barnett, E. H.; Bye-A-Jee, H.; Carpentier, D.; Denny, P.; Fan, J.; Garmiri, P.; Gonzales, L. J. D. C.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Joshi, V.; Jyothi, D.; Kandasaamy, S.; Lock, A.; Luciani, A.; Luo, J.; Lussi, Y.; Marin, J. S. M.; Raposo, P.; Rice, D. L.; Santos, R.; Speretta, E.; Stephenson, J.; Totoo, P.; Tyagi, N.; Urakova, N.; Vasudev, P.; Warner, K.; Wijerathne, S.; Yu, C. W.-H.; Zaru, R.; Bridge, A. J.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A. H.; Axelsen, K. B.; Bansal, P.; Baratin, D.; Batista Neto, T. M.; Blatter, M.-C.; Bolleman, J. T.; Boutet, E.; Breuza, L.; Gil, B. C.; Casals-Casas, C.; Echioukh, K. C.; Coudert, E.; Cuche, B.; De Castro, E.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gaudet, P.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz, N.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Kerhornou, A.; Mercier, P. L.; Lieberherr, D.; Masson, P.; Morgat, A.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Poux, S.; Pozzato, M.; Pruess, M.; Redaschi, N.; Rivoire, C.; Sigrist, C. J. A.; Sonesson, K.; Sundaram, S.; Sveshnikova, A.; Wu, C. H.; Arighi, C. N.; Chen, C.; Chen, Y.; Huang, H.; Laiho, K.; Lehvaslaiho, M.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Y.; Zhang, J. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Research* **2025**, *53*, D609–D617.

(35) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500.

(36) Boudreaux, D. A.; Maiti, T. K.; Davies, C. W.; Das, C. Ubiquitin Vinyl Methyl Ester Binding Orients the Misaligned Active Site of the Ubiquitin Hydrolase UCHL1 into Productive Conformation. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 9117–9122.