ACS Synthetic Biology

pubs.acs.org/synthbio
Research Article

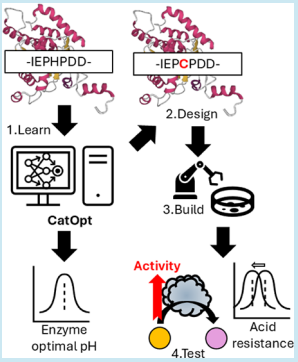
Deep Learning-Based Prediction of Enzyme Optimal pH and Design of Point Mutations to Improve Acid Resistance

Sizhe Qiu,[#] Nan-Kai Wang,[#] Yishun Lu,[#] Jin-Song Gong,[*] Jin-Song Shi, and Aidong Yang[*]

Cite This: *ACS Synth. Biol.* 2025, 14, 4897−4906
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: An accurate deep learning predictor of enzyme optimal pH is essential to quantitatively describe how pH influences the enzyme catalytic activity. CatOpt, developed in this study, outperformed existing predictors of enzyme optimal pH (RMSE = 0.833 and R^2 = 0.479), and could provide good interpretability with informative residue attention weights. The classification of acidophilic and alkaliphilic enzymes and prediction of enzyme optimal pH shifts caused by point mutations showcased the capability of CatOpt as an effective computational tool for identifying enzyme pH preferences. Furthermore, a single point mutation designed with the guidance of CatOpt successfully enhanced the activity of *Pyrococcus horikoshii* diacetylchitobiose deacetylase at low pH (pH = 4.5/5.5) by approximately 7%, suggesting that CatOpt is a promising *in silico* enzyme design tool for pH-dependent enzyme activities.



KEYWORDS: *enzyme optimal pH, deep learning, sequence-based prediction, self-attention, enzyme engineering, acid resistance*

1. INTRODUCTION

In the era of synthetic biology, enzymes play a crucial role in industrial processes, such as food fermentation, waste transformation, and eco-friendly bio-manufacturing of chemical products.[1] In those industrial processes, pH is an important influencing factor of enzyme catalytic activity, as the increase or decrease of pH can affect enzyme protein conformations.[2−4] Each enzyme has an optimal pH (pH_{opt}) where its maximum catalytic rate is attained. Therefore, an accurate enzyme pH_{opt} predictor is highly desirable for enzyme mining and engineering, enabling the discovery of enzymes suited to specific environmental pH and supporting the efforts to enhance catalytic activity within targeted pH ranges.

To fill the knowledge gap of enzyme pH_{opt} in enzyme databases (e.g., BRENDA,[5] uniprot[6]) caused by the high cost of enzyme assays,[7] several machine learning models were developed to make predictions from protein sequences, but most of them could only predict pH_{opt} ranges (acidophilic or alkaliphilic).[8−11] MeTarEnz[12] used random forest regression to quantitatively predict pH_{opt} values, but the accuracy was low (MSE = 1.648 and R^2 = 0.195). Recently, the use of protein language models improved the prediction accuracy of pH_{opt}. EpHod[13] and Seq2pHopt[14] achieved RMSE scores close to 1 using ESM-1[15] and ESM-2,[16] respectively. Subsequently, OphPred[17] surpassed EpHod and Seq2pHopt using ESM-2 and XGBoost,[18] but it lacked interpretability for protein residues. Besides, none of the existing models has been applied to enzyme engineering.

With the aim to build a predictor of enzyme pH_{opt} with good accuracy and interpretability, this study constructed a deep learning model, named CatOpt, with a pre-trained language model of proteins, multi-scale convolutional neural network (CNN), multi-head self-attention, and residual dense neural networks. CatOpt allows the interpretation of residue attention weights, which helps to decipher the key sequence information for enzyme pH_{opt}. Case studies on classifying acidophilic and alkaliphilic enzymes and predicting enzyme pH_{opt} changes by point mutations were carried out to examine CatOpt's performance on *in silico* enzyme selection. The predictor-guided engineering of*Pyrococcus horikoshii*diacetylchitobiose deacetylase to enhance acid resistance, suited to its acidic working environment,[19] demonstrated that CatOpt can function as a useful computational tool for enzyme engineering.

2. METHODS

2.1. Construction of the Deep Learning Model. The training and test datasets of optimal pH (pH_{opt}) were obtained from the Zenodo repository of EpHod.[13] The training set in this study was merged with the original training and validation sets of EpHod. Using the same training and test datasets as EpHod avoided data leakage in model comparison (Section

Received: September 11, 2025
Revised: October 21, 2025
Accepted: November 10, 2025
Published: November 21, 2025

https://doi.org/10.1021/acssynbio.5c00679
ACS Synth. Biol. 2025, 14, 4897−4906

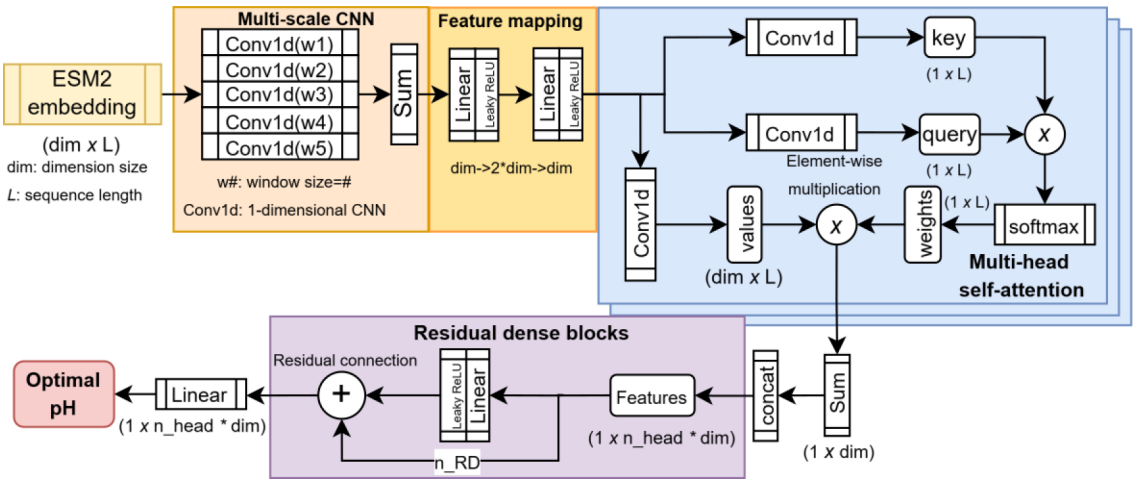


Figure 1. Model architecture of CatOpt. L: protein sequence length; dim: embedding dimension size; Conv1d: 1-D convolutional layer; ⊗: element-wise multiplication; *n*_head: the number of heads in multi-head attention; concat: concatenation; RD: residual dense block, a dense layer with residual connection.

3.1). Sequence identity scores between all protein sequences in the test and training datasets were calculated using MMseqs2[20] for subsequent model evaluation. The model architecture of CatOpt consisted of protein sequence embedding by ESM-2,[16] multi-scale convolutional neural network (CNN), multi-head self-attention network, and residual dense blocks (Figure 1). As ESM-2 had good performance in Seq2pHopt and OphPred, it was chosen to generate protein sequence embeddings in this work.[13,17]

First, the protein sequence embeddings ($r \in R^{dim \times L}$, L: sequence length, dim = 320) were computed using the esm2_t6_8M_UR50D model.[16] The embeddings (r) were first passed to the multi-scale CNN to extract features with sliding window sizes of 1, 2, 3, 4, and 5. The outputs from CNNs with different sliding window sizes were added in an element-wise way. Then, the outputs of the multi-scale CNN were passed to a 2-layer linear featuremap. As alternatives to the multi-scale CNN, CNNs with fixed sliding window sizes were also tested (see SI, Figure S3). In linear feature mapping, the feature dimension was transformed from dim to 2*dim, and then back to dim.

In multi-head self-attention, the outputs of the 2-layer linear feature map were transformed to values ($v^i \in R^{dim \times L}$, i: attention head index), keys ($k^i \in R^{1 \times L}$) and queries ($q^i \in R^{1 \times L}$) via 1-D CNNs. The self-attention weights ($w^i \in R^{1 \times L}$) were computed with element-wise multiplication of keys and queries and a softmax function (eq 1). Next, the element-wise products of values and weights were computed and summed at the dimension of sequence length. The weighted features ($x^i \in R^{dim}$) from all attention heads were concatenated as the inputs ($x_{concat} \in R^{n\_head \times dim}$, *n*_head: number of attention heads) for residual dense blocks. Each residual dense block consisted of a linear layer, Leaky ReLU,[21] and a residual addition operator, ⊕. In the end, a linear layer used the outputs from residual dense blocks to regress for enzyme pH_{opt} values.

$$w^i = \text{softmax}(q^i \times k^i), \quad \forall \text{ attention head } i = 1 - n\_\text{head} \tag{1}$$

2.2. Deep Learning Model Training. For the training process, batch training was used (batch size = 32) for the efficiency and generalizability of the deep learning neural network. Adam optimization algorithm[22] was used to update neural network weights iteratively. The loss function was mean squared error (MSE). The initial learning rate was 0.0005, and the learning rate decayed by 50% for every 10 epochs to prevent overfitting. Before model training started, 10% of the training set was split out as the validation set, and target values were rescaled as $\frac{pH_{opt}}{14}$. During the training process, the prediction accuracy of the model was evaluated with root mean squared error (RMSE), mean average error (MAE), and r-squared (R^2) (see SI, eq S1-3). For details of software and hardware, please see the Section S1.1 of the Supplementary Information. There were 2 hyperparameters in CatOpt, number of attention heads (*n*_head) and number of residual dense blocks (*n*_RD), and hyperparameter optimization was performed on *n*_head = 4, 5, 6 and *n*_RD = 3, 4, 5 (see SI, Figure S2).

2.3. Interpretation of Residue Attention Weights. To investigate how enzyme pH_{opt} was predicted from the amino acid sequence, the average residue attention weights ($w_{avg} \in R^{1*L}$) were computed by averaging the weights across all attention heads (eq 2). Then, the average residue attention weights were mapped to the protein sequence, together with annotated acidic/basic residues, active and binding sites obtained from the uniprot database.[6] The spatial distribution of residue attention weights and annotated protein sequence features could assist in revealing the key sequence information influencing the enzyme pH_{opt}.

$$w_{\text{avg}} = \sum_{i=1}^{n\_\text{head}} w^i / n\_\text{head}, \quad i\text{: attention head index} \tag{2}$$

2.4. Predictor-Guided Design of Single Point Mutations. CatOpt was used to design single point mutations to enhance the acid resistance of*Pyrococcus horikoshii* diacetylchitobiose deacetylase (PhDac), an enzyme catalyzing the production of glucosamine (GlcN) from N-acetylglucosamine (GlcNAc).[19] Because the fermentation environment of diacetylchitobiose deacetylase is usually acidic, it is desirable to enhance its enzyme activity under low pH.[23] For 3 × 21 sites centered at three substrate binding sites (D46, R92, and H152, −10 site to +10 site), all possible amino acid substitutions were considered, and thus there were totally 1197 mutated sequences. For all mutants, pH_{opt} and turnover

https://doi.org/10.1021/acssynbio.5c00679
ACS Synth. Biol. 2025, 14, 4897−4906

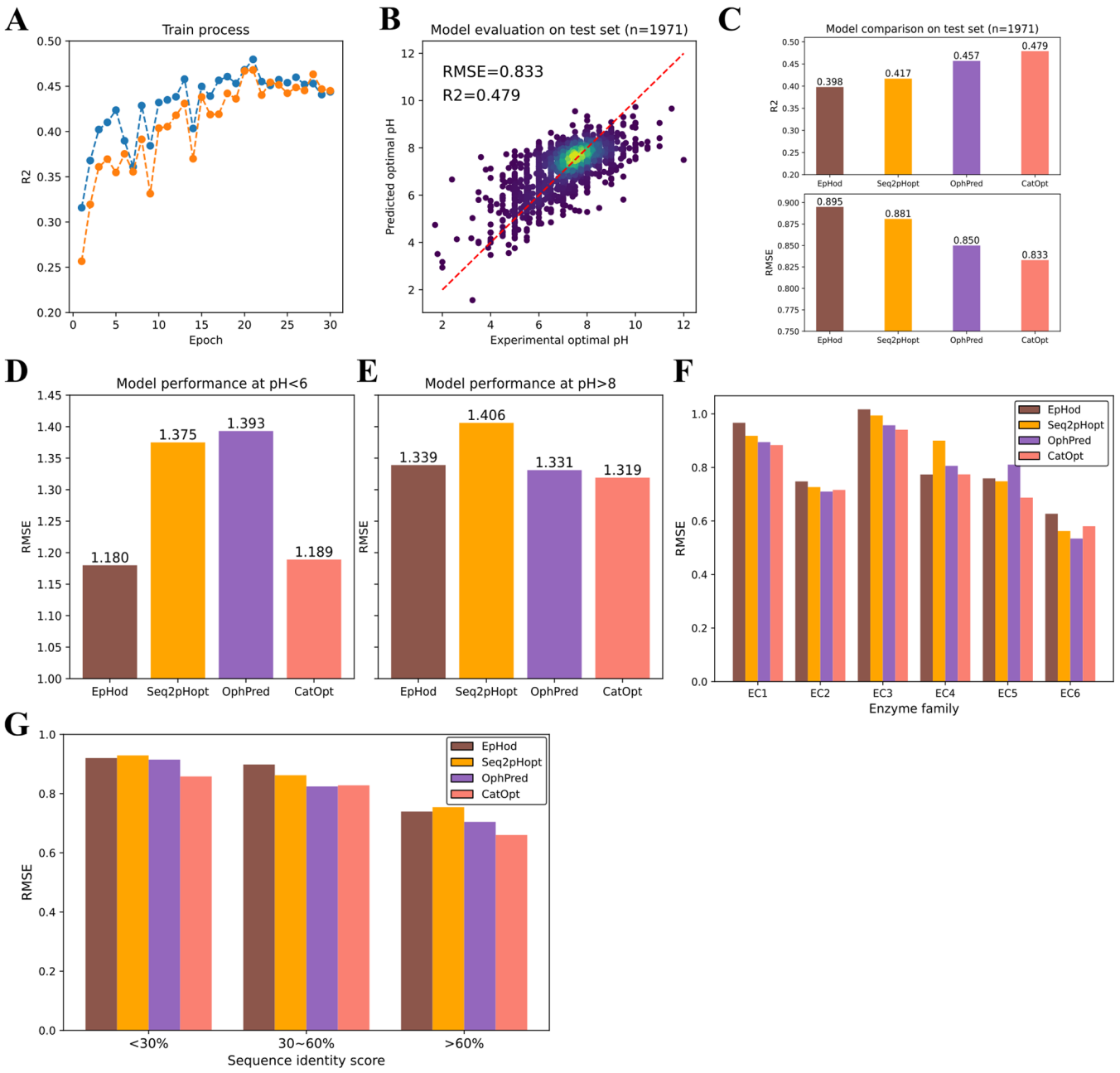


Figure 2. Model performance evaluation of CatOpt. (A) The R^2 scores of pH_{opt} prediction during the training process. Blue dotted curve: test set; Orange dotted curve: dev set (validation set). (B) Experimental and predicted pH_{opt} by CatOpt (RMSE = 0.833 and R^2 = 0.479). (C) Prediction accuracy comparison of EpHod (R^2 = 0.399, RMSE = 0.895), Seq2pHopt (R^2 = 0.417, RMSE = 0.881), OphPred (R^2 = 0.457, RMSE = 0.85), and CatOpt (R^2 = 0.479, RMSE = 0.833) on the same test set. (D) Prediction accuracy comparison of 4 models at low (pH_{opt} < 6) value range. (E) Prediction accuracy comparison of 4 models at high (pH_{opt} > 8) value range. (F) Prediction accuracy comparison of EpHod, Seq2pHopt, OphPred, and CatOpt for different enzyme classes (EC 1−6). (G) Prediction accuracy comparison of EpHod, Seq2pHopt, OphPred, and CatOpt for different sequence identity score ranges (<30%, 30−60%, >60%).

numbers (k_{cat}) were predicted by CatOpt and DLTKcat,[24] respectively. DLTKcat was used to predict turnover numbers of mutants and filter out mutants with low catalytic efficiency. Designed point mutations were selected with an arbitrary threshold of predicted pH_{opt} < 7. In this study, site directed mutagenesis, protein expression and purification of PhDac mutants followed the same procedure as in Huang et al., 2021.[19]

2.5. Measurement of Enzyme Activity. The enzyme activity measurement method was adapted from Jiang et al., 2019[25] with several modifications. In summary, the reaction solution comprised 1 mL of citrate buffer (50 mM, pH 4.5 to 5.5), 50 g/L of GlcNAc, and 100 μL crude enzyme (i.e., designed mutants). The reaction was performed in a metal bath at 40 °C and 900 rpm for 20 min. Subsequently, the reaction was halted by the addition of 50 μL of 0.5 M HCl. The mixture was then centrifuged at 12,000 rpm for 5 min. After centrifugation, 10 μL of the supernatant was combined with 100 μL of the OPA detection reagent (composed of 5 mg OPA, 10 μL of 1.0 M dithiothreitol, and 100 μL of alcohol in 10 mL of sodium carbonate buffer), and the absorbance was measured at 330 nm using the Infinite M200 PRO Spectrum

https://doi.org/10.1021/acssynbio.5c00679
ACS Synth. Biol. 2025, 14, 4897−4906

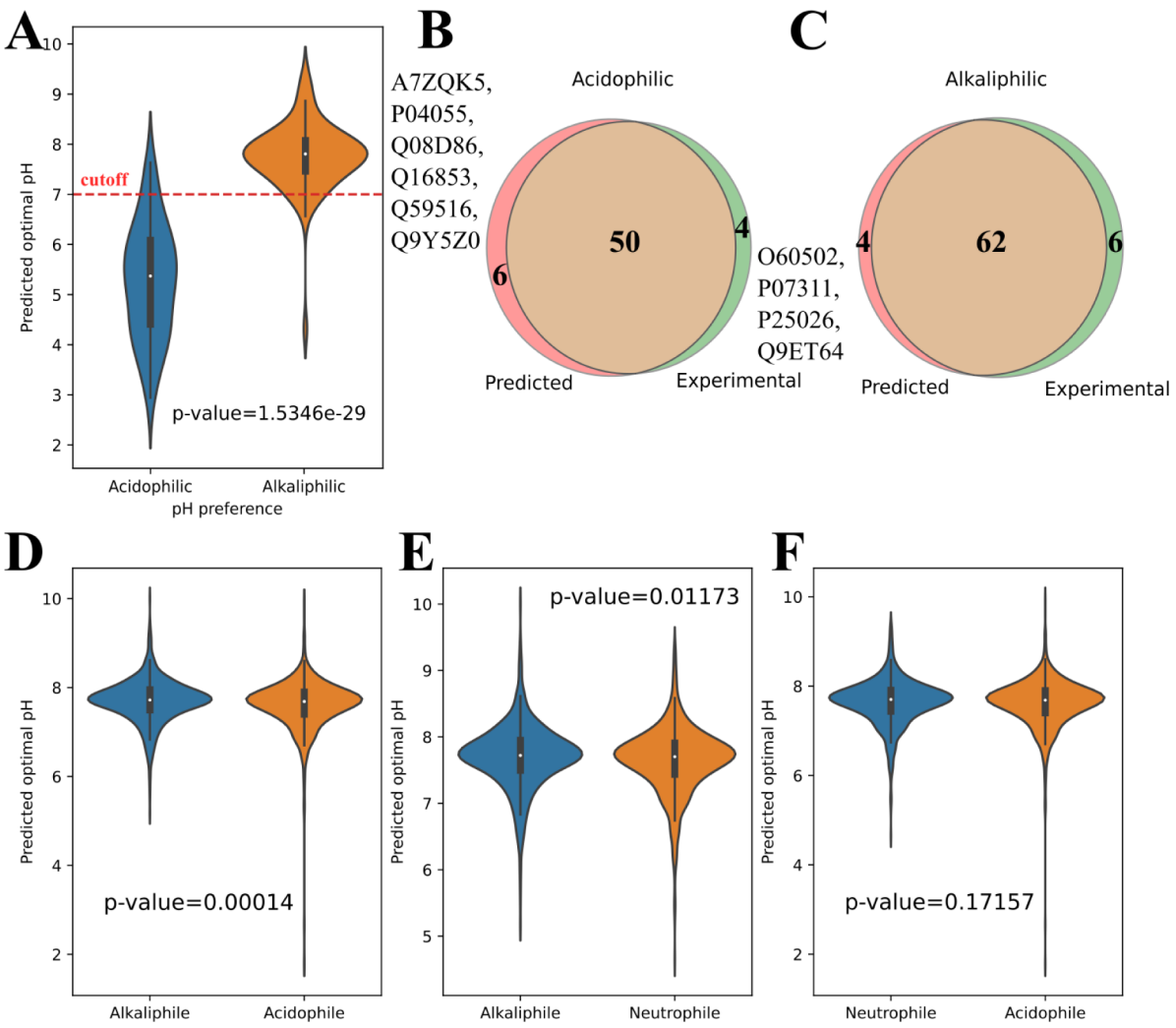


Figure 3. Performance of CatOpt on the pH preference of enzymes and microorganisms. (A) The distribution of predicted pH_{opt} values of 54 acidophilic enzymes and 68 alkaliphilic enzymes (p-value < 0.001). (B) The Venn diagram of predicted and experimental acidophilic enzymes. (C) The Venn diagram of predicted and experimental alkaliphilic enzymes. (D) The distribution of predicted pH_{opt} values of enzymes from alkaliphilic and acidophilic microorganisms (p-value < 0.001). (E) The distribution of predicted pH_{opt} values of enzymes from alkaliphilic and neutrophilic microorganisms (p-value < 0.05). (F) The distribution of predicted pH_{opt} values of enzymes from neutrophilic and acidophilic microorganisms (p-value > 0.05).

spectrophotometer (Tecan Trading AG; Switzerland). One unit of enzyme activity was defined as the amount of the enzyme liberating 1 μM GlcN in 1 h at 40 °C.

3. RESULTS

3.1. CatOpt Outperformed Existing Predictive Models. First, the hyperparameter optimization on the number of attention heads (n_head) and residual dense blocks (n_RD) found that n_head = 4 and n_RD = 4 is the best set of hyperparameters in the search scope (see SI, Figure S2). Also, the performance comparison of CNNs with fixed sliding window sizes (3, 4, 5) and the multi-scale CNN of 5 different window sizes justified the use of multi-scale CNN in CatOpt (see SI, Figure S3). Then, CatOpt, with the optimal set of hyperparameters, achieved a prediction accuracy of R^2 = 0.479, MAE = 0.607 and RMSE = 0.833 on the hold-out test set (Figure 2AB and see SI, Figure S4). In model comparison on the same test set provided by the Zenodo repository of EpHod,[13] CatOpt outperformed Seq2pHopt, EpHod, and OphPred (Figure 2C). With respect to prediction errors at pH_{opt} < 6, CatOpt had a RMSE of 1.189, close to EpHod (RMSE = 1.180) and lower than OphPred (RMSE = 1.393) and Seq2pHopt (RMSE = 1.375) (Figure 2D). At pH_{opt} > 8, CatOpt had a RMSE of 1.319, slightly lower than EpHod (RMSE = 1.339) and OphPred (RMSE = 1.331) (Figure 2E). For 6 enzyme classes (EC1-6), CatOpt had lower RMSEs than the other 3 models in oxidoreductases (EC1), hydrolases (EC3), lyases (EC4), and isomerases (EC5), and close RMSEs to EpHod and OphPred in transferases (EC2) and ligases (EC6) (Figure 2F). There were only 22 translocases (EC7) in the test set (see SI, Figure S1), much fewer than the other enzyme classes, therefore EC7 was not included in model performance evaluation. For enzymes with varying identity scores to those in the training dataset, CatOpt exhibited lower RMSEs than the other 3 models in <30% and >60% identity score ranges, and a comparable RMSE to OphPred in the 30−60% range (Figure 2G). In general, CatOpt exhibited good accuracy and outperformed existing predictive models of enzyme pH_{opt}.

https://doi.org/10.1021/acssynbio.5c00679
ACS Synth. Biol. 2025, 14, 4897−4906

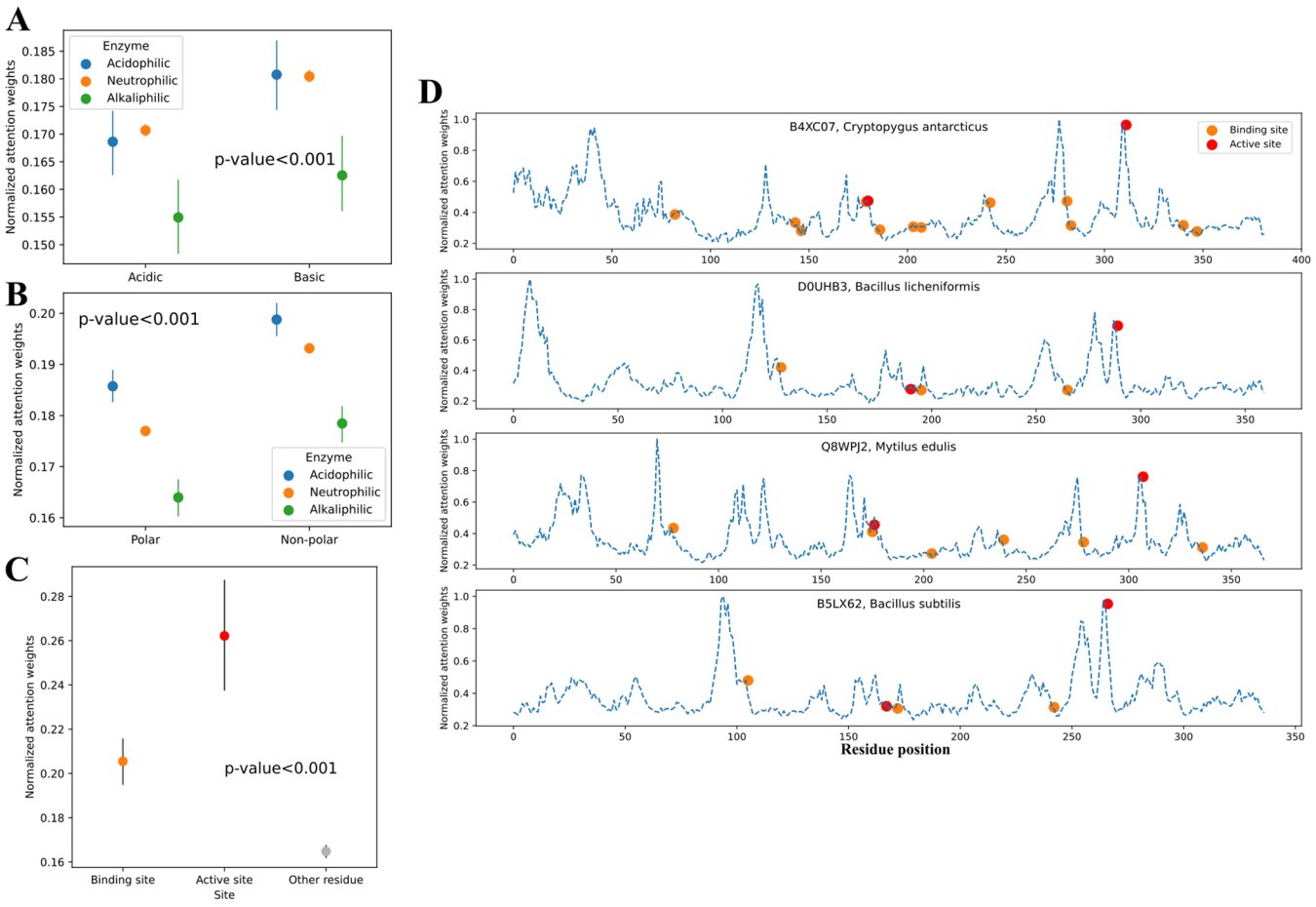


Figure 4. Analysis of residue attention weights. (A) The attention weights of acidic and basic residues of acidophilic, neutrophilic, and alkaliphilic enzymes. (B) The attention weights of polar and non-polar residues of acidophilic, neutrophilic, and alkaliphilic enzymes. (C) The attention weights on binding sites, active sites and other residues. (D) Representative examples of residue attention weights, positions of binding and active sites of 4 mannan endo-1,4-beta-mannosidases (EC:3.2.1.78). The uniprot IDs are B4XC07, D0UHB3, Q8WPJ2, and B5LX62. Blue dashed curve: normalized attention weights; orange dot: binding site; red dot: active site.

3.2. Identifying the pH Preferences of Enzymes and Microorganisms. After the prediction accuracy of CatOpt had been validated (Section 3.1), this study proceeded to examine its ability to identify the pH preferences of enzymes and microorganisms. The benchmark dataset of AcalPred (54 acidophilic enzymes and 68 alkaliphilic enzymes)[8] was used in the identification of acidophilic and alkaliphilic enzymes. Notably, this dataset had no overlap with the training dataset of CatOpt. The predicted pH_{opt} values of alkaliphilic enzymes were significantly higher than those of acidophilic enzymes (Figure 3A). With the cutoff of pH_{opt} = 7.0, CatOpt classified acidophilic and alkaliphilic enzymes with an accuracy of 91.8%, 50 acidophilic enzymes and 62 alkaliphilic enzymes were accurately identified (Figure 3BC). In 10 misclassified enzymes, CatOpt mislabelled O60502, P07311, P25026, Q9ET64 as alkaliphilic enzymes, and A7ZQK5, P04055, Q08D86, Q16853, Q59516, Q9Y5Z0 as acidophilic enzymes. EpHod performed slightly worse than CatOpt, misclassifying 13 enzymes (see SI, Figure S9). The host organisms of misclassified enzymes were mainly neutrophilic, such as*Homo sapiens*and*Pseudomonas pyrrocinia*. This is likely because enzymes from neutrophiles typically have near-neutral pH_{opt} values, making it difficult to classify them as either acidophilic or alkaliphilic. Overall, this case study demonstrated that CatOpt could discriminate enzymes with different pH preferences.

Next, pH_{opt} values were predicted for catalytic enzymes belonging to 3 acidophilic, 3 neutrophilic, and 3 alkaliphilic microorganisms (see SI, Figure S5). Enzyme protein sequences were all obtained from the uniprot database.[6] CatOpt could discriminate alkaliphilic and acidophilic microorganisms with significantly different distributions of predicted enzyme pH_{opt} values, the same for alkaliphilic and neutrophilic microorganisms (Figure 3D,E). However, there was no significant difference between predicted pH_{opt} values of enzymes from acidophilic and neutrophilic microorganisms (Figure 3F). Despite that CatOpt could identify the pH preferences of enzymes, it could not accurately classify acidophilic, neutrophilic, and alkaliphilic microorganisms based on distributions of predicted enzyme pH_{opt} values.

3.3. Residue Attention Weights Capture Key Sequence Information. To investigate how residue attention weights capture important sequence information, this study compared attention weights on different types of residues across acidophilic, neutrophilic, and alkaliphilic enzymes in the hold-out test set. For both acidic and basic residues, the attention weights of acidophilic and neutrophilic enzymes were significantly higher than the weights of acidophilic enzymes (Figure 4A), which revealed the importance of ionizable

https://doi.org/10.1021/acssynbio.5c00679
ACS Synth. Biol. 2025, 14, 4897−4906

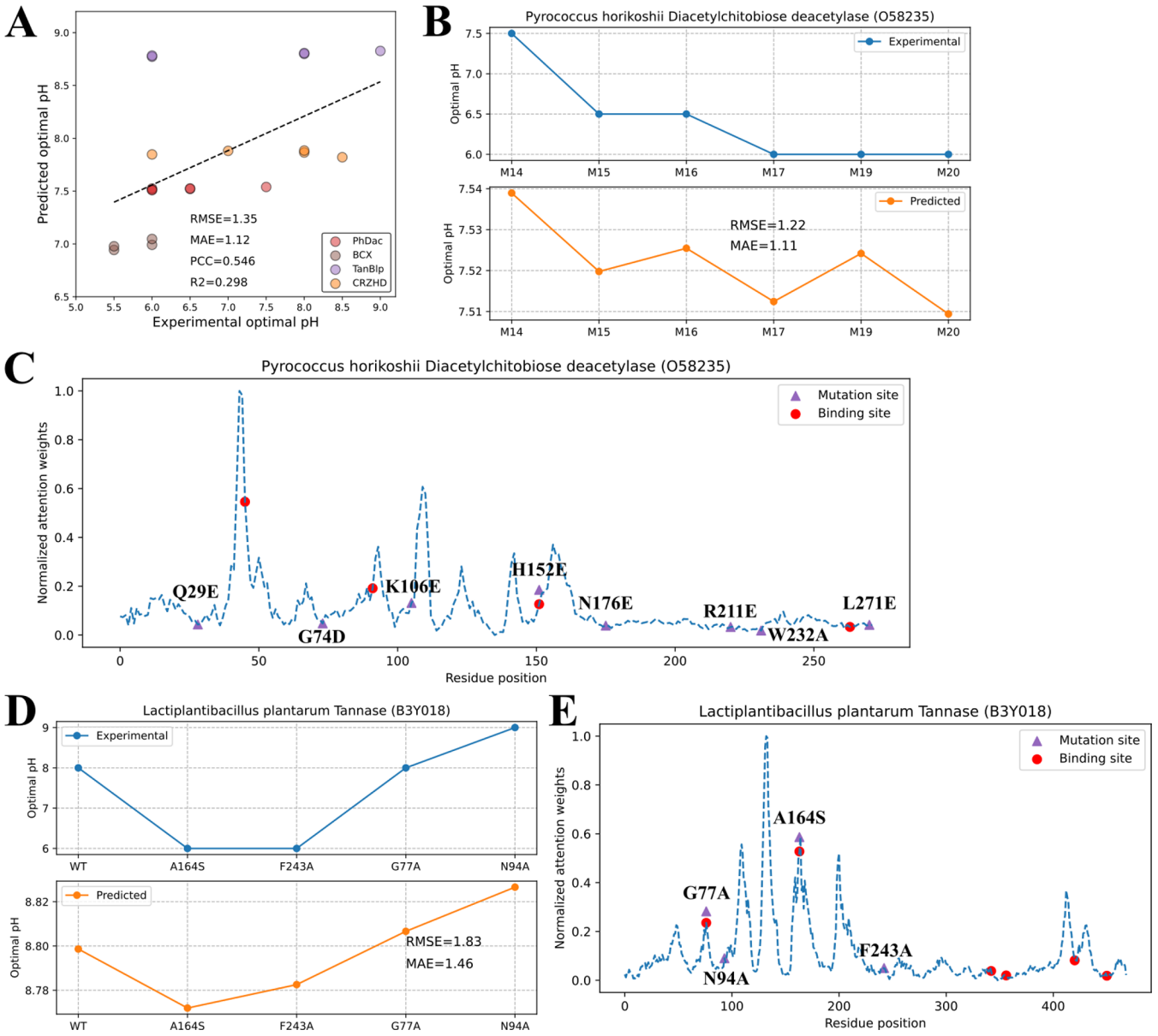


Figure 5. Prediction of enzyme pH_{opt} shifts caused by mutations. (A) Experimental and predicted enzyme pH_{opt} of WTs and mutants of 4 different enzymes (RMSE = 1.35, MAE = 1.12, PCC = 0.546, R^2 = 0.298). PCC: Pearson correlation coefficient, PhDac:*Pyrococcus horikoshii*diacetylchitobiose deacetylase, BCX:*Bacillus circulans* xylanase, TanBlp:*Lactiplantibacillus plantarum*tannase, CRZHD:*Clonostachys rosea*zearalenone hydrolase, WT: wild-type. (B) Experimental and predicted enzyme pH_{opt} of 6 mutants of PhDac. (C) Residue attention weights of PhDac and positions of substrate binding sites and point mutations in 6 mutants. (D) Experimental and predicted enzyme pH_{opt} of WT and mutants of TanBlp. (E) Residue attention weights of TanBlp and positions of substrate binding sites and point mutations in 4 mutants.

residues on enzyme pH_{opt}. In all three types of enzymes, the attention weights on nonpolar residues were significantly higher than the weights on polar residues (Figure 4B). Previous studies have shown that ionizable and polar residues are associated with pH-dependent transient states of enzymes, where catalysis occurs.[26−28] Therefore, the significantly different distributions of attention weights on ionizable, polar, and other residues indicated that CatOpt effectively identified key residues related to pH-dependent enzyme catalytic activity.

Next, 173 enzymes with annotated active and binding sites were selected from the test set. The active sites are regions where chemical reactions happen, and the binding sites are residues where substrates bind. The attention weights on active and binding sites were significantly higher than attention weights on other residues (Figure 4C), suggesting that residue attention weights could capture important residues for enzyme catalysis. For example, the active sites of 4 mannan endo-1,4-beta-mannosidases were mostly close to peaks of residue attention weights, although the weights were not indicative for binding sites (Figure 4D). In presented 4 enzymes, there were high attention weight peaks that did not correspond to annotated active and binding sites. Unfortunately, these high weight peaks could not be interpreted with existing sequence annotations. In a nutshell, residue attention weights in CatOpt provided good interpretability by capturing key sequence information, i.e., ionizable residues, active and binding sites.

3.4. Prediction of Optimal pH Shifts Caused by Point Mutations. To examine the inference ability of CatOpt on how point mutations affect enzyme pH_{opt}, this study used it to

https://doi.org/10.1021/acssynbio.5c00679
ACS Synth. Biol. 2025, 14, 4897−4906

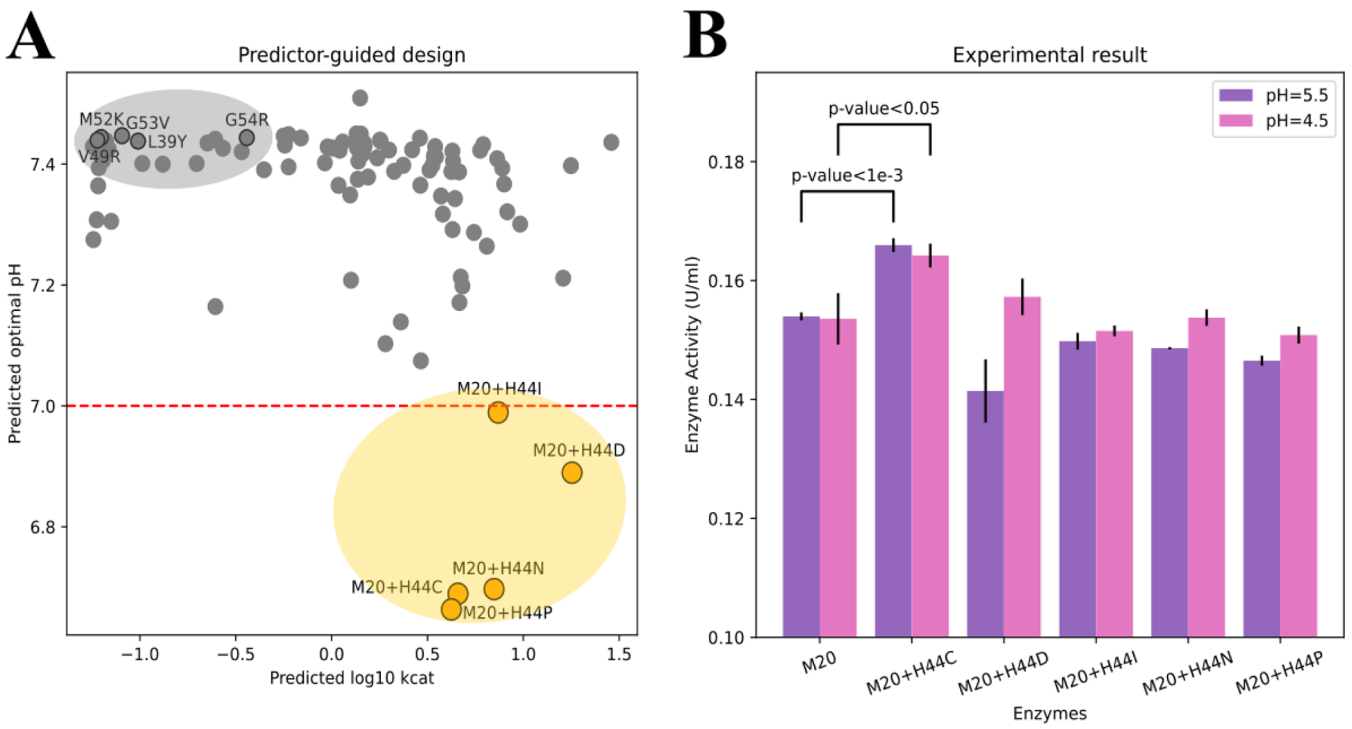


Figure 6. Predictor-guided engineering of*Pyrococcus horikoshii*diacetylchitobiose deacetylase (PhDac). (A) Predicted enzyme pH_{opt} and turnover number values (log_{10} k_{cat}) of mutants with single point mutations based on M20 (only lower 10% of predicted pH_{opt} values were presented). Orange dots represent 5 selected mutants: M20 + H44I, M20 + H44D, M20 + H44N, M20 + H44C, M20 + H44P. Grey dots represent the remaining un-selected mutants, grey dots with black edges represent M20 + G53 V, M20 + M52K, M20 + G54R, M20 + V49R, M20 + L39Y. (B) Enzyme activities (U/mL) at pH = 4.5 and 5.5 for M20 and 5 selected mutants. All measurements have 3 replicates (see SI Table S2, S3). The error bars represent standard deviations.

predict pH_{opt} values of wild-types (WTs) and mutants for *Pyrococcus horikoshii*diacetylchitobiose deacetylase (PhDac),[19] *Bacillus circulans*xylanase (BCX),[29]*Lactiplantibacillus plantarum*tannase (TanBlp),[30] and*Clonostachys rosea* zearalenone hydrolase (CRZHD)[31] (see SI, Table S1). The experimental data of those enzymes was not included in the training set of CatOpt. The overall prediction error for WTs and mutants of those 4 different enzymes was RMSE = 1.35, MAE = 1.12, and R^2 = 0.298 (Figure 5A).

For PhDac, the prediction by CatOpt qualitatively accounted for the downshift of pH_{opt} in M15 and M16 in comparison to M14, and in M17 and M20 in comparison to M15 and M16, but the numerical difference of predicted pH_{opt} among 6 mutants was small (Figure 5B). The residue attention weights of PhDac captured substrate binding sites with high peaks (e.g., D46) (Figure 5C), although the spatial distribution of weights could not explain effective point mutations causing down-shift of pH_{opt} (e.g., Q29E). The highest peak of attention weights was at residue 44, suggesting a potential effective site for point mutations. For TanBlp, CatOpt quantitatively predicted the downshift of pH_{opt} caused by A164S and F243A in comparison to WT, and the upshift of pH_{opt} by G77A and N94A in comparison to A164S and F243A (Figure 5D). Two substrate binding sites of TanBlp (G77 and A164) and two effective point mutations (G77A and A164S) were identified by high peaks of residue attention weights (Figure 5E). BCX and CRZHD were not included in further analysis (see SI, Figure S7), due to lack of protein sequence annotation. Generally speaking, this case study demonstrated CatOpt's capability to predict the effect of point mutations on enzyme pH_{opt}, although the numerical differences of predicted enzyme pH_{opt} for WTs and mutants were smaller than those observed in experimental measurements.

3.5. Predictor-Guided Engineering of Diacetylchitobiose Deacetylase to Enhance Acid Resistance. CatOpt was used as a computational design tool to enhance the acid resistance of PhDac (Figure 5BC), which is used in the environmentally-friendly manufacturing of GlcN.[19] Enzyme pH_{opt} and turnover number values of 1197 mutants with single point mutations based on M20 (see SI, Table S1) were predicted (Figure 6A). 5 mutants with lowest predicted pH_{opt} values, which also had relatively high turnover number values, were selected to examine their activities at pH = 4.5 and 5.5. Compared to M20, H44C improved the activities of PhDac at pH = 4.5 and 5.5 by around 7% (p-value < 0.05), H44D improved the activity at pH = 4.5 by around 2% (non-significant) (Figure 6B). The other 3 selected mutants (M20 + H44I, M20 + H44N, M20 + H44P) did not enhance the acid resistance, but their activities at pH = 4.5 were all higher than activities at pH = 5.5 (Figure 6B), which suggested the down-shift of pH_{opt}. Both effective point mutations (H44C and H44D) substituted a basic residue (histidine (H)) with an acidic residue (cysteine (C) and aspartate (D)), possibly stabilizing the transition state at the substrate binding site, D46, under acidic pH conditions.[32] However, the stabilizing effect of acidic residues close to PhDac's substrate binding sites remained a hypothesis, pending future experimental validation. In addition, enzyme activities of 5 un-selected mutants (Figure 6A) were also measured, and they showed weaker acid resistance than M20 (Tables S2, S3, Figure S8). In short, the success of computationally designed point mutations showed

https://doi.org/10.1021/acssynbio.5c00679
ACS Synth. Biol. 2025, 14, 4897−4906

the usefulness of CatOpt as a design tool of enzyme engineering.

4. DISCUSSION

To address the limitations of existing predictive models of pH_{opt} (e.g., low accuracy or lack of interpretability), this study developed CatOpt, a deep learning model capable of predicting enzyme pH_{opt} directly from protein sequences. Main components of CatOpt were ESM-2 embedding generation, multi-scale CNN, multi-head self-attention, and residual dense neural networks (Figure 1). Compared with one-hot encoding[33] or k-mer based dictionary embedding,[24,34] ESM-2, as a pre-trained large language model of proteins, can transfer the knowledge of protein structures and functions from a large dataset of millions of protein sequences to this prediction task using thousands of protein sequences.[16] The advantage of multi-scale CNN with different window sizes over CNN with a fixed window size lies in ensembling information to enrich the representation of protein sequence features.[35] In contrast to multi-head light attention in Seq2Topt/Seq2-pHopt,[14] self-attention in CatOpt can model dependencies between different regions of the protein sequence.[36] Additionally, the use of residue dense neural networks instead of multiple linear layers could effectively reduce the vanishing and exploding gradient issues in deep neural networks.[37] Consequently, CatOpt outperformed existing enzyme pH_{opt} predictors (e.g., OphPred) with RMSE = 0.833 and R^2 = 0.479, and provided good interpretability with informative residue attention weights (Section 3.3).

The classification of acidophilic/alkaliphilic enzymes (Section 3.2), prediction of enzyme pH_{opt} shifts by point mutations (Section 3.4), and predictor-guided engineering of PhDac (Section 3.5) demonstrated that CatOpt could be applied to enzyme mining and computational design of enzymes via fast screening the effect of point mutations. Besides *in silico* screening, the combination of generative deep learning and CatOpt might lead to automatic generation of novel acidophilic or alkaliphilic enzymes through predictor-guided generator optimization.[38] Moreover, like EpHod[13] and Seq2Topt,[14] CatOpt can also be trained to predict multiple protein properties with a single sequence-based model. Besides, the informative attention weighted protein features extracted by CatOpt (Section 3.3) could be used in other prediction tasks of enzyme catalytic activity, such as the prediction of pH-dependent enzyme turnover numbers,[39,40] via transfer learning. Therefore, beyond predicting a specific protein property, CatOpt holds the potential to serve as a protein foundation model[41] that can be adapted for diverse downstream tasks in protein science.

Despite the achievement of CatOpt outlined above, some limitations still exist and hinder its performance. Similar to Seq2Topt/Seq2pHopt,[14] the accuracy of CatOpt was also affected by the imbalance of the training dataset. Oversampling and loss reweighting can mitigate the imbalance, but the prediction accuracy at low and high value ranges will still be relatively low.[13,14] Using high-throughput enzyme assays to append entries at the ranges of $pH_{opt} < 6$ and pHopt > 8 to the dataset is necessary to further improve the performance of pH_{opt} prediction. The neglect of environmental factors influencing enzyme pH_{opt} is another shortcoming, such as temperature[42] and ionic strength.[43] These environmental factors can affect the protein conformation and thereby complicate the relationship between pH and enzyme catalytic activity. The inclusion of enzyme assay metadata could help account for environmental factors and improve the prediction accuracy of enzyme pH_{opt}, although a large portion of enzyme assay results curated from databases currently lack such information. In the attempt to identify organismal pH preferences (Section 3.2), CatOpt failed to differentiate neutrophilic and acidophilic microorganisms with distributions of predicted enzyme pH_{opt} values. The distribution of experimental enzyme pH_{opt} values indicates that most enzymes in acidophilic microorganisms have pH_{opt} in the range of 6−8, just like neutrophilic microorganisms (see SI, Figure S6). Therefore, the relationship between microbial growth pH_{opt} and enzyme pH_{opt} still remains to be investigated.

In conclusion, CatOpt is an interpretable deep learning predictor of enzyme pH_{opt} that demonstrated improved accuracy compared to existing tools, despite the limitations discussed above. As envisaged, CatOpt can potentially accelerate enzyme discovery for desired properties from "biological dark matter" and enzyme engineering with *in silico* design.

■ ASSOCIATED CONTENT

Data Availability Statement

The code and data are openly available at https://github.com/SizheQiu/CatOpt and Supporting Information.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.5c00679.

Supplementary Methods: S1.1 Software and code availability. Supplementary Methods: S1.2 Evaluation metrics. Supplementary Figures: Figure S1. Distributions of enzyme optimal pH values in the training set (7884 entries) and test set (1971 entries), and distributions of enzymes from different EC families in both sets. Supplementary Figures: Figure S2. Hyperparameter optimization of enzyme optimal pH prediction for number of attention heads (n_head) and residual dense blocks (n_RD). n_RD = 4, n_head = 4 was chosen as the best set of hyperparameters. Supplementary Figures: Figure S3. Performance comparison of CNNs with fixed sliding window sizes (3, 4, 5) and multi-scale CNN used in CatOpt. n_RD = 4, n_head = 4. Supplementary Figures: Figure S4. The RMSE, MAE, R^2 scores of CatOpt during the training process. Supplementary Figures: Figure S5. Predicted optimal pH values for enzymes in 9 microorganisms. Acidophiles: *Acidithiobacillus ferrooxidans*, *Lactobacillus acidophilus*, and *Thermoplasma acidophilum*; Neutrophiles: *Bacillus subtilis*, *Salmonella enterica*, and *Staphylococcus aureus*; Alkaliphiles: *Clostridium paradoxum*, *Alkalihalobacillus alcalophilus*, and *Natronomonas pharaonis*. Supplementary Figures: Figure S6. Experimental optimal pH values for enzymes in 3 acidophilic microorganisms and 2 neutrophilic microorganisms. Supplementary Figures: Figure S7. (A) Experimental and predicted enzyme pH_{opt} of wild-types and mutants of BCX. (B) Experimental and predicted enzyme pH_{opt} of wild-types and mutants of CRZHD. Supplementary Figures: Figure S8. Enzyme activities (U/mL) at pH = 4.5 and 5.5 for M20 and M20 + G53V, M20 + M52K, M20 + G54R, M20 + V49R, M20 + L39Y. Supplementary Figures: Figure S9. The performance of EpHod on the

https://doi.org/10.1021/acssynbio.5c00679
ACS Synth. Biol. 2025, 14, 4897−4906

pH preference of enzymes in the AcalPred dataset. Supplementary Tables : Table S1. Information of wild-type and mutated enzymes used in case studies. Supplementary Tables : Table S2. The measured enzyme activities (U/mL) of PhDac mutants at pH = 5.5. Supplementary Tables : Table S3. The measured enzyme activities (U/mL) of PhDac mutants at pH = 4.5 (PDF)

AUTHOR INFORMATION

Corresponding Authors

Jin-Song Gong − *Key Laboratory of Carbohydrate Chemistry and Biotechnology, Ministry of Education, School of Life Sciences and Health Engineering, Jiangnan University, Wuxi 214122, PR China; Yixing Institute of Food and Biotechnology Co. Ltd, Yixing 214200, PR China*; Email: jinsonggong.bio@hotmail.com

Aidong Yang − *Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, United Kingdom*; orcid.org/0000-0001-5974-247X; Email: aidong.yang@eng.ox.ac.uk

Authors

Sizhe Qiu − *Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, United Kingdom*; orcid.org/0000-0002-1936-1223

Nan-Kai Wang − *Key Laboratory of Carbohydrate Chemistry and Biotechnology, Ministry of Education, School of Life Sciences and Health Engineering, Jiangnan University, Wuxi 214122, PR China; Yixing Institute of Food and Biotechnology Co. Ltd, Yixing 214200, PR China*

Yishun Lu − *Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, United Kingdom; Oxford E-Research Centre, Department of Engineering Science, University of Oxford, Oxford OX1 3QG, United Kingdom*; orcid.org/0000-0003-2345-4470

Jin-Song Shi − *Key Laboratory of Carbohydrate Chemistry and Biotechnology, Ministry of Education, School of Life Sciences and Health Engineering, Jiangnan University, Wuxi 214122, PR China; Yixing Institute of Food and Biotechnology Co. Ltd, Yixing 214200, PR China*; orcid.org/0000-0001-8514-3112

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.5c00679

Author Contributions

#S.Q., N.-K.W., and Y.L. contributed equally. S.Q. constructed the deep learning model, designed point mutations, and produced the first draft. N.W. conducted the experiment and contributed to the first draft. Y.L. assisted in model construction and contributed to model optimization. J.G. participated in the writing and review of the first draft. A.Y. and J.S. supervised this research project and critically reviewed the manuscript.

Notes

The authors declare no competing financial interest.

Biography

S.Q. is a researcher in the fields of deep learning models of enzymes, metabolic modeling, and multi-omics analysis.

ACKNOWLEDGMENTS

This work was financially supported by the National Key Research and Development Program of China (No. 2023YFA0914500), and the National Natural Science Foundation of China (No. 32171261). The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility (10.5281/zenodo.22558) in carrying out this work.

ABBREVIATIONS

BCX	*Bacillus circulans*xylanase
CNN	convolutional neural network
CRZHD	*Clonostachys rosea*zearalenone hydrolase
GlcN	glucosamine
GlcNAc	N-acetylglucosamine
Leaky ReLU	leaky rectified linear unit
MAE	mean absolute error
MSE	mean squared error
PhDac	*Pyrococcus horikoshii*diacetylchitobiose deacetylase
pH_{opt}	optimal pH
RD	residual dense block
R^2	r-squared, the coefficient of determination
RMSE	root mean squared error
TanBlp	*Lactiplantibacillus plantarum*tannase
WT	wild-type

REFERENCES

(1) Kirk, O.; Borchert, T. V.; Fuglsang, C. C. Industrial Enzyme Applications. *Curr. Opin. Biotechnol.* **2002**, *13* (4), 345−351.

(2) Yang, A. S.; Honig, B. On the pH Dependence of Protein Stability. *J. Mol. Biol.* **1993**, *231* (2), 459−474.

(3) Gratacós-Cubarsí, M.; Lametsch, R. Determination of Changes in Protein Conformation Caused by pH and Temperature. *Meat Sci.* **2008**, *80* (2), 545−549.

(4) Di Russo, N. V.; Estrin, D. A.; Martí, M. A.; Roitberg, A. E. pH-Dependent Conformational Changes in Proteins and Their Effect on Experimental pK(a)s: The Case of Nitrophorin 4. *PLoS Comput. Biol.* **2012**, *8* (11), No. e1002761.

(5) Schomburg, I.; Jeske, L.; Ulbrich, M.; Placzek, S.; Chang, A.; Schomburg, D. The BRENDA Enzyme Information system−From a Database to an Expert System. *J. Biotechnol.* **2017**, *261*, 194−206.

(6) The UniProt Consortium UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D523−D531.

(7) Nilsson, A.; Nielsen, J.; Palsson, B. O. Metabolic Models of Protein Allocation Call for the Kinetome. *Cell Syst.* **2017**, *5* (6), 538−541.

(8) Lin, H.; Chen, W.; Ding, H. AcalPred: A Sequence-Based Tool for Discriminating between Acidic and Alkaline Enzymes. *PLoS One* **2013**, *8* (10), No. e75726.

(9) Khan, Z. U.; Hayat, M.; Khan, M. A. Discrimination of Acidic and Alkaline Enzyme Using Chou's Pseudo Amino Acid Composition in Conjunction with Probabilistic Neural Network Model. *J. Theor. Biol.* **2015**, *365*, 197−203.

(10) Wang, X.; Li, H.; Gao, P.; Liu, Y.; Zeng, W. Combining Support Vector Machine with Dual G-Gap Dipeptides to Discriminate between Acidic and Alkaline Enzymes. *Lett. Org. Chem.* **2019**, *16* (4), 325−331.

(11) Li, X.; Dou, Z.; Sun, Y.; Wang, L.; Gong, B.; Wan, L. A Sequence Embedding Method for Enzyme Optimal Condition Analysis. *BMC Bioinf.* **2020**, *21* (1), 512.

(12) Shahraki, M. F.; Atanaki, F. F.; Ariaeenejad, S.; Ghaffari, M. R.; Norouzi-Beirami, M. H.; Maleki, M.; Salekdeh, G. H.; Kavousi, K. A Computational Learning Paradigm to Targeted Discovery of Biocatalysts from Metagenomic Data: A Case Study of Lipase Identification. *Biotechnol. Bioeng.* **2022**, *119* (4), 1115−1128.

(13) Gado, J. E.; Knotts, M.; Shaw, A. Y.; Marks, D.; Gauthier, N. P.; Sander, C.; Beckham, G. T. Machine Learning Prediction of Enzyme Optimum pH. *Nat. Mach. Intell.* **2025**, *7*, 716.
(14) Qiu, S.; Hu, B.; Zhao, J.; Xu, W.; Yang, A. Seq2Topt: A Sequence-Based Deep Learning Predictor of Enzyme Optimal Temperature. *Briefings Bioinf.* **2025**, *26* (2), bbaf114.
(15) Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; Rives, A. Language Models Enable Zero-Shot Prediction of the Effects of Mutations on Protein Function. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 29287−29303.
(16) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; Dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379* (6637), 1123−1130.
(17) Zaretckii, M.; Buslaev, P.; Kozlovskii, I.; Morozov, A.; Popov, P. Approaching Optimal pH Enzyme Prediction with Large Language Models. *ACS Synth. Biol.* **2024**, *13* (9), 3013−3021.
(18) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System: ACM, **2016**.
(19) Huang, Z.; Mao, X.; Lv, X.; Sun, G.; Zhang, H.; Lu, W.; Liu, Y.; Li, J.; Du, G.; Liu, L. Engineering Diacetylchitobiose Deacetylase from Pyrococcus Horikoshii towards an Efficient Glucosamine Production. *Bioresour. Technol.* **2021**, *334*, 125241.
(20) Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat. Biotechnol.* **2017**, *35* (11), 1026−1028.
(21) Maas, A. L.; Hannun, A. Y.; Ng, A. Y. *Rectifier nonlinearities improve neural network acoustic models*. http://robotics.stanford.edu/amaas/papers/relu_hybrid_icml2013_final.pdf (accessed 05 August 2023).
(22) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv*, **2014**.
(23) Wang, L.; Hu, M.; Tao, Y. Efficient Production of D-Glucosamine by Diacetylchitobiose Deacetylase Catalyzed Deacetylation of N-Acetyl-D-Glucosamine. *Biotechnol. Lett.* **2022**, *44* (3), 473−483.
(24) Qiu, S.; Zhao, S.; Yang, A. DLTKcat: Deep Learning-Based Prediction of Temperature-Dependent Enzyme Turnover Rates. *Briefings Bioinf.* **2023**, *25* (1), bbad506.
(25) Jiang, Z.; Niu, T.; Lv, X.; Liu, Y.; Li, J.; Lu, W.; Du, G.; Chen, J.; Liu, L. Secretory Expression Fine-Tuning and Directed Evolution of Diacetylchitobiose Deacetylase by Bacillus Subtilis. *Appl. Environ. Microbiol.* **2019**, *85* (17), No. e01076-19.
(26) Goh, G. B.; Laricheva, E. N.; Brooks, C. L. 3rd. Uncovering pH-Dependent Transient States of Proteins with Buried Ionizable Residues. *J. Am. Chem. Soc.* **2014**, *136* (24), 8496−8499.
(27) Mishra, P.; Patni, D.; Jha, S. K. A pH-Dependent Protein Stability Switch Coupled to the Perturbed pKa of a Single Ionizable Residue. *Biophys. Chem.* **2021**, *274*, 106591.
(28) Olland, A. M.; Strand, J.; Presman, E.; Czerwinski, R.; Joseph-McCarthy, D.; Krykbaev, R.; Schlingmann, G.; Chopra, R.; Lin, L.; Fleming, M.; Kriz, R.; Stahl, M.; Somers, W.; Fitz, L.; Mosyak, L. Triad of Polar Residues Implicated in pH Specificity of Acidic Mammalian Chitinase. *Protein Sci.* **2009**, *18* (3), 569−578.
(29) Kim, S. H.; Pokhrel, S.; Yoo, Y. J. Mutation of Non-Conserved Amino Acids Surrounding Catalytic Site to Shift pH Optimum of Bacillus Circulans Xylanase. *J. Mol. Catal. B: Enzym.* **2008**, *55* (3−4), 130−136.
(30) Pan, H.; Zhan, J.; Yang, H.; Wang, C.; Liu, H.; Zhou, H.; Zhou, H.; Lu, X.; Su, X.; Tian, Y. Improving the Acid Resistance of Tannase TanBLp (AB379685) from Lactobacillus Plantarum ATCC14917T by Site-Specific Mutagenesis. *Indian J. Microbiol.* **2022**, *62* (1), 96−102.
(31) Dotsenko, A.; Sinelnikov, I.; Zorov, I.; Denisenko, Y.; Rozhkova, A.; Shcherbakova, L. The Protein Engineering of Zearalenone Hydrolase Results in a Shift in the pH Optimum of the Relative Activity of the Enzyme. *Toxins* **2024**, *16* (12), 540.
(32) Paoli, P.; Taddei, N.; Fiaschi, T.; Veggi, D.; Camici, G.; Manao, G.; Raugei, G.; Chiti, F.; Ramponi, G. The Contribution of Acidic Residues to the Conformational Stability of Common-Type Acylphosphatase. *Arch. Biochem. Biophys.* **1999**, *363* (2), 349−355.
(33) Zhang, Y.; Guan, F.; Xu, G.; Liu, X.; Zhang, Y.; Sun, J.; Yao, B.; Huang, H.; Wu, N.; Tian, J. A Novel Thermophilic Chitinase Directly Mined from the Marine Metagenome Using the Deep Learning Tool Preoptem. *Bioresour. Bioprocess.* **2022**, *9* (1), 54.
(34) Li, M.; Lu, Z.; Wu, Y.; Li, Y. BACPI: A Bi-Directional Attention Neural Network for Compound-Protein Interaction and Binding Affinity Prediction. *Bioinformatics* **2022**, *38* (7), 1995−2002.
(35) Lauriola, I.; Gallicchio, C.; Aiolli, F. Enhancing Deep Neural Networks via Multiple Kernel Learning. *Pattern Recognit.* **2020**, *101*, 107194.
(36) Yang, B.; Wang, L.; Wong, D.; Chao, L. S.; Tu, Z. Convolutional Self-Attention Networks. *arXiv*, **2019**.
(37) Borawar, L.; Kaur, R. ResNet: Solving Vanishing Gradient in Deep Networks. Springer Nature Singapore, 2023, 235−247.
(38) Killoran, N.; Lee, L. J.; Delong, A.; Duvenaud, D.; Frey, B. J. Generating and Designing DNA with Deep Generative Models. *arXiv*, **2017**.
(39) Jiang, H.; Wang, J.; Yang, Z.; Chen, C.; Yao, G.; Bao, S.; Wan, X.; Ding, J.; Wang, L. MPEK: A Multi-Task Learning Based on Pre-Trained Language Model for Predicting Enzymatic Reaction Kinetic Parameters. *Res. Square* **2024**, bbae387.
(40) Yu, H.; Deng, H.; He, J.; Keasling, J. D.; Luo, X. UniKP: A Unified Framework for the Prediction of Enzyme Kinetic Parameters. *Nat. Commun.* **2023**, *14* (1), 8211.
(41) Bjerregaard, A.; Groth, P. M.; Hauberg, S.; Krogh, A.; Boomsma, W. Foundation Models of Protein Sequences: A Brief Overview. *Curr. Opin. Struct. Biol.* **2025**, *91*, 103004.
(42) Hazel, J. R.; Garlick, W. S.; Sellner, P. A. The Effects of Assay Temperature upon the pH Optima of Enzymes from Poikilotherms: A Test of the Imidazole Alphastat Hypothesis. *J. Comp. Physiol.* **1978**, *123* (2), 97−104.
(43) Maurel, P.; Douzou, P.; Waldmann, J.; Yonetani, T. Enzyme Behaviour and Molecular Environment. The Effects of Ionic Strength, Detergents, Linear Polyanions and Phospholipids on the pH Profile of Soluble Cytochrome Oxidase. *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol.* **1978**, *525* (2), 314−324.

https://doi.org/10.1021/acssynbio.5c00679
ACS Synth. Biol. 2025, 14, 4897−4906