Supporting Information:

Protein Electrostatic Properties are Fine-Tuned Through Evolution

Mingzhe Shen,[†] Guy W. Dayhoff II,[†,‡] 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

List of Tables

S1 Effect of pretraining and separation of acid and base models S-10

S2 Transformer layer analysis for ESM_650M S-10

S3 Transformer layer analysis for ESM_15B S-11

S4 Transformer layer analysis for ESMC S-11

S5 Overall performance of KaML-ESMs S-12

List of Figures

S1 Data split for the pretraining dataset S-13

S2 Data split for the fine-tune (PKAD-3) dataset S-14

S3 t-SNE analysis for ESMC embeddings S-15

S4 Predicted vs. experimental pK_a's in 20 hold-out tests. S-16

S5 Performance of KaML-ESM2 and KaML-ESMC S-17

S6 External evaluation of KaML-ESM S-18

S7 KaML-ESM2 predicted vs. experimental pK_a's in all 20 hold-out tests S-19

S8 KaML-ESMC vs. experimental pK_a's in all 20 hold-out tests S-20

S9 Screen shots of the browser-based KaML GUI S-21

Materials and Methods

Development of the KaML-ESM models

Construction of the pretraining dataset. We used KaML-CBTree[S1] to predict the pK_a values of all proteins in the pkPDB[S2] database (structures fixed using PDBFixer from OpenMM[S3]) which are composed of proteins from the PDB. In total, 5,616,944 pK_a values were predicted for 1,570,916 unique residues in 32,418 unique proteins. To reduce computational cost, we created a subset as the model pretraining dataset. For each titratable amino acid, we randomly sampled predicted pK_a values at a ratio of 20:1 relative to the number of entries in PKAD-3. In total, the pretraining set contains 35,080 pK_a's from 10,400 Asp, 11,580 Glu, 5,860 His, 2,503 Cys, 820 Tyr, and 3,917 Lys in 9,945 proteins. Residues in the PKAD-3 database were excluded from the pretraining dataset to avoid data leakage in later model training/testing.

To ensure a precise correspondence between the sequence and the protein studied by experiment regardless of the residue-specific resolution, we extracted sequences directly from the PDB files. This extraction process involved identifying resolved residues from ATOM records while detecting unresolved segments through multiple methods: analyzing discontinuities in residue numbering, identifying unusually large distances between adjacent backbone atoms, and incorporating segments explicitly annotated as missing in the PDB metadata (`REMARK 465`). Non-standard amino acids were carefully mapped to their closest canonical equivalents (e.g., selenomethionine [MSE] to methionine [M], methyllysine [MLY] to lysine [K], phosphoserine [SEP] to serine [S]), or otherwise represented as `X` when no direct canonical mapping exists. This rigorous and automated sequence extraction procedure was implemented using the `extract_sequence.py` script from *PDB doctor*, an unpublished in-house software suite available at https://github.com/wayyne/pdb_doctor.

Training dataset and protocol for data splitting. Following pretraining, the PKAD-3 database[S1] was used for model training (i.e., fine-tuning), validation, and testing. We used the same data splitting protocol as in the development of KaML-CBTree and KaML-GAT models.[S1] Briefly, we labeled a unique residue as a unique combination of Uniprot ID + the Uniprot residue ID (Uni_resid). To account for mutants and multiple conformational states, we defined these proteins as 'Uniprot ID-mutation or conformational state'. For example, P0AEG4-H32L represents H32L mutation on DsbA (P0AEG4). For SNase (P00644), which has multiple background constructs, e.g., WT, PHS, and Δ+PHS, we used P00644-V66D, P00644-WT-V66D, and P00644-PHS-V66D to represent the Δ+PHS, WT, and PHS constructs, respectively. Data splitting into training, validation, and test sets was conducted using the 'StratifiedGroupKFold' strategy adapted from the sklearn python package.[S4] Here group refers to a unique residue, and stratification was made based on the experimental pK_a values. The stratification bins were manually selected such that they can separate residues with different pK_a values, while at the same time each bin can be further divided into training, validation, and test sets.

From the PKAD-3 database,[S1] we first randomly sampled 10% as the unseen test set with the 'StratifiedGroupKFold' strategy. The remaining 90% of the data set was then split into training and validation sets in 9:1 ratio with the same splitting protocol. This data splitting protocol was repeated 20 times independently. For the pretraining dataset, we used a 9:1 ratio for training and validation data with the same strategy.

Architecture of the KaML-ESM models. The KaML-ESM models were built and trained using the PyTorch package.[S5] A multilayer perceptron (MLP) with 3 hidden layers was built as a task head on top of an ESM model. The number of neurons in KaML-ESM2_650M, KaML-ESM2_15B, and KaML_ESMC models was 1280-512-256-32-1, 5120-2048-1024-256-1, and 2560-1024-512-64-1, respectively. The rectified linear unit (ReLU) activation function is applied to each hidden layer to add non-linearity. The output of the last neuron

is converted to the prediction through a linear activation function. Batch normalization is applied to all layers except the last one. For model training, we used a learning rate of 0.0005 and a dropout rate of 0.2 with a batch size of 64 and a maximum epoch of 200. The Adam optimizer is used to minimize the mean squared error (MSE) loss. Early stopping is applied to prevent overfitting: model training is terminated if the MSE loss does not decrease within the next 10 epochs with a tolerance of 0.1.

The evolutionary scale models (ESMs) are protein large language models (pLLMs) based on the BERT[S6] style transformer[S7] architecture. We used ESM2[S8] and ESMC,[S9] which are encoder-only transformers trained using masked learning of 50 million protein sequences. ESM2_t33_650M_UR50D (ESM_650M) model has 33 transformer layers (20 attention heads each) and 650 million parameters. ESM2_t48_15B_UR50D (ESM2_15B) has 48 transformer layers (40 attention heads each) and 15 billion parameters. ESMC-6B-2024-12 (ESMC) has 80 transformer layers (40 attention heads each) and 6 billion parameters. This model, which is focused on representation learning of the underlying biology, is a parallel model to the generative model ESM3.[S10] For titratable residues of interest, the token embeddings were extracted from ESM2_650M, ESM2_15B, or ESMC, which have dimensions of 1280, 5120, and 2560, respectively. Note, ESM2_650M can be loaded to GPUs with ~ 10 GB memory while the ESM2-15B requires a memory space of ~ 100 GB and ESMC embeddings can only be extracted using the forge-API (https://forge.evolutionaryscale.ai/) with daily token and speed limitations.

Training the ensemble KaML-ESM models. Due to the distinct physiochemical properties of acidic (Asp, Glu, Cys, and Tyr) and basic (His and Lys) residues, we trained acid and base models separately. This strategy has been shown to enhance model performance (main text). The models were trained on the pK_a shifts relative to the solution pK_a values of model peptides $CH_3COGXGNH_2$[S11] or $CH_3COAAXAANH_2$.[S12] Accordingly, the solution pK_a's are 3.7/3.9 for Asp, 4.3/4.3 for Glu, 6.5/6.5 for His, 8.6/8.5 for Cys, 9.8/9.8

for Tyr, and 10.4/10.3 for Lys. The pK_a shifts are converted back to pK_a values after the model training or prediction. We pretrained our model using a synthetic dataset composed of pK_a predicted by the KaML-CBtree model (Fig. S1). The Glorot algorithm[S13] is used to initialize the model weights in the pre-training stage. In the model fine-tuning stage, for each training/test data splitting, 10 models (with different random seeds) are independently trained based on different 9:1 training:cross-validation splits. Thus, each prediction is made by an ensemble model that gives the average predictions made by 10 models. In the production stage, we retrained KaML-ESM (which comprises KaML-ESM2a and KaML-ESMCb) using the entire PKAD-3 dataset. Here we performed 20 training:cross-validation splits and for each split 10 models were trained. Therefore, the final KaML-ESM is an ensemble of 200 models.

Implementation of the KaML platform

The KaML-ESM platform comprises multiple modular stages managed within a unified Python framework, facilitating ease of use and reproducibility. Protein sequences or structures (specified via UniProt identifiers, PDB files, PDB IDs, or FASTA sequences) are initially retrieved through established web services (UniProt, RCSB) or processed directly when provided by users. For sequences without structural information, computational folding is performed using the remote ESM3-medium model accessible through the Forge service.

Following retrieval and preprocessing, sequences and structures are parsed to identify titratable residues. Embeddings are subsequently computed from these sequences using a pLLM, specifically ESM2 (esm2_t33_650M_UR50D)[S8] or ESMC (esmc-6b-2024-12),[S9] with model selection determined by residue type (acidic or basic). By default, embeddings for acidic residues are extracted from ESM2 and processed using a multi-layer perceptron (MLP) configured as 1280→512→256→32→1 neurons with dropout regularization. Embeddings for basic residues are derived from ESMC and processed through a sep-

arate MLP architecture of 2560→1024→512→64→1 neurons, similarly utilizing dropout regularization. Details are given in paragraph Architecture of the KaML-ESM models.

Predictions from acidic and basic channels are generated via parallelized inference using ensemble models, supported by statistical normalization strategies developed from extensive training datasets. The acidic and basic channel predictions are consolidated to yield the predicted pK_a values, shifts, and standard error metrics. An additional, independently operated modular conformer channel is integrated, allowing users to employ alternative conformer-sensitive predictive models in cases where distinct conformations significantly impact predicted pK_a values. Currently, this channel employs the previously validated KaMLs-CBTree model.[S1] Predictions from the conformer channel are reported independently in the final output.

The command-line KaML platform. The KaML platform (https://github.com/JanaShenLab/KaML-ESM) is openly available, accompanied by pretrained MLP ensemble weights. Specifically provided are pretrained weights for the acidic ESMC channel and for cysteine residues using layer 33 of the ESM2 model. The platform is designed to offer users flexibility in specifying foundational models independently for each prediction channel. For example, users can override default settings to utilize ESMC for both acidic and basic channels when analyzing sequences exceeding the 1022-residue limit of ESM2, as ESMC supports sequence lengths up to 2046 residues. In these scenarios, the platform automatically utilizes ESMC for the acidic channel, emphasizing the practical benefit of including these pretrained weights. Additionally, users may optionally select alternative layers within ESM2, such as employing layer 33 specifically for cysteine residues, reflecting enhanced performance demonstrated in the primary analysis. By default, however, the platform utilizes layer 31.

Finally, predictions are integrated back into structural outputs by annotating the structure files with residue-specific predictions in the B-factor field, facilitating visualization and

downstream structural analysis. Prediction results, distinguishing between acidic/basic channel outputs and independent conformer channel predictions, are documented in CSV files, enhancing clarity and accessibility for further analysis.

The browser-based KaML application. We provide an easy-to-use online browser-based GUI (https://kaml.computchem.org/) for non-commercial academic and research purposes. Users can provide either the protein sequence, Uniprot ID, PDB ID, or PDB files of interest. By clicking the 'Run pipeline' button, the complete KaML-ESM model will run on our web server. A PDB format file with the B-factor column filled with the predicted pK_a shifts and a CSV file with predicted pK_a values for all titratable residues can be downloaded. Note, to use the KaML-ESMC models, users need to provide ESM forge API tokens, which is freely applicable at https://www.evolutionaryscale.ai/blog/esm-cambrian. The screenshots of the web KaML application is given in Fig. S9.

Data for external evaluations

Newly curated experimental data. The recently published PKAD-3 database includes 1,167 pK_a's of 992 unique residues in 247 proteins (wild type or mutant).[S1] Note, this database only includes experimental pK_a values published before April 2023. In order to compile a dataset for external evaluation, we conducted a literature search for publications after April 2023 by applying the keyword 'pK_a' in the title and/or abstract on the PubMed website. We then manually verified the pK_a values in the publications. Furthermore, we found a few SNase mutant pK_a's based on NMR measurements[S14] that were missed when we curated PKAD-3. In total, 55 pK_a's from 16 proteins were found. These new entries were used in the external evaluation of the models and are now added to the PKAD-3 database (http://database.computchem.org/pkad-3). Currently, only the first 50 entries are displayed on the website, but the entire database is searchable and downloadable.

Data for proteome-wide pK_a predictions. The activity-based protein profiling (ABPP) a chemical proteomic technology for discovering covalently ligandable sites across the proteome.[S15] We collected an ABPP dataset from 10 publications.[S16–S25] In total, the ABPP dataset contains 98,921 Asp, 147,092 Glu, 48,705 His, 41,634 Cys, 51,825 Tyr, and 121,660 Lys (sum up to 509,837 unique residues) from 3,892 unique genes, most of which do not have available crystal structures.

Supplemental Tables

Table S1: Effect of pretraining and separating acid and base models on the pK_a prediction errors of KaML-ESM2

	All	Asp	Glu	Cys	Tyr	His	Lys
No PT + No AB	0.93 ± 0.04	0.95 ± 0.09	0.71 ± 0.04	1.03 ± 0.11	1.54 ± 0.19	0.96 ± 0.05	0.99 ± 0.09
PT only	0.89 ± 0.03	0.93 ± 0.08	0.73 ± 0.04	1.05 ± 0.11	1.44 ± 0.17	0.82 ± 0.04	0.88 ± 0.10
AB only	0.76 ± 0.02	0.71 ± 0.05	0.70 ± 0.04	1.00 ± 0.09	1.74 ± 0.19	0.79 ± 0.12	0.59 ± 0.05
PT + AB	0.73 ± 0.03	0.72 ± 0.06	0.67 ± 0.04	1.01 ± 0.09	1.65 ± 0.17	0.67 ± 0.03	0.55 ± 0.04

Overall and amino acid-specific RMSE±standard error of the pK_a predictions by the KaML-ESM2 model using 20 20 hold-out tests (same as in our previous work[S1]). PT: pretraining, AB: training acidic and basic models separately. Residue embeddings were extracted from ESM2_650M layer 33.

Table S2: pK_a prediction errors of the models trained with the embeddings from different layers of ESM_650M

	All	Asp	Glu	Cys	Tyr	His	Lys
	Without pretraining						
layer33	0.76 ± 0.02	0.71 ± 0.05	0.70 ± 0.04	1.00 ± 0.09	1.74 ± 0.19	0.79 ± 0.04	0.59 ± 0.05
layer32	0.77 ± 0.02	0.70 ± 0.04	0.69 ± 0.04	1.04 ± 0.09	1.75 ± 0.20	0.84 ± 0.04	0.61 ± 0.05
layer31	0.75 ± 0.02	0.68 ± 0.04	0.62 ± 0.03	1.23 ± 0.11	1.58 ± 0.17	0.81 ± 0.05	0.71 ± 0.07
layer30	0.79 ± 0.02	0.71 ± 0.04	0.74 ± 0.03	1.15 ± 0.10	1.43 ± 0.18	0.85 ± 0.04	0.65 ± 0.05
layer29	0.82 ± 0.02	0.74 ± 0.03	0.74 ± 0.04	1.30 ± 0.14	1.50 ± 0.20	0.92 ± 0.05	0.62 ± 0.06
layer28	0.81 ± 0.02	0.73 ± 0.04	0.75 ± 0.04	1.23 ± 0.11	1.49 ± 0.22	0.89 ± 0.05	0.61 ± 0.06
layer27	0.81 ± 0.02	0.74 ± 0.04	0.74 ± 0.04	1.25 ± 0.11	1.56 ± 0.22	0.89 ± 0.05	0.64 ± 0.06
layer26	0.86 ± 0.03	0.78 ± 0.04	0.79 ± 0.04	1.36 ± 0.13	1.66 ± 0.25	0.90 ± 0.04	0.70 ± 0.07
layer25	0.87 ± 0.03	0.81 ± 0.04	0.79 ± 0.04	1.48 ± 0.13	1.71 ± 0.26	0.91 ± 0.04	0.68 ± 0.06
layer24	0.87 ± 0.03	0.83 ± 0.04	0.81 ± 0.04	1.36 ± 0.13	1.58 ± 0.24	0.91 ± 0.04	0.68 ± 0.07
layer23	0.89 ± 0.03	0.85 ± 0.04	0.82 ± 0.04	1.42 ± 0.13	1.57 ± 0.23	0.90 ± 0.04	0.71 ± 0.07
layer22	0.89 ± 0.03	0.84 ± 0.04	0.81 ± 0.04	1.45 ± 0.14	1.49 ± 0.23	0.91 ± 0.04	0.72 ± 0.07
layer21	0.87 ± 0.02	0.84 ± 0.03	0.76 ± 0.04	1.43 ± 0.12	1.49 ± 0.20	0.92 ± 0.04	0.74 ± 0.08
layer20	0.89 ± 0.03	$0.0.83 \pm 0.04$	0.80 ± 0.04	1.45 ± 0.16	1.64 ± 0.21	0.91 ± 0.04	0.79 ± 0.07
layer19	0.88 ± 0.03	0.83 ± 0.04	0.77 ± 0.04	1.41 ± 0.16	1.76 ± 0.25	0.92 ± 0.04	0.79 ± 0.07
layer18	0.87 ± 0.03	0.84 ± 0.04	0.77 ± 0.04	1.43 ± 0.14	1.60 ± 0.22	0.88 ± 0.04	0.80 ± 0.08
layer17	0.87 ± 0.03	0.85 ± 0.04	0.76 ± 0.04	1.58 ± 0.12	1.51 ± 0.22	0.89 ± 0.04	0.79 ± 0.07
	With pretraining						
layer33	0.73 ± 0.03	0.72 ± 0.06	0.67 ± 0.04	1.01 ± 0.09	1.65 ± 0.17	0.67 ± 0.03	0.55 ± 0.04
layer32	0.71 ± 0.03	0.69 ± 0.05	0.65 ± 0.04	1.14 ± 0.10	1.47 ± 0.14	0.67 ± 0.03	0.61 ± 0.04
layer31	0.68 ± 0.02	0.61 ± 0.04	0.58 ± 0.04	1.11 ± 0.09	1.54 ± 0.16	0.68 ± 0.03	0.69 ± 0.07
layer30	0.75 ± 0.02	0.71 ± 0.04	0.75 ± 0.04	1.18 ± 0.11	1.34 ± 0.16	0.65 ± 0.04	0.62 ± 0.06
layer29	0.75 ± 0.02	0.70 ± 0.04	0.72 ± 0.04	1.32 ± 0.12	1.35 ± 0.18	0.69 ± 0.03	0.55 ± 0.05
layer28	0.74 ± 0.02	0.70 ± 0.04	0.72 ± 0.04	1.15 ± 0.08	1.39 ± 0.19	0.73 ± 0.03	0.50 ± 0.04

Overall and amino acid-specific RMSE±standard error of the pK_a predictions by the KaML-ESM2 model using 20 hold-out tests (same as in our previous work[S1]). Residue embeddings were extracted from the last 6 layers of ESM2_650M. Layer 31 gives the lowest overall RMSE, while the lowest RMSEs (red) for Cys, Tyr, His, and Lys pK_a predictions are from other layers. Pre-training was performed for all models.

Table S3: pK_a prediction errors of the models trained with the embeddings from different layers of ESM2_15B (no pretraining)

	All	Asp	Glu	Cys	Tyr	His	Lys
layer48	0.74 ± 0.03	0.70 ± 0.05	0.60 ± 0.03	1.16 ± 0.08	1.70 ± 0.20	0.84 ± 0.04	0.61 ± 0.05
layer47	0.73 ± 0.02	0.68 ± 0.04	0.60 ± 0.02	1.07 ± 0.08	1.73 ± 0.20	0.85 ± 0.04	0.64 ± 0.06
layer46	0.74 ± 0.02	0.69 ± 0.04	0.61 ± 0.03	1.09 ± 0.09	1.62 ± 0.20	0.85 ± 0.04	0.62 ± 0.06
layer45	0.74 ± 0.03	0.70 ± 0.04	0.62 ± 0.04	1.07 ± 0.10	1.51 ± 0.20	0.85 ± 0.04	0.63 ± 0.05
layer44	0.76 ± 0.03	0.68 ± 0.04	0.66 ± 0.04	1.11 ± 0.10	1.53 ± 0.21	0.87 ± 0.04	0.65 ± 0.06
layer43	0.76 ± 0.03	0.70 ± 0.04	0.66 ± 0.04	1.10 ± 0.11	1.48 ± 0.22	0.87 ± 0.04	0.65 ± 0.06
layer42	0.77 ± 0.03	0.70 ± 0.04	0.66 ± 0.04	1.10 ± 0.12	1.47 ± 0.23	0.90 ± 0.04	0.64 ± 0.06

Overall and amino acid-specific RMSE±standard error of the pK_a predictions by the KaML-ESM2 model using 20 hold-out tests (same as in our previous work[S1]). The layer that gives the lowest overall RMSE is highlighted in bold font. The lowest amino acid-specific RMSE is highlighted in red.

Table S4: pK_a prediction errors of the the models trained with the embeddings from different layers of KaML-ESMC (no pretraining)

	All	Asp	Glu	Cys	Tyr	His	Lys
layer80	0.70 ± 0.03	0.64 ± 0.04	0.61 ± 0.03	1.22 ± 0.13	1.46 ± 0.17	0.74 ± 0.04	0.53 ± 0.04
layer79	0.74 ± 0.03	0.68 ± 0.05	0.64 ± 0.03	1.47 ± 0.12	1.41 ± 0.16	0.74 ± 0.04	0.58 ± 0.04
layer78	0.73 ± 0.03	0.67 ± 0.05	0.62 ± 0.03	1.30 ± 0.12	1.58 ± 0.21	0.77 ± 0.04	0.57 ± 0.04
layer77	0.76 ± 0.03	0.69 ± 0.05	0.64 ± 0.03	1.31 ± 0.15	1.73 ± 0.26	0.79 ± 0.04	0.58 ± 0.04
layer76	0.75 ± 0.03	0.67 ± 0.05	0.64 ± 0.03	1.35 ± 0.14	1.75 ± 0.25	0.76 ± 0.04	0.60 ± 0.04
layer75	0.75 ± 0.03	0.67 ± 0.05	0.66 ± 0.04	1.36 ± 0.15	1.71 ± 0.25	0.78 ± 0.04	0.59 ± 0.04
layer74	0.77 ± 0.03	0.69 ± 0.05	0.65 ± 0.04	1.39 ± 0.14	1.68 ± 0.25	0.83 ± 0.04	0.57 ± 0.04
layer73	0.78 ± 0.03	0.70 ± 0.06	0.68 ± 0.04	1.42 ± 0.12	1.61 ± 0.23	0.84 ± 0.04	0.57 ± 0.05
layer72	0.78 ± 0.03	0.70 ± 0.06	0.70 ± 0.04	1.33 ± 0.13	1.55 ± 0.24	0.84 ± 0.03	0.58 ± 0.06
layer71	0.78 ± 0.03	0.70 ± 0.05	0.69 ± 0.04	1.36 ± 0.11	1.52 ± 0.24	0.83 ± 0.03	0.60 ± 0.05
layer70	0.79 ± 0.03	0.70 ± 0.05	0.71 ± 0.04	1.40 ± 0.12	1.59 ± 0.23	0.83 ± 0.04	0.63 ± 0.06
layer69	0.80 ± 0.03	0.72 ± 0.06	0.71 ± 0.04	1.47 ± 0.11	1.53 ± 0.22	0.84 ± 0.03	0.64 ± 0.06
layer68	0.80 ± 0.03	0.72 ± 0.06	0.72 ± 0.04	1.44 ± 0.12	1.47 ± 0.22	0.84 ± 0.04	0.66 ± 0.06
layer67	0.81 ± 0.03	0.73 ± 0.06	0.73 ± 0.04	1.46 ± 0.13	1.49 ± 0.23	0.85 ± 0.04	0.66 ± 0.07
layer66	0.81 ± 0.03	0.74 ± 0.05	0.73 ± 0.04	1.46 ± 0.14	1.44 ± 0.23	0.85 ± 0.03	0.67 ± 0.06
layer65	0.81 ± 0.03	0.73 ± 0.05	0.74 ± 0.04	1.44 ± 0.13	1.47 ± 0.24	0.85 ± 0.04	0.67 ± 0.06
layer64	0.82 ± 0.03	0.75 ± 0.06	0.74 ± 0.04	1.46 ± 0.12	1.49 ± 0.24	0.86 ± 0.04	0.67 ± 0.06
layer63	0.82 ± 0.03	0.75 ± 0.06	0.74 ± 0.04	1.45 ± 0.12	1.55 ± 0.24	0.85 ± 0.04	0.69 ± 0.06
layer62	0.82 ± 0.03	0.76 ± 0.06	0.74 ± 0.04	1.49 ± 0.13	1.52 ± 0.23	0.85 ± 0.04	0.69 ± 0.06
layer61	0.83 ± 0.03	0.77 ± 0.06	0.73 ± 0.04	1.48 ± 0.12	1.47 ± 0.21	0.86 ± 0.04	0.71 ± 0.06
layer60	0.83 ± 0.03	0.76 ± 0.06	0.74 ± 0.04	1.46 ± 0.11	1.49 ± 0.23	0.87 ± 0.04	0.71 ± 0.06

Overall and amino acid-specific RMSE±standard error of the pK_a predictions by the KaML-ESM2 model using 20 hold-out tests (same as in our previous work[S1]). The layer that gives the lowest overall RMSE is highlighted in bold font. The lowest amino acid-specific RMSE is highlighted in red.

Table S5: Overall performance of KaML-ESMs for predicting protein pK_a's and protonation states and comparison to the structure-based KaML-CBTree and popular empirical PROPKA3 models[a]

	KaML-ESM2	KaML-ESMC	KaML-CBTree	PROPKA3
RMSE	0.68 ± 0.02	0.68 ± 0.03	0.77 ± 0.02	1.20 ± 0.03
PCC	0.96 ± 0.01	0.96 ± 0.01	0.95 ± 0.01	0.87 ± 0.01
MAXE	3.07 ± 0.15	3.25 ± 0.23	3.40 ± 0.10	5.04 ± 0.10
Classification of protonation states at pH 7[b]				
Pre (prot)	0.97	0.97	0.97	0.85
Rec (prot)	0.96	0.93	0.93	0.86
Pre (dep)	0.99	0.98	0.99	0.97
Rec (dep)	0.99	0.99	0.98	0.97
CER[c]	$33/2607$	$43/2653$	$46/2635$	$143/2673$

[a]The metrics of KaML-CBTree[S1] and PROPKA3[S26] are taken from Ref.[S1] [b]Prediction is based on the probability of protonation given a predicted pK_a (see main text). [c]Critical error rate (CER) refers to the percentage of predictions misclassifying protonated as deprotonated or vice versa. Precision (Pre) and recall (Rec) were calculated for protonated (prot) and deprotonated (dep) states after accumulating the predictions from all 20 holdout test sets.

Supplemental Figures

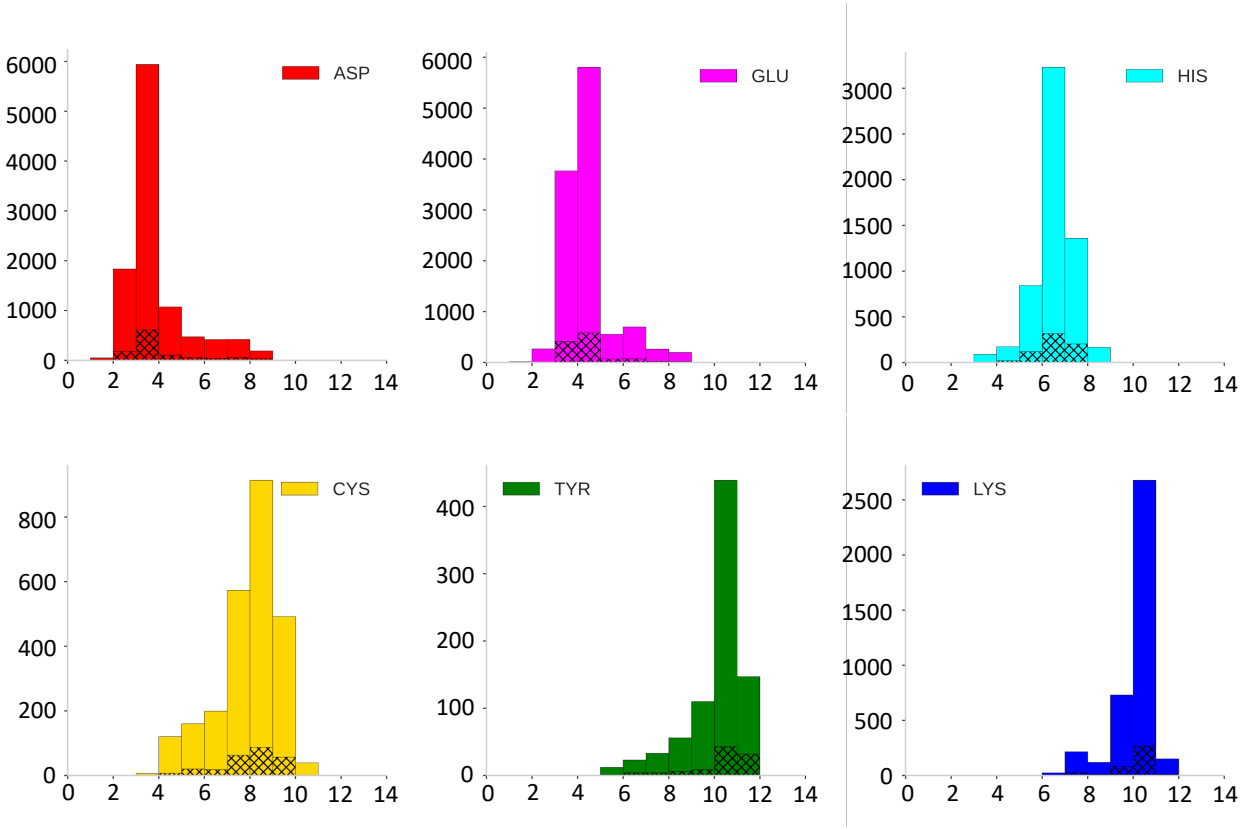

Figure S1: **Train-validation split for the pretraining dataset.** Histograms of the KaML-CBTree calculated pK_a values for pre-training. The dashed bins represent the validation dataset

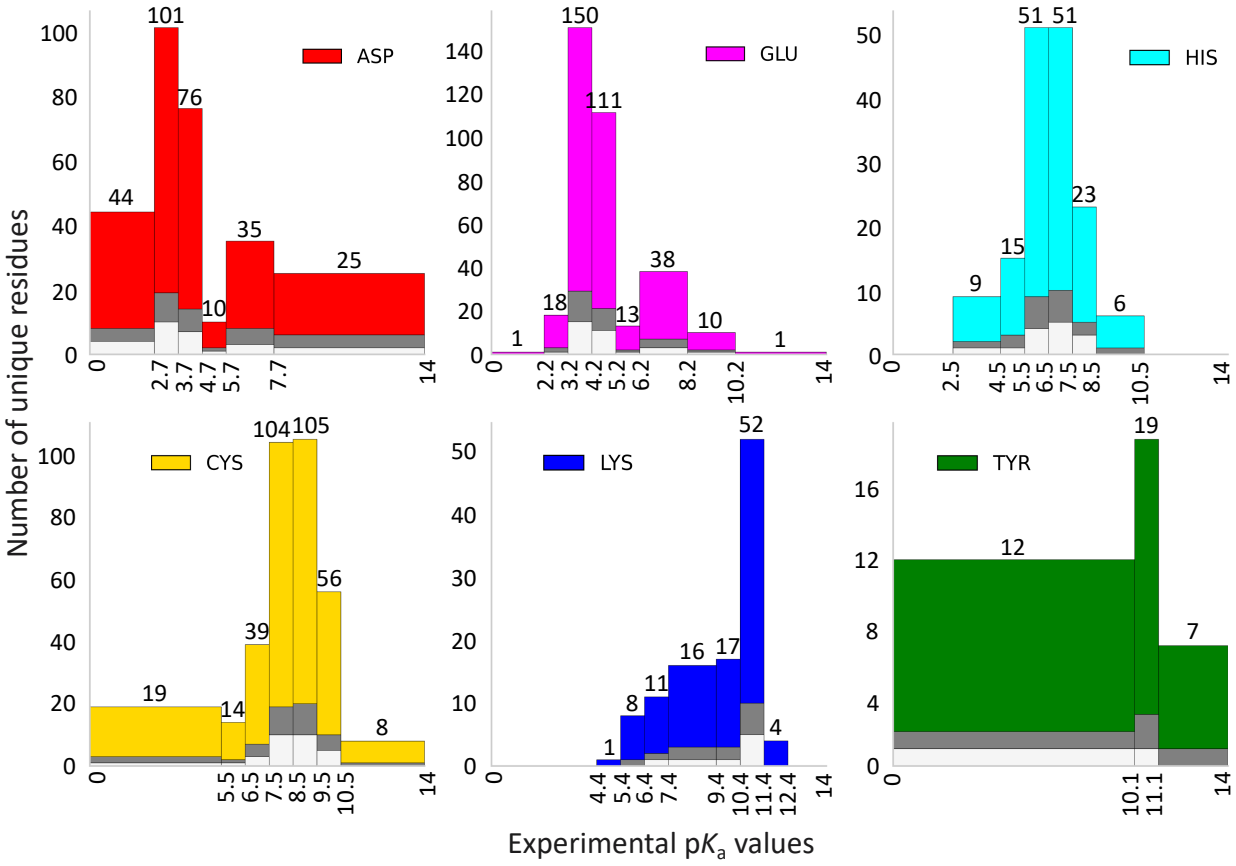

Figure S2: **Train-validation-test split.** The binning scheme used in this study for each residue type. The colored bins represent the overall dataset, gray bins represent the validation dataset, and the white bins represent the test dataset. The data splitting protocol is repeated 20 times independently and the first spiting is shown here as an illustration.

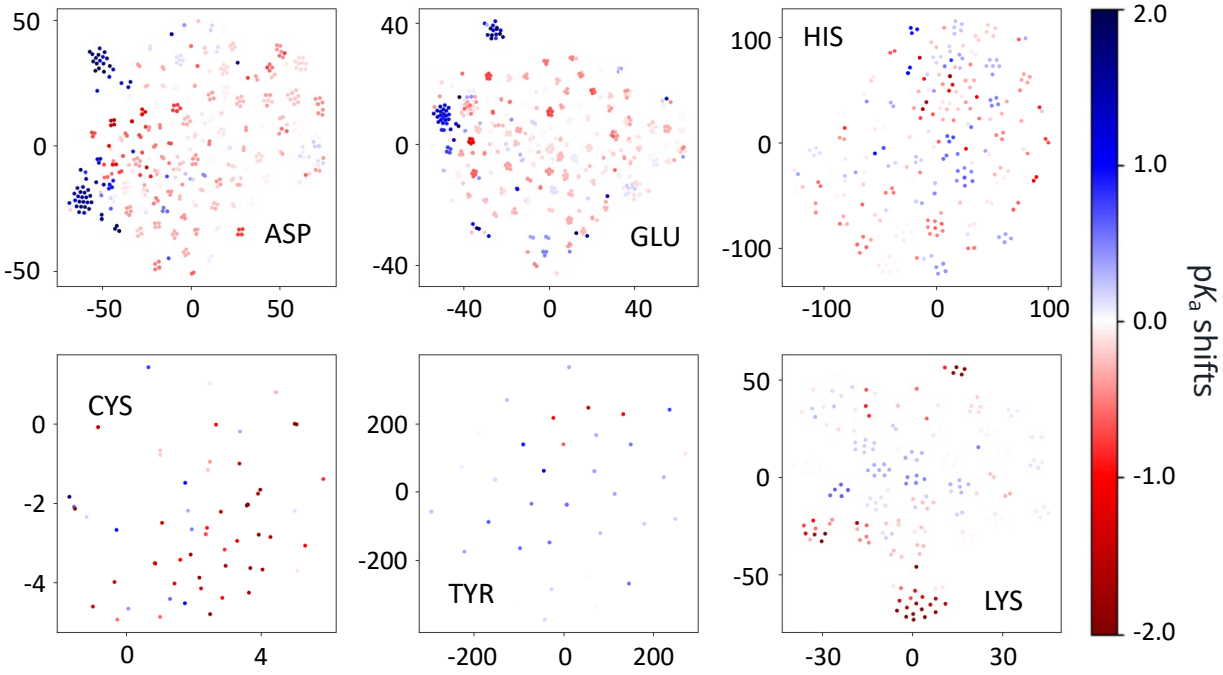

Figure S3: **t-SNE analysis of the residue embeddings from ESMC.** t-SNE visualization of the residue embeddings (2560-digit) extracted from layer 80 of the ESMC model. Residues with up-shifted and down-shifted pK_a's are colored red and blue, respectively.

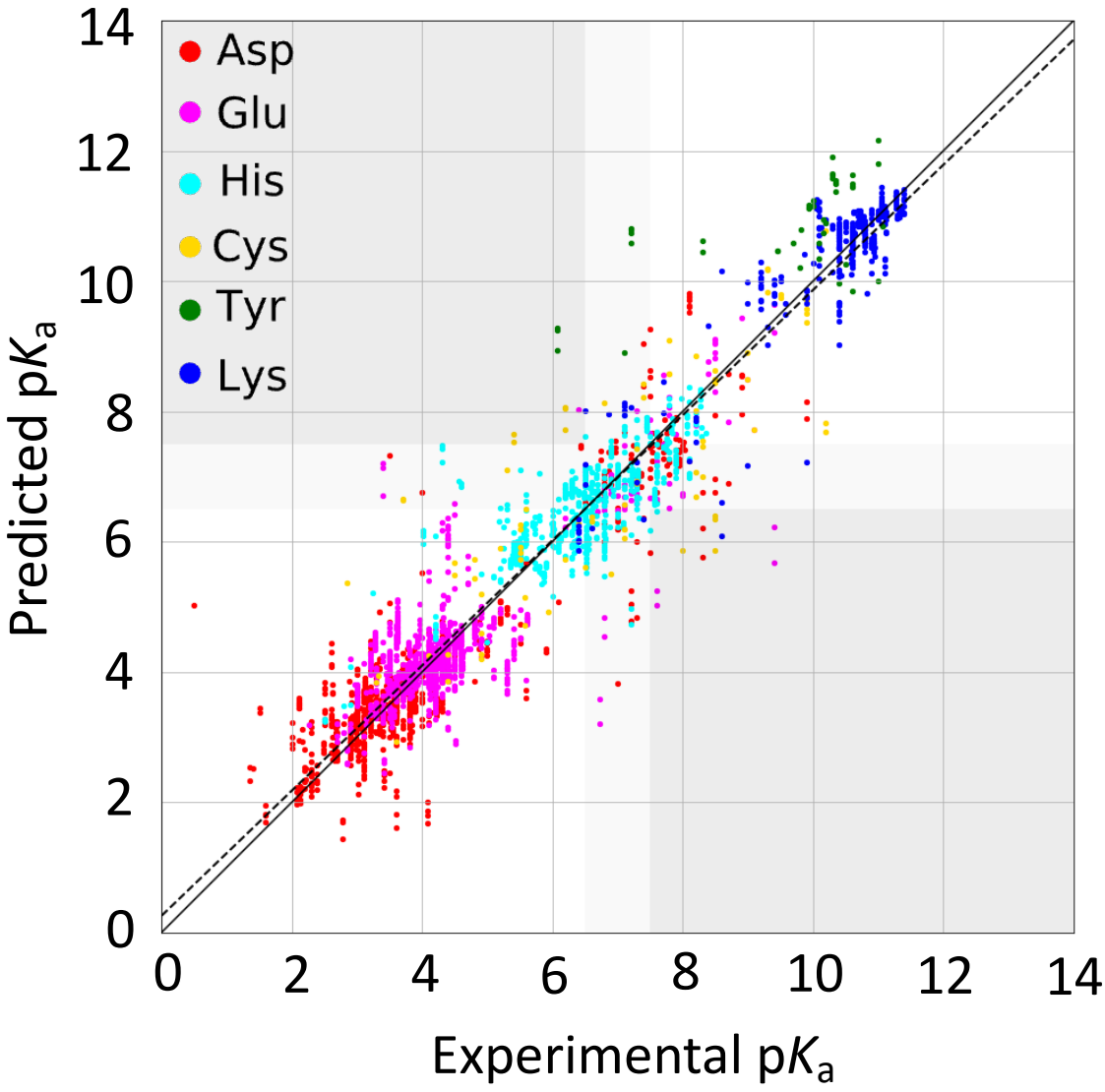

Figure S4: **Experimental vs predicted pK_a's in all 20 hold-out tests.** The RMSE and PCC is 0.65 and 0.96, respectively. Solid line is the identity and the dotted line is a linear fit. Data points are color-coded by amino acid. The white regions indicate correct protonation state classifications. The dark gray regions highlight the critical errors. The light gray regions indicate titrating pK_a's which were excluded from the classification analysis.

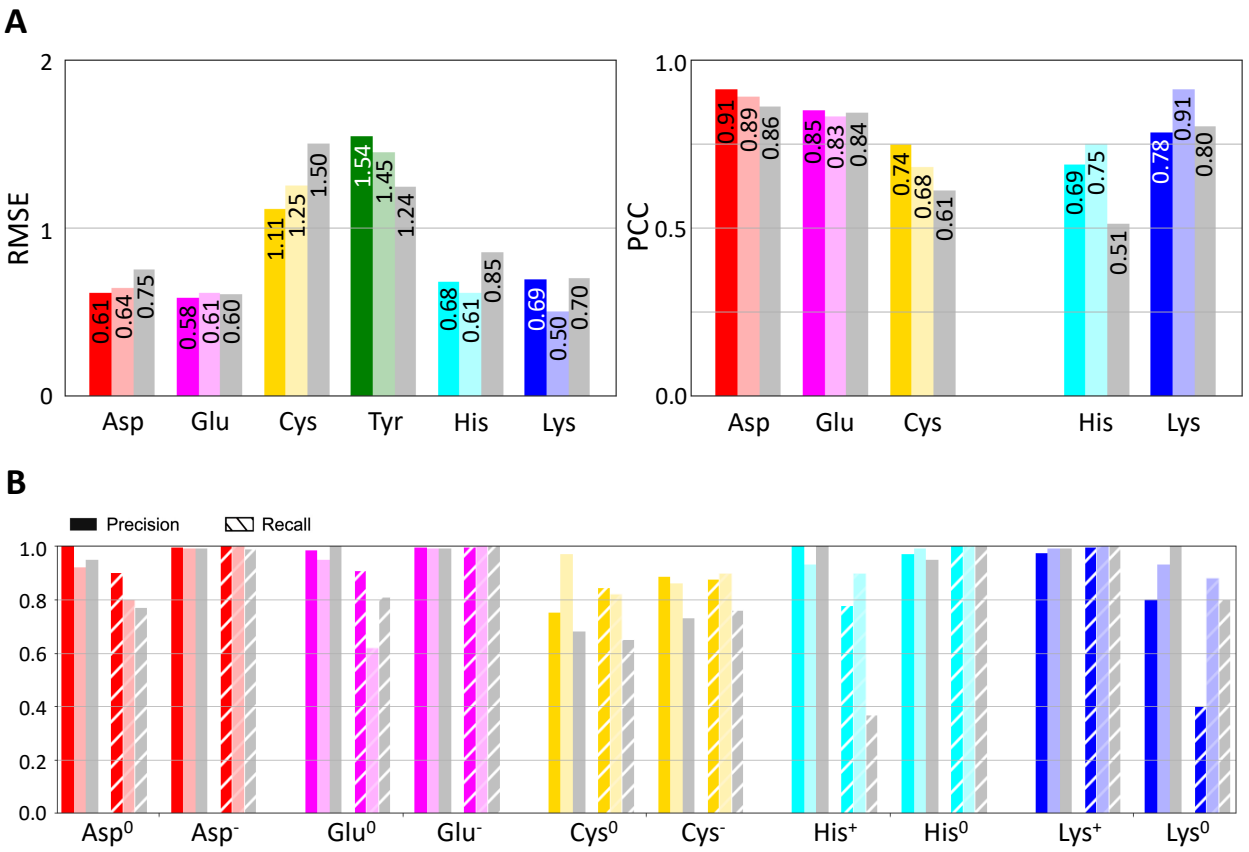

Figure S5: **Evaluation of KaML-ESM2 and KaML-ESMC for predicting pK_a's and protonation state for different titratable residues. A.** Regression metrics for the KaML-ESM2 model (color-coded), KaML-ESMC model (light color-coded), and KaML-CBtree model (grey) in terms of PCC (left) and RMSE (right). The mean value of the 20 splits is given. **B.** The precision (solid) and recall (stripe) for the protonated and deprotonated residues under pH 7 derived from the predictions from the KaML-ESM2 (color-coded), KaML-ESMC model (light color-coded), and KaML-CBtree model (grey)) models. The metrics for Tyr are not shown due to the small test dataset.

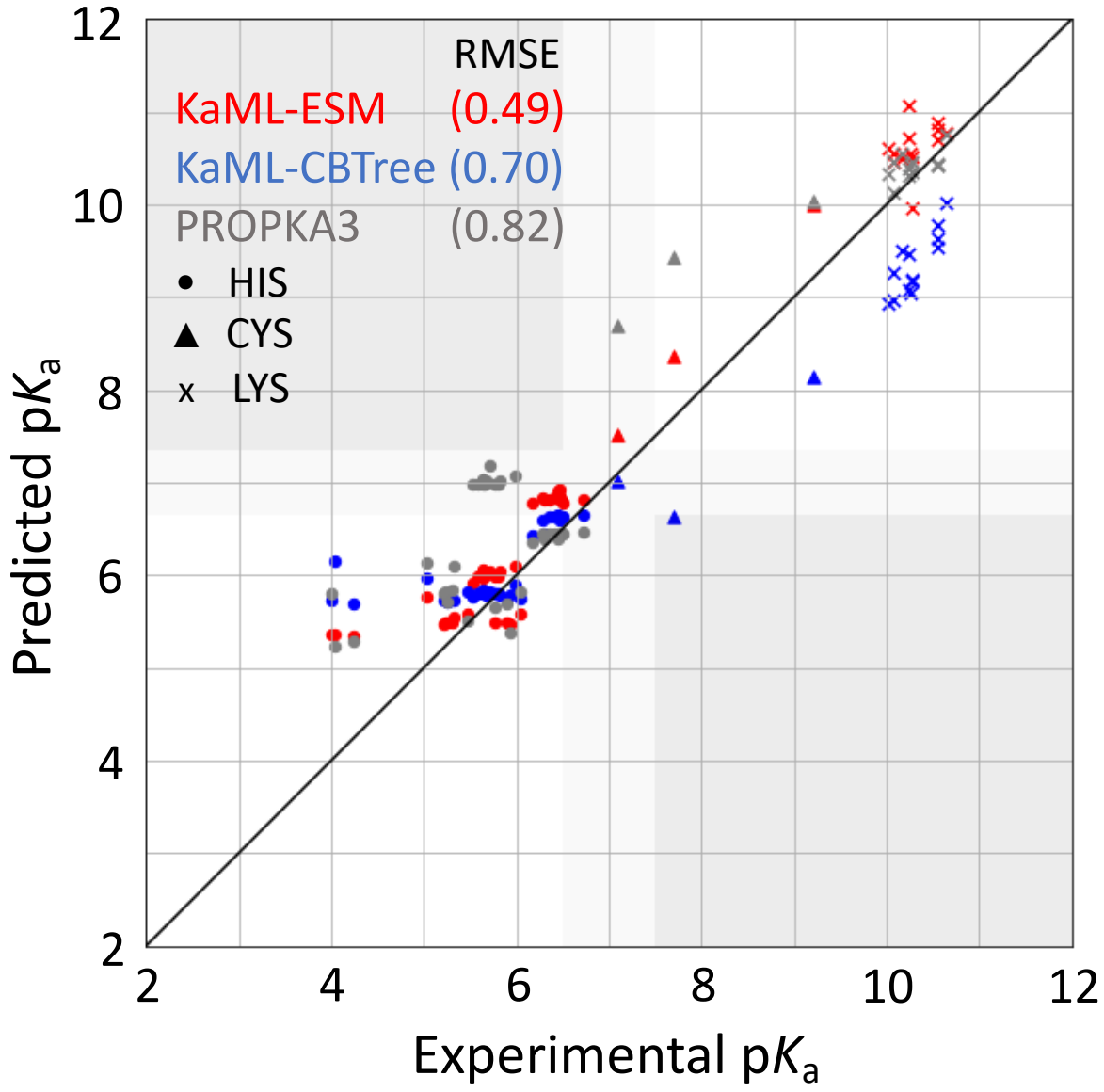

Figure S6: **External evaluation of KaML-ESM against KaML-CBTree, and PROPKA3** Amino acids are represented as shapes: His (•), Cys (▲), and Lys (x). The overall RMSE is given for each model. Following our previous work,[S1] we divided the pK_a range into different regions to illustrate whether a predicted pK_a corresponds to a correct protonation state prediction at pH 7. White regions indicate correct protonation state predictions; gray regions indicate critical errors (i.e., protonated predicted as deprotonated or vice versa); and light gray regions indicate that the predicted pK_a belongs in titration range.

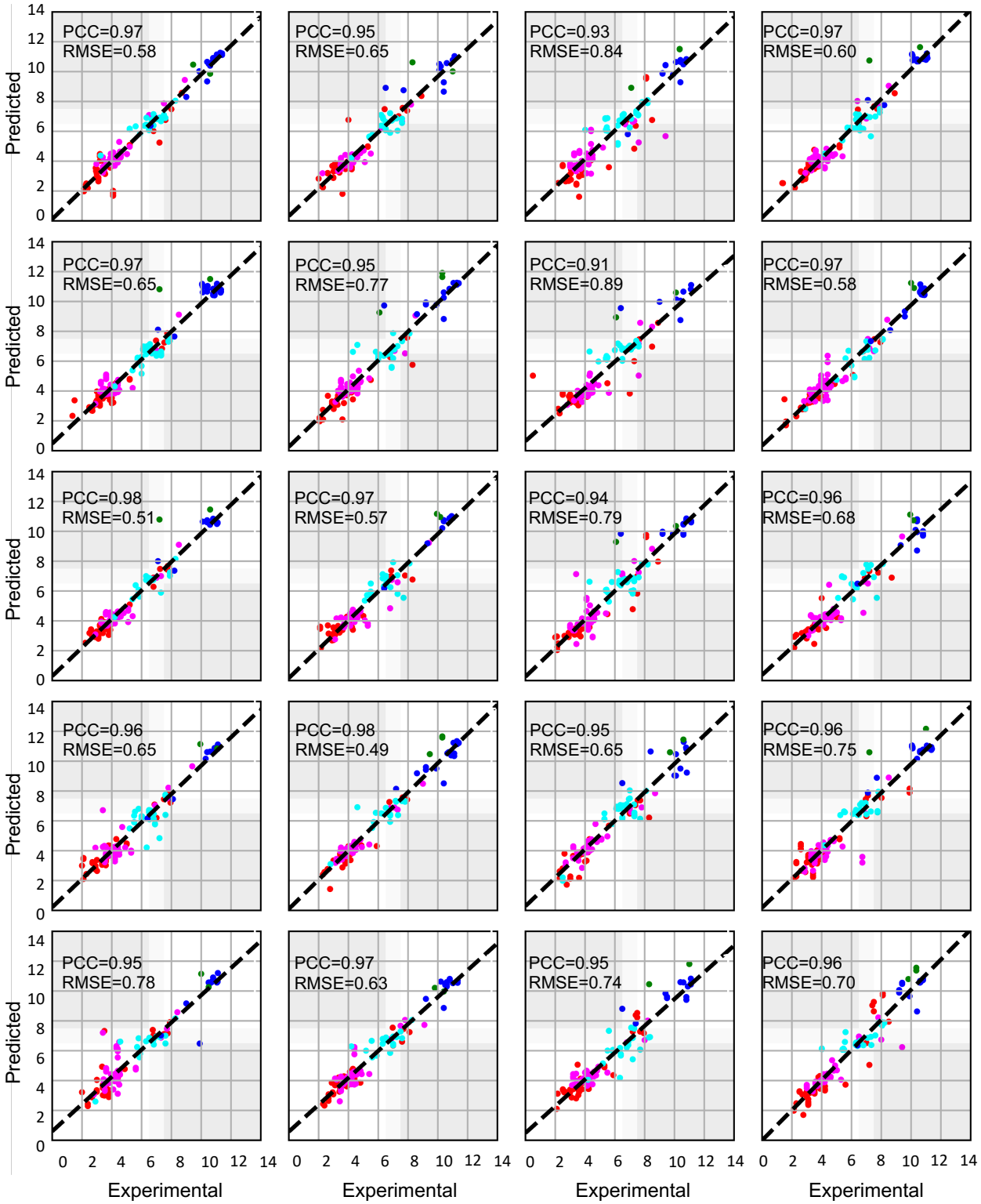

Figure S7: **KaML-ESM2 predicted vs. experimental pK_a's in all 20 hold-out tests.** Predicted vs. experimental pK_a values for all test splits. The dashed line is a linear fit.

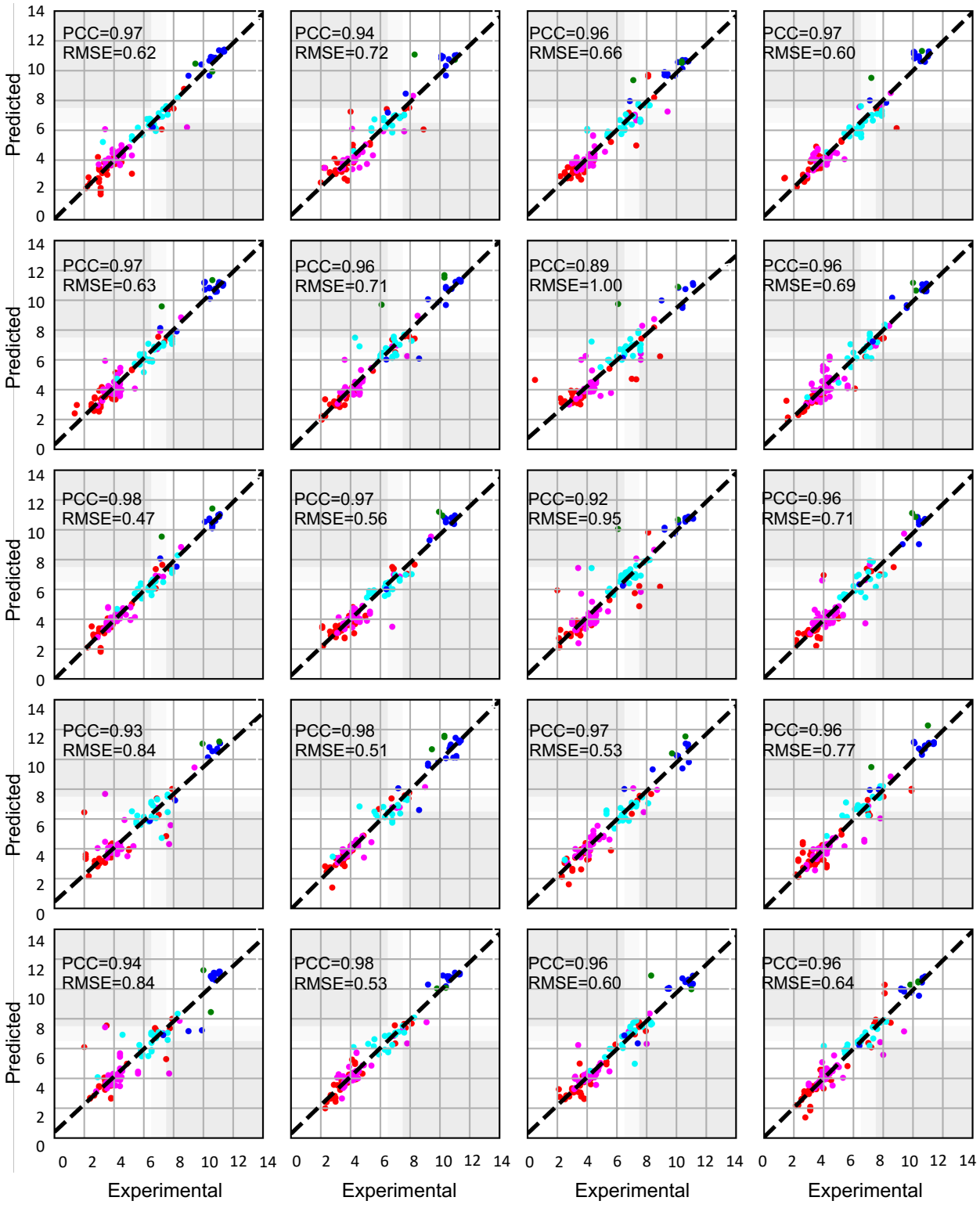

Figure S8: **Comparison of KaML-ESMC predicted and experimental pK_a's in all 20 hold-out tests.** Predicted vs. experimental pK_a values for all test splits. The dashed line is a linear fit.

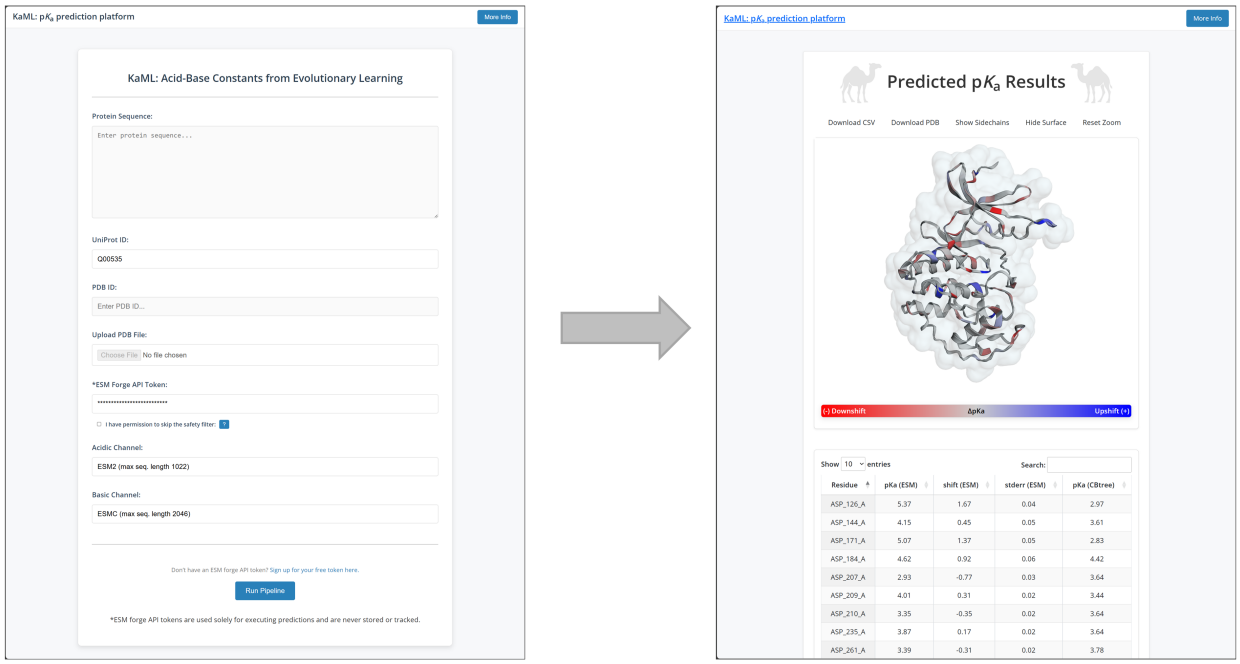

Figure S9: **Screenshot of the browser-based KaML GUI.** An end-to-end web application which takes either the protein sequence, Uniprot ID, PDB ID, or a PDB file as input, along with the end-user's ESM forge API token, to make predictions. Available online for non-commerical use (kaml.computchem.org).

References

(S1) 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.

(S2) Reis, P. B. P. S.; Clevert, D.-A.; Machuqueiro, M. pKPDB: A Protein Data Bank Extension Database of pK_a and pI Theoretical Values. *Bioinformatics* **2021**, *38*, 297–298.

(S3) 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.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput. Biol.* **2017**, *13*, e1005659.

(S4) 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.

(S5) Fey, M.; Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. International Conference on Learning Representations. 2019.

(S6) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019.

(S7) 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.

(S8) 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.

(S9) ESM Team ESM Cambrian: Revealing the Mysteries of Proteins with Unsupervised Learning. https://www.evolutionaryscale.ai/blog/esm-cambrian.

(S10) 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.

(S11) 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.

(S12) Platzer, G.; Okon, M.; McIntosh, L. P. pH-dependent Random Coil 1H, 13C, and 15N Chemical Shifts of the Ionizable Amino Acids: A Guide for Protein pK a Measurements. *J. Biomol. NMR* **2014**, *60*, 109–129.

(S13) Glorot, X.; Bengio, Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. Proceed. Mach. Learn. Res. 2010.

(S14) Baran, K. L.; Chimenti, M. S.; Schlessman, J. L.; Fitch, C. A.; Herbst, K. J.; Garcia-Moreno, B. E. Electrostatic Effects in a Network of Polar and Ionizable Groups in Staphylococcal Nuclease. *J. Mol. Biol.* **2008**, *379*, 1045–1062.

(S15) Ngo, C.; Ekanayake, A.; Zhang, C. Identification of Covalent Ligands – from Single Targets to Whole Proteome. *Israel J. Chem.* **2023**,

(S16) Backus, K. M.; Correia, B. E.; Lum, K. M.; Forli, S.; Horning, B. D.; González-Páez, G. E.; Chatterjee, S.; Lanning, B. R.; Teijaro, J. R.; Olson, A. J.; Wolan, D. W.;

Cravatt, B. F. Proteome-Wide Covalent Ligand Discovery in Native Biological Systems. *Nature* **2016**, *534*, 570–574.

(S17) Vinogradova, E. V.; Zhang, X.; Remillard, D.; Lazar, D. C.; Suciu, R. M.; Wang, Y.; Bianco, G.; Yamashita, Y.; Crowley, V. M.; Schafroth, M. A.; Yokoyama, M.; Konrad, D. B.; Lum, K. M.; Simon, G. M.; Kemper, E. K.; Lazear, M. R.; Yin, S.; Blewett, M. M.; Dix, M. M.; Nguyen, N.; Shokhirev, M. N.; Chin, E. N.; Lairson, L. L.; Melillo, B.; Schreiber, S. L.; Forli, S.; Teijaro, J. R.; Cravatt, B. F. An Activity-Guided Map of Electrophile-Cysteine Interactions in Primary Human T Cells. *Cell* **2020**, *182*, 1009–1026.e29.

(S18) Kuljanin, M.; Mitchell, D. C.; Schweppe, D. K.; Gikandi, A. S.; Nusinow, D. P.; Bulloch, N. J.; Vinogradova, E. V.; Wilson, D. L.; Kool, E. T.; Mancias, J. D.; Cravatt, B. F.; Gygi, S. P. Reimagining High-Throughput Profiling of Reactive Cysteines for Cell-Based Screening of Large Electrophile Libraries. *Nat. Biotechnol.* **2021**, *39*, 630–641.

(S19) Yang, F.; Jia, G.; Guo, J.; Liu, Y.; Wang, C. Quantitative Chemoproteomic Profiling with Data-Independent Acquisition-Based Mass Spectrometry. *J. Am. Chem. Soc.* **2022**, *144*, 901–911.

(S20) Cao, J.; Boatner, L. M.; Desai, H. S.; Burton, N. R.; Armenta, E.; Chan, N. J.; Castellón, J. O.; Backus, K. M. Multiplexed CuAAC Suzuki–Miyaura Labeling for Tandem Activity-Based Chemoproteomic Profiling. *Anal. Chem.* **2021**, *93*, 2610–2618.

(S21) Yan, T.; Desai, H. S.; Boatner, L. M.; Yen, S. L.; Cao, J.; Palafox, M. F.; Jami-Alahmadi, Y.; Backus, K. M. SP3-FAIMS Chemoproteomics for High-Coverage Profiling of the Human Cysteinome. *ChemBioChem* **2021**, *22*, 1841–1851.

(S22) Koo, T.-Y.; Lai, H.; Nomura, D. K.; Chung, C. Y.-S. N-Acryloylindole-alkyne (NAIA)

Enables Imaging and Profiling New Ligandable Cysteines and Oxidized Thiols by Chemoproteomics. *Nat. Commun.* **2023**, *14*, 3564.

(S23) Yan, T.; Boatner, L. M.; Cui, L.; Tontonoz, P. J.; Backus, K. M. Defining the Cell Surface Cysteinome Using Two-Step Enrichment Proteomics. *JACS Au* **2023**, *3*, 3506–3523.

(S24) Njomen, E.; Hayward, R. E.; DeMeester, K. E.; Ogasawara, D.; Dix, M. M.; Nguyen, T.; Ashby, P.; Simon, G. M.; Schreiber, S. L.; Melillo, B.; Cravatt, B. F. Multi-Tiered Chemical Proteomic Maps of Tryptoline Acrylamide–Protein Interactions in Cancer Cells. *Nat. Chem.* **2024**, *16*, 1592–1604.

(S25) Biggs, G. S.; Cawood, E. E.; Vuorinen, A.; McCarthy, W. J.; Wilders, H.; Riziotis, I. G.; Van Der Zouwen, A. J.; Pettinger, J.; Nightingale, L.; Chen, P.; Powell, A. J.; House, D.; Boulton, S. J.; Skehel, J. M.; Rittinger, K.; Bush, J. T. Robust Proteome Profiling of Cysteine-Reactive Fragments Using Label-Free Chemoproteomics. *Nat. Commun.* **2025**, *16*, 73.

(S26) 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.