DeepPH : A Multimodal Deep Learning Model for Predicting Enzyme Optimal pH Range

Anonymous Author(s)

Abstract

Enzymes are essential biological catalysts whose activity is strongly influenced by pH, making accurate prediction of pH optima critical for both industrial and biochemical applications. Recent state-of-the-art (SOTA) methods have improved prediction of optimal pH value. However, enzymes typically function across a range of pH values rather than at a fixed point. In addition, three-dimensional (3D) structural features, key determinants of enzyme behavior, are often overlooked due to their scarcity and bias toward well-characterized enzymes, which limits the model generalizability. To address these challenges, we introduce DeepPH, regression framework that integrates sequence embeddings with predicted structural representations to capture spatial determinants of pH sensitivity. DeepPH employs a spatial radius-based structural encoding scheme, a multimodal attention pooling for fusing sequence and structure information, and a novel loss function designed to handle supervision imbalance. We further propose an interval-aware prediction setting to better reflect the inherent nature of optimal pH as a range. Extensive experiments on test sets demonstrate that DeepPH outperforms existing SOTA models in both average-based and interval-aware pH prediction tasks, demonstrating strong generalization even on enzyme sequences with extreme lengths. The code and datasets used in this study are publicly available at https://anonymous.4open.science/r/DeepPH-anon/.

CCS Concepts

• **Applied computing** → **Bioinformatics**; **Computational biology**; • **Computing methodologies** → **Neural networks**.

Keywords

Deep learning, Enzyme, pH prediction

ACM Reference Format:

Anonymous Author(s). 2025. DeepPH : A Multimodal Deep Learning Model for Predicting Enzyme Optimal pH Range. In *Proceedings of 16th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB '25).* ACM, Philadelphia, PA, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Enzymes are biological catalysts that accelerate chemical reactions, playing a crucial role in various biochemical and industrial processes[4]. They have a wide array of applications, from food processing and pharmaceutical manufacturing to biofuel production [14, 16]. However, a major challenge in the effective application of enzymes is their strong sensitivity to the pH of the reaction environment[1, 28]. Enzymes generally function most efficiently within an optimal pH range, where their catalytic activity is at its peak. Deviations from this optimal range can significantly reduce their activity or even lead to complete inactivation [2, 8, 23, 24]. This phenomenon is primarily attributed to changes in the protonation state of the catalytic center and the loss of protein structural stability caused by pH variations [22, 32]. Therefore, accurately predicting and optimizing of pH-dependent enzyme behavior is critical for improving their industrial applications.

In recent years, protein language models (PLMs), such as the ESM [17, 20] and ProtTrans [6] series, have shown increasing potential in enzyme research, particularly in modeling enzyme behavior under varying biochemical conditions[25]. For example, the EpHod method predicts the optimal pH of enzymes using protein sequence embeddings such as ESM-1v, providing a valuable tool for studies of enzyme function [9, 20]. Based on this foundation, Zhang et al. proposed VENUS-DREAM[34], a retrieval-enhanced meta-learning framework that uses sequence embeddings from ESM-2[17], combining k-nearest neighbors (KNN) and few-shot learning. Using related enzymes as contextual support, VENUS-DREAM improves prediction accuracy, especially in limited data scenarios. Both methods utilize PLMs to extract rich sequence features and have demonstrated promising performance in large-scale pH data sets.

However, despite their success, existing optimal pH estimation models rely primarily on sequence-derived information and often neglect explicit three-dimensional (3D) structural characteristics, which are well-established determinants of enzymatic function and pH sensitivity [3, 5, 13]. The structural conformation of proteins, including enzymes, plays a fundamental role in understanding their physicochemical properties and functional mechanisms. Recent studies, such as GPSFun[33], GraphEC[29], and DeepFRI[10], have shown that the incorporation of predicted structures can significantly improve performance in functional classification tasks. These findings highlight the importance of integrating statistical information to improve predictive accuracy and broaden the scope of computational approaches in enzyme characterization.

Although structural features are critical in enzyme pH study, their incorporation at scale remains challenging due to the limited availability of high-quality experimental structures, especially for newly discovered or poorly annotated enzymes [19]. In one estimate, as little as 22% of the enzyme sequence had an experimentally solved structure in the Protein Data Bank (PDB) [31], highlighting the gap between sequence databases and structural data. The sparsity could cause the AI models training on structural features to have a restricted and potentially biased sample of enzymes —

ACM-BCB '25, Philadelphia, PA

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

well-studied enzymes dominate the structural databases, whereas newly discovered or poorly annotated enzymes are underrepresented. As a result, models can learn biased patterns that reflect only the properties of these well-characterized enzymes, ignoring the diversity of enzyme space. In fact, previous enzyme property predictors often avoided using 3D structure data, precisely because such data were scarce and non-representative. By relying mostly on sequence based training sets, early models still inherently the same biases — too many example of enzymes with common shapes and pH value close to neutral. As a result, the model could predict properties accurately for enzymes that look like the ones in the training data but performed poorly when trying to predict properties for rarer or less familiar enzymes with different structures.

To address the limitation of existing methods, we proposed DeepPH, a regression framework for predicting the optimal pH range of enzymes through the integrated use of sequence and structural data. Unlike prior models that rely exclusively on sequence-derived features, DeepPH explicitly incorporates 3D structural representation, including predicted conformations, to capture spatial determinants of enzyme behavior that are critical for pH sensitivity and often overlooked. The contributions of our method are as follows:

- Spatial radius-based enzyme structure representation learning to incorporate structural information ignored by prior studies.
- Multimodal Attention-based Alignment for effective integration of Sequence and Structure Features.
- A novel loss to handle diverse supervision signals and improve robustness under label imbalance.
- Interval-aware prediction settings tailored to the nature of optimal pH as a range.

We extensively evaluated our proposed method, DeepPH, on several benchmark enzyme datasets against multiple state of art models, and the results show that incorporating structural information improves predictive accuracy over sequence-only models under both evaluation modes. This demonstrates the effectiveness of DeepPH in predicting pH ranges and highlights its potential for broad applicability in enzyme characterization.

2 Method

In this study, we proposed DeepPH, a multimodal regression framework designed to predict the optimal pH range — specifically the lower and upper bounds — for enzyme activity by integrating both sequence and structural information. The model consists of three components (Fig. 1): (1) a residue level sequence embedding model, (2) a spatial radius structure module capture spatial context; and (3) an attention-based mechanism for fusing sequence and structure representations.

2.1 Contextual Feature Projection of Enzyme Residues

2.1.1 Residue Level Sequence Embedding. The sequence module utilized the ProtT5 pre-trained protein language model to extract contextual embeddings for each amino acid residue. Given a primary enzyme sequence represented as $E_{\text{seq}} \in \mathbb{R}^L$, where L is the length of the sequence, a set of generic embedding functions $f^*(\cdot)$ maps it to a high-dimensional representation $f^*(E_{\text{seq}}) \in \mathbb{R}^{L\times D}$, where $D = 1024$ denotes the fixed output dimension of the ProtT5-XL-U50 model [7]. This model is a widely used protein language model that encodes semantic information for each residue based on the full sequence context, capturing both local motifs and long-range dependencies, and supporting inference on long sequences without imposing strict length restrictions [11, 29]. To enable feature fusion with the structure representation, the embeddings are further projected into a lower-dimensional space via a linear layer, yielding the final sequence feature $\mathbf{h}_i^{\text{seq}} \in \mathbb{R}^{d_q}$ for the i-th residue, where d_q is equal to 128.

2.1.2 Spatial Radius Structure EGNN. To capture structural and biochemical contexts beyond the primary sequence, we constructed residue-level representations $\mathbf{v}_i = [\mathbf{f}_i^{ss} \| \mathbf{f}_i^{geo} \| \mathbf{f}_i^{chem}] \in \mathbb{R}^F$. Specifically, $\mathbf{f}_i^{ss}$, $\mathbf{f}_i^{geo}$ and $\mathbf{f}_i^{chem}$ represent secondary structure, spatial geometry, and physicochemical properties of residue i, respectively.

Graph Construction. We denote the set of residues in an enzyme of length L by $\{1, 2, \ldots, L\}$. Let $\mathbf{x}_i = \{\mathbf{x}_i^N, \mathbf{x}_i^{C_\alpha}, \mathbf{x}_i^C, \mathbf{x}_i^O, \mathbf{x}_i^R\}$, where each $\mathbf{x}_i^a \in \mathbb{R}^3$ represents the predicted Cartesian coordinates of the backbone atoms $\{N, C_\alpha, C, O\}$ and the side-chain center $\{R\}$ for residue i, obtained by ESMFold[18]. We then define an undirected spatial graph:

$$G = (V, E), \tag{1}$$

where $V = \{1, 2, ..., L\}$ and $E = \{(i, j) | \|\mathbf{x}_i - \mathbf{x}_j\|_2 \le r\}$, with threshold $r = 15$Å[29]. Equivalently, the adjacency matrix $A \in \{0, 1\}^{L\times L}$ is given by:

$$A_{ij} = \begin{cases} 1, & \text{if } \|x_i - x_j\|_2 \le 15\,\text{Å}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

This radius-based connectivity ensures that each node is linked to all residues within a 15Å spatial neighborhood, thereby encoding local structural context for downstream message passing.

Secondary Structure Features. The descriptors $\mathbf{f}_i^{ss} \in \mathbb{R}^8$ were extracted using DSSP[12]. For an enzyme length L, DSSP calculates the secondary structure for each residue i, which we represent as $\mathbf{f}_i^{ss} \in \{H, B, E, G, I, T, S, C\}$, denotes the secondary structure label assigned to the residue i based on hydrogen bond energies and local geometry. The energy of a hydrogen bond between residues i and j is defined as:

$$EN_{ij} = 0.084\{\frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}}\} \cdot C \quad kcal/mol, \tag{3}$$

where $r_{i,j} = \|x_i - x_j\| \in \mathbb{R}$ denotes the Euclidean distance between the atom x_i and x_j, and $C = 332$ is a constant that converts electrostatic energy to $kcal/mol$. A hydrogen bond is considered significant if $EN_{i,j} < -0.5\ kcal/mol$, typically. Based on recurring hydrogen bond patterns, a rule-based mapping assigns the secondary structure label.

3D Geometrical Features. The geometric feature vector for each residue i, denoted as $\mathbf{f}_i^{\text{geo}} \in \mathbb{R}^{185}$, designed to encode spatial relationships with neighboring residues, is composed of four components:

(1) Angular features. To capture the local geometric conformation of the protein backbone, we computed six angular features for

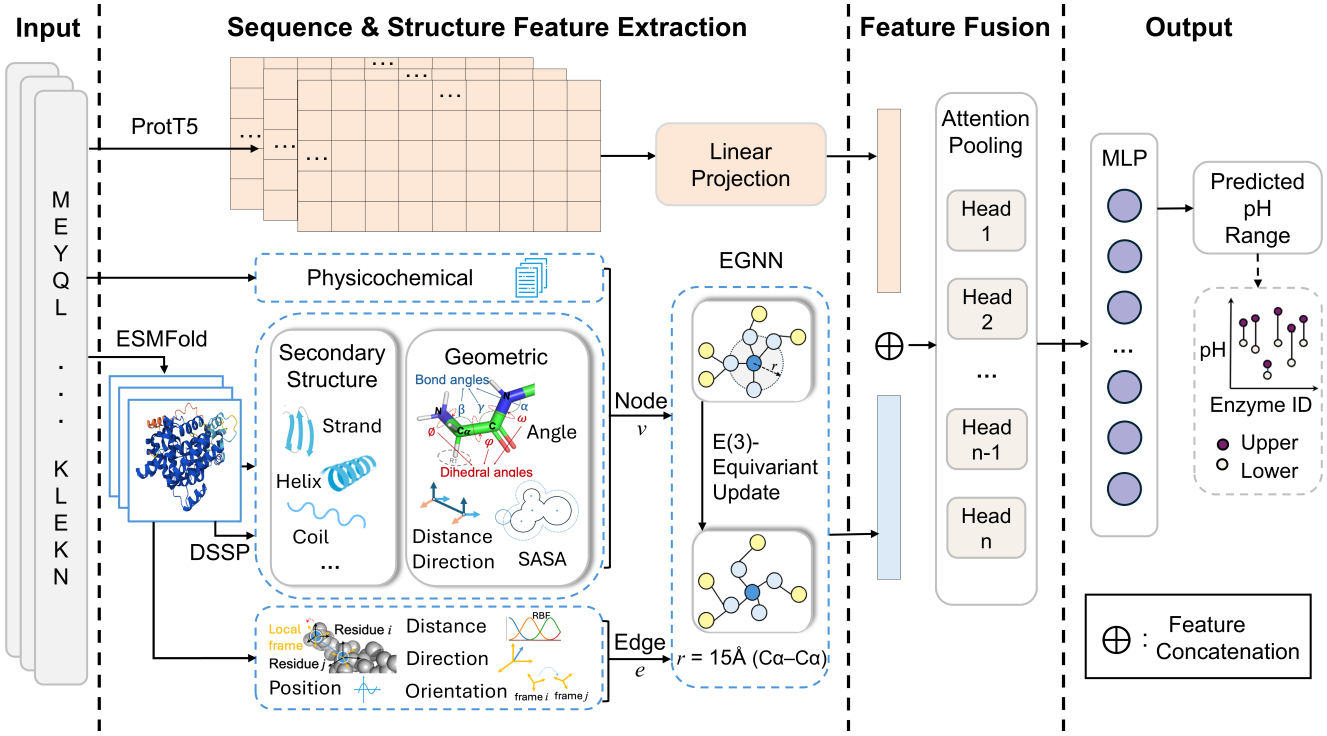

Figure 1: Overview of the DeepPH framework for enzyme pH range prediction. The model takes an amino acid sequence as input and extracts both sequence and structure features. Sequence features were obtained by projecting ProtT5 embeddings to a lower dimension using a linear layer. Structural features were derived from an EGNN, with each nodes features v encoded by comprehensive information drived from ESMFold-predicted structures and each edge e is represented by a vector combining position, distance, direction, and rotational orientation features between residues. Sequence and structure features were concatenated ($\bigoplus$) and integrated via a multi-head attention pooling layer. The fused representation was then decoded by a multilayer perceptron to predict the upper and lower bounds of the optimal pH range.

each residue, including three dihedral angles (ϕ_i, ψ_i, ω_i) and three bond angles (α_i, β_i, γ_i).

The three dihedral angles are defined as $\phi_i = \mathrm{D}(x_{i-1}^C, x_i^N, x_i^{C_\alpha}, x_i^C)$, $\psi_i = \mathcal{D}(x_i^N, x_i^{C_\alpha}, x_i^C, x_{i+1}^N)$, and $\omega_i = \mathcal{D}(x_i^{C_\alpha}, x_i^C, x_{i+1}^N, x_{i+1}^{C_\alpha})$, where the dihedral angle function is defined as

$$\mathcal{D}(p_1, p_2, p_3, p_4) = \arctan 2\left(\mathbf{b} \cdot (\mathbf{n}_1 \times \mathbf{n}_2),\ \mathbf{n}_1 \cdot \mathbf{n}_2\right), \tag{4}$$

with $\mathbf{n}_1 = (p_2 - p_1) \times (p_3 - p_2)$, $\mathbf{n}_2 = (p_3 - p_2) \times (p_4 - p_3)$, and $\mathbf{b} = p_3 - p_2$. These torsion angles describe the rotation between consecutive peptide planes.

The three bond angles are defined as $\alpha_i = \mathcal{B}(x_{i-1}^C, x_i^N, x_i^{C_\alpha})$, $\beta_i = \mathcal{B}(x_i^N, x_i^{C_\alpha}, x_i^C)$, and $\gamma_i = \mathcal{B}(x_i^{C_\alpha}, x_i^C, x_{i+1}^N)$. Each bond angle is computed using

$$\mathcal{B}(q_1, q_2, q_3) = \arccos\left(\frac{(q_1 - q_2) \cdot (q_3 - q_2)}{\|q_1 - q_2\| \cdot \|q_3 - q_2\|}\right), \tag{5}$$

with q_2 being the central atom. To ensure rotational continuity, each angle is further encoded as a two-dimensional vector $[\cos(\theta), \sin(\theta)]$, resulting in a 12-dimensional angular feature vector per residue.

(2) Distance features. We extracted intra-residue spatial relationships by computing the Euclidean distances between all 10 unique atom pairs among $Atom = \{N, C_\alpha, C, O, R\}$ within each residue. For each residue i, the distances are denoted as $[\|x_i^{a_1} - x_i^{a_2}\|] \in \mathbb{R}^{\binom{5}{2}} = \mathbb{R}^{10}$, where $a_1, a_2 \in Atom$ and $a_1 \neq a_2$. These distances were further expanded using radial basis function (RBF) encoding. In our implementation, we set the dimension of RBF encoding as 16, resulting in 160-dimensional distance features per residue.

(3) Directional features. To capture the local spatial orientation of atoms within each residue, we constructed a local coordinate frame centered at the C_α atom. Specifically, we utilized two primary backbone direction vectors: from C_α to N, and from C to C_α, and normalize them as orthogonal basis vectors to build the local coordinate system (hereafter referred to as the local frame) . The local frame matrix is denoted as $\mathbf{R}_i = [\mathbf{b}, \mathbf{n}, \mathbf{b} \times \mathbf{n}] \in \mathbb{R}^{3\times3}$, where

$$\mathbf{u} = \frac{\mathbf{x}_i^{C_\alpha} - \mathbf{x}_i^N}{\|\mathbf{x}_i^{C_\alpha} - \mathbf{x}_i^N\|}, \quad \mathbf{v} = \frac{\mathbf{x}_i^C - \mathbf{x}_i^{C_\alpha}}{\|\mathbf{x}_i^C - \mathbf{x}_i^{C_\alpha}\|}, \quad \mathbf{b} = \frac{\mathbf{u} - \mathbf{v}}{\|\mathbf{u} - \mathbf{v}\|}, \quad \mathbf{n} = \frac{\mathbf{u} \times \mathbf{v}}{\|\mathbf{u} \times \mathbf{v}\|}. \tag{6}$$

Next, for each atom $a \neq C_\alpha$, we computed the unit direction vector from C_α to a and project it onto the local frame:

$$\mathbf{d}_a = \mathbf{R}_i^\top \cdot \frac{\mathbf{x}_i^a - \mathbf{x}_i^{C_\alpha}}{\|\mathbf{x}_i^a - \mathbf{x}_i^{C_\alpha}\|} \in \mathbb{R}^3. \tag{7}$$

Finally, we concatenated the projected vectors of the four atoms to form a 12-dimensional directional feature.

(4) Solvent-Accessible Surface Area (SASA). The SASA of residue i is computed by simulating a virtual probe rolling over the molecular surface and summing the resulting exposed surface areas. The values are obtained from DSSP.

Physicochemical Features. To complement geometric information with the chemical context specific to the residues, we defined a characteristic vector $\mathbf{f}_i^{chem}$, indicating its physicochemical category: aliphatic, aromatic, uncharged polar, acidic, or basic. Furthermore, we included continuous descriptors, which were represented as vector $\mathbf{k}_i = \{mW_i, pKa_i, pKb_i, pK_{X_i}, pI_i, Hyd_i^2, Hyd_i^7\}$[15]. Here, mW_i denotes the molecular weight, pKa_i and pKb_i are the dissociation constants of the carboxyl and amino groups, pK_{X_i} represents the ionization constant of the side chain, pI_i is the isoelectric point at which the residue does not carry a net charge, and Hyd_i^2 and Hyd_i^7 are the hydrophobicity values measured at pH 2 and pH 7. The final physicochemical representation for the residue i is given by concatenating the five-dimensional one-hot encoding of physicochemical category and the seven-dimensional continuous descriptor vector, denoted as $\mathbf{f}_i^{chem} \in \mathbb{R}^{12}$. Min-max normalization was applied to the continuous descriptors.

In addition to the node features, each edge $(i, j) \in E$ is associated with a feature vector $\mathbf{e}_{ij} = [\mathbf{f}^{pos}{}_{ij} \| \mathbf{f}^{dist}{}_{ij} \| \mathbf{f}^{dir}{}_{ij} \| \mathbf{f}^{ori}{}_{ij}] \in \mathbb{R}^{450}$ that encodes pairwise geometric relationships between residues. Here, $\mathbf{f}^{pos}{}_{ij} \in \mathbb{R}^{16}$ denotes the positional embedding based on sequence distance using sinusoidal functions; $\mathbf{f}^{dist}{}_{ij} \in \mathbb{R}^{400}$ represents radial basis function (RBF) encoded Euclidean distances between all atom-type pairs; $\mathbf{f}^{dir}{}_{ij} \in \mathbb{R}^{30}$ captures directional vectors in local coordinate frames; and $\mathbf{f}_{ij}^{ori} \in \mathbb{R}^4$ encodes relative rotational orientations using quaternion representations.

The node features $\mathbf{v}_i$ and edge feature $\mathbf{e}_{ij}$ are mapped to structure embedding $\mathbf{h}_i^{str} \in \mathbb{R}^{d_s}$, where d_s is equal to 128, through a single layer EGNN encoder, detailed in Algorithm 1.

Algorithm 1 Structure Embedding via EGNN

Require: Node features $\{\mathbf{v}_i\}$, node coordinates $\{\mathbf{x}_i\}$, edge set $\mathcal{E}$, edge features $\{\mathbf{e}_{ij}\}$
Ensure: Structure embeddings $\{\mathbf{h}_i^{str}\}$

Define $\mathcal{N}(i) = \{j : (i, j) \in \mathcal{E}\}$ as neighbors of node i
$\mathbf{h}_i \leftarrow \text{Linear}(\mathbf{v}_i)$ ▷ Node feature embedding
for $l = 1$ to L **do**
  Initialize $\Delta\mathbf{x}_i \leftarrow \mathbf{0}$ and $\mathbf{m}_i^{agg} \leftarrow \mathbf{0}$ for all nodes i
  for each edge $(i, j) \in \mathcal{E}$ **do**
    $\mathbf{r}_{ij} \leftarrow \mathbf{x}_i - \mathbf{x}_j$
    $r_{ij} \leftarrow \|\mathbf{r}_{ij}\|^2$
    $\mathbf{m}_{ij} \leftarrow \text{EdgeMLP}([\mathbf{h}_i, \mathbf{h}_j, r_{ij}, \mathbf{e}_{ij}])$
    $\Delta\mathbf{x}_i \mathrel{+}= \mathbf{r}_{ij} \cdot \text{CoordMLP}(\mathbf{m}_{ij})$ ▷ Accumulate coord changes
    $\mathbf{m}_i^{agg} \mathrel{+}= \mathbf{m}_{ij}$ ▷ Accumulate edge messages
  end for
  for each node i **do**
    $\mathbf{x}_i \leftarrow \mathbf{x}_i + \frac{1}{|\mathcal{N}(i)|}\Delta\mathbf{x}_i$ ▷ Mean aggregation over neighbors
    $\mathbf{h}_i \leftarrow \mathbf{h}_i + \text{NodeMLP}([\mathbf{h}_i, \mathbf{m}_i^{agg}])$ ▷ Sum aggregation for features
  end for
end for
$\mathbf{h}_i^{str} \leftarrow \text{Linear}(\mathbf{h}_i)$ ▷ Output projection

2.2 Multi-head Feature Attention Pooling

To integrate structural and sequence information at residue level, we defined a representation by concatenating the projected sequence embedding $\mathbf{h}_i^{\text{seq}} \in \mathbb{R}^{d_q}$ projected from ProtT5 and the structure embedding $\mathbf{h}_i^{\text{str}} \in \mathbb{R}^{d_s}$ from EGNN. The resulting multimodal residue feature vector is given by:

$$\mathbf{h}_i = \left[\mathbf{h}_i^{\text{seq}} \,\|\, \mathbf{h}_i^{\text{str}}\right] \in \mathbb{R}^{d_q + d_s}. \tag{8}$$

In our implementation, we set the sequence and structure embedding dimension to $d_q = d_s = 128$. The residue-wise embeddings are then concatenate into a matrix $\mathbf{H} \in \mathbb{R}^{L \times d}$, where $d = d_q + d_s$ denotes the dimensionality of the residue representation. We applied a multi-head attention pooling mechanism to aggregate these features into a global enzyme representation:

$$\text{AttPooling}(\mathbf{H}) = \sum_{i=1}^{n} \mathbf{A}_i^T \cdot \mathbf{H}, \tag{9}$$

where n representing the number of attention heads. In our configuration, we set $n = 4$ attention heads. $\mathbf{A}_i$ denotes the attention weight matrix of the i-th head, defined as:

$$\mathbf{A} = \text{softmax}\left(\mathbf{W}_2 \cdot \tanh(\mathbf{W}_1 \cdot \mathbf{H})\right), \tag{10}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times h}$ and $\mathbf{W}_2 \in \mathbb{R}^{h \times n}$ are learnable transformation matrices, with h being the intermediate attention dimension. We set $h = 16$. The softmax operation is applied along the sequence dimension to ensure proper attention weight normalization.

This attention mechanism enables the model to automatically identify and emphasize the most informative residues for the prediction task, providing both interpretability and improved performance through adaptive feature selection. After attention pooling, the resulting enzyme representation is passed through a two-layer MLP The first layer maps the input to a hidden space with ELU activation, followed by a second layer that outputs a vector corresponding to the predicted pH range:

$$\mathbf{y} = \left[\hat{y}^{\min}, \hat{y}^{\max}\right]. \tag{11}$$

2.3 Loss function

To accommodate both precise ($y^{\max} - y^{\min} < 1$) and interval-based ($y^{\max} - y^{\min} \geq 1$) pH annotations, we designed a compound loss function. This loss consists of four components:

Average Deviation Loss. For non-interval samples $\mathcal{P}$, we compute the normalized squared error between the predicted midpoint and the true midpoint:

$$\mathcal{L}_{\text{avg}} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \frac{(y_i - \hat{y}_i)^2}{\text{Var}(y)}, \tag{12}$$

$$y_i = \frac{y_i^{\min} + y_i^{\max}}{2}, \quad \hat{y}_i = \frac{\hat{y}_i^{\min} + \hat{y}_i^{\max}}{2}, \tag{13}$$

where $y_i^{\min}$ and $y_i^{\max}$ denote the annotated lower and upper bounds of the optimal pH range for enzyme i, while $\hat{y}_i^{\min}$ and $\hat{y}_i^{\max}$ are the corresponding predicted values produced by the model. $\mathrm{Var}(y)$ is the variance of ground-truth midpoints across the training set, used for normalization.

Range Endpoint Loss. For interval samples $\mathcal{R}$, we penalize the deviation of predicted endpoints from ground-truth:

$$\mathcal{L}_{\text{range}} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \left[\frac{\left(y_i^{\min} - \hat{y}_i^{\min}\right)^2}{\mathrm{Var}(y^{\min})} + \frac{\left(y_i^{\max} - \hat{y}_i^{\max}\right)^2}{\mathrm{Var}(y^{\max})} \right]. \quad (14)$$

Range Direction Loss. To ensure $\hat{y}_i^{\min} \le \hat{y}_i^{\max}$, we apply a directional constraint:

$$\mathcal{L}_{\text{dir}} = \sum_i \max\left(0, \hat{y}_i^{\min} - \hat{y}_i^{\max}\right). \quad (15)$$

Range Size Matching Loss. We also encourage predicted range widths to match ground-truth intervals:

$$\mathcal{L}_{\text{size}} = \sum_{i \in \mathcal{R}} \left| (\hat{y}_i^{\max} - \hat{y}_i^{\min}) - (y_i^{\max} - y_i^{\min}) \right|. \quad (16)$$

Overall Loss. The final loss is a weighted combination:

$$\mathcal{L}_{\text{RangeR2}} = \mathcal{L}_{\text{avg}} + \mathcal{L}_{\text{range}} + \mathcal{L}_{\text{dir}} + 0.1 \cdot \mathcal{L}_{\text{size}}. \quad (17)$$

This design enables the model to handle diverse supervision signals and improves its robustness in pH interval prediction, especially when experimental uncertainty is involved. Implementation details, including training hyperparameters and optimization setup, are provided in Supplementary Sections A.1 and A.2.

3 Experimental

3.1 Dataset

The overall dataset partitioning and filtering strategy is illustrated in Fig. 2. The dataset used in this study was obtained from the BRENDA[27] database (released in January 2023). To construct a diverse and non-redundant dataset, we constructed a diverse and non-redundant dataset by applying MMseqs2[21] filtering in two stages. At first, enzyme entries with pairwise sequence identity above 20% were removed to ensure diversity. Secondly, to avoid overlaps with the EpHod and VENUS-DREAM training set and ensure fair model evaluation, we further applied MMseqs2 clustering with a 20% identity and 80% alignment coverage threshold. This resulted in 4,054 enzyme entries. Based on this filtered BRENDA-derived dataset, we employed holdout method by partitioning 3,251 sequences for training and 803 for testing using the same identity and coverage thresholds. The training set was further split into training and validation subsets at an 8:2 ratio. This test set was used to evaluate the performance of DeepPH and compare it with baseline models including VENUS-DREAM.

Notably, EpHod further performs large-scale pretraining on over 1.71 million enzyme sequences. To prevent potential data leakage from this pretraining corpus, we conducted an additional MMseqs2-based filtering step (20% identity, 80% coverage) to exclude test samples with substantial similarity. This yielded a stricter subset of 534 test sequences, which was used exclusively for fair comparison

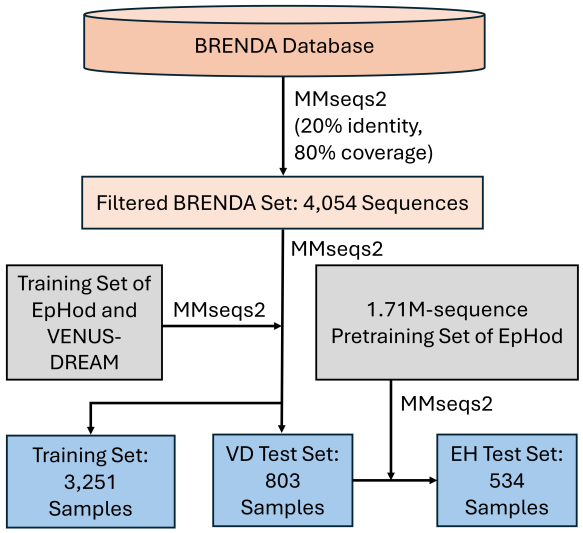

Figure 2: Overview of dataset partitioning and filtering process. A total of 4,054 enzyme sequences were filtered from the BRENDA database using MMseqs2 (20% identity, 80% coverage). From this set, three subsets were constructed: (1) training set (3,251 samples); (2) VD test set (803 samples) for evaluating DeepPH and VENUS-DREAM; and (3) EH test set (534 samples) for EpHod and DeepPH.

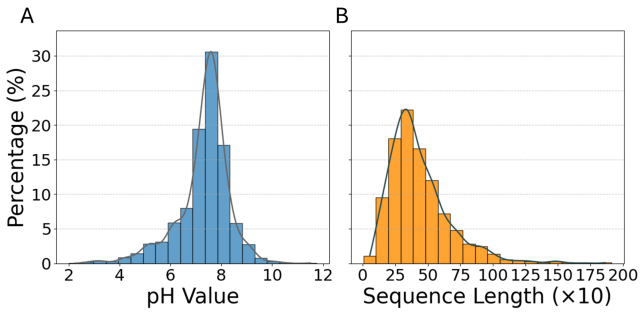

Figure 3: Training data overview. (A) Optimal pH values are mainly distributed between pH 6 and 9, peaking near neutral pH. (B) Most enzyme sequences are short, though some extend to much longer lengths. The dataset exhibits both diversity and skewness, which are relevant for subsequent model training.

with EpHod. We denoted the two test set as VD (803 samples) and EH (534 samples) test set for evaluation against VENUS-DREAM and EpHod respectively.

The dataset distribution of pH values and the sequence lengths is shown in Fig. 3. As shown in the left panel (Fig. 3A), most enzymes exhibit optimal activity within the pH 6–9 range, with a peak close to neutral, indicating a general preference for this mildly acidic to slightly alkaline window. The right panel (Fig. 3B) shows the sequence length distribution, indicating that most enzyme sequences range from 200 to 600 amino acids, with a peak around 350 residues.

A small number of sequences extend beyond 1,000 residues, forming a long-tailed distribution. These observations suggest a degree of imbalance in the dataset, particularly in the prevalence of neutral pH enzymes and shorter sequences, which may affect model generalization. Recognizing such imbalances is critical to designing robust models and interpreting predictive performance.

3.2 Baseline Models

To comprehensively assess the predictive performance of DeepPH, we compare it against two categories of representative baselines: (1) EpHod-based models, including Support Vector Regression (SVR), Regularized Linear Attention Transformer (RLATtr), and their ensemble (Ensemble), and (2) meta-learning models from the VENUS-DREAM framework (MAML, Reptile). Since the original implementations of these baselines are not publicly available, we reproduced and trained them based on the descriptions in their respective publications to ensure comparability (Details are provided in the Supplementary Section A.3). Specifically, evaluation against EpHod-based models is conducted on the EH test set (534 samples), which excludes sequences that potentially overlap with EpHod's training data, while comparison with VENUS-DREAM models is performed on the VD test set (803 samples). This dual test setup enables a fair and informative evaluation across different modeling assumptions and data dependencies.

3.3 Evaluation Metrics

Unlike existing methods that output a single-point prediction for the optimal pH, our model predicts a range $[\hat{y}_i^{\min}, \hat{y}_i^{\max}]$, reflecting both the uncertainty and variability observed in enzyme annotations. To enable fair and informative evaluation, we adopt two complementary strategies:

Average-based Evaluation. For baseline models producing a single predicted value $\hat{y}_i \in \mathbb{R}$, we directly utilize their predictions. Conversely, for our interval-based predictions, we represent each interval using its midpoint as defined in Section 2.3 (Equation (13)). Similarly, the midpoint of the ground-truth interval is computed following the same definition. Notably, when the ground-truth label is a single value ($y_i^{\min} = y_i^{\max}$), this formulation still applies, and y_i is equal to that value. This ensures that point-based labels are naturally handled within the interval framework. Using the predicted and reference means, we compute the following regression metrics:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}(|\hat{y}_i - y_i|), \tag{18}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{19}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{20}$$

where $\bar{y} = \mathbb{E}(y_i)$ denotes the mean of all ground-truth midpoints. This unified formulation allows direct comparison across single-output and interval-output models, while retaining compatibility with datasets containing both precise and uncertain annotations.

Interval-aware Evaluation. We additionally evaluate how well the predicted mean falls within the true annotated interval. If it lies inside the range, the error is set to zero; otherwise, we penalize the distance to the nearest boundary:

$$\text{Int-Error}_i = \max(0, y_i^{\min} - \hat{y}_i^{min}) + \max(0, \hat{y}_i^{max} - y_i^{\max}), \tag{21}$$

$$\text{MAE}_{\text{int}} = \frac{1}{N}\sum_{i=1}^{N}\text{Int-Error}_i, \tag{22}$$

$$\text{RMSE}_{\text{int}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{Int-Error}_i)^2}. \tag{23}$$

4 Result

4.1 Overall Performance Comparison

Under the interval-aware evaluation setting, where predictions are compared directly against the true interval bounds, DeepPH consistently achieves superior performance. As shown in Table 1, on the EH test set, DeepPH achieves a 15.6% lower MAE and 12.6% lower RMSE compared to the best EpHod baseline (Ensemble). On the VD test set, DeepPH achieves a lower MAE of 0.443 compared to 0.704 from VENUS-DREAM-MAML, and a reduced RMSE of 0.696 versus 0.993, indicating strong generalization across diverse enzyme types.

Table 1: Performance comparison on the EH and VD test sets under interval-aware evaluation.

Model	MAE_{int}	RMSE_{int}	Test Set
EpHod-SVR	0.593	0.853	EH
EpHod-RLAT	0.502	0.749	
EpHod-Ensemble	0.499	0.746	
DeepPH (Ours)	**0.421**	**0.652**	
VENUS-DREAM-MAML	0.704	0.993	VD
VENUS-DREAM-Reptile	4.473	4.798	
DeepPH (Ours)	**0.443**	**0.696**	

Best score in each block is in **bold.*

In the Average-based Evaluation setting—where both predictions and ground truths are treated as midpoint values—DeepPH again delivers state-of-the-art results. As summarized in Table 2, it achieves an MAE of 0.592 and an RMSE of 0.799 on the EH test set, along with an R^2 score of 0.271, which represents an 84.4% relative improvement over the best EpHod variant, EpHod-Ensemble ($R^2 = 0.147$). However, EpHod-RLAT provides more accurate predictions for acidic and alkaline enzymes, with MAEs of 1.032 and 1.927, respectively. On the VD test set, DeepPH consistently outperforms all baselines across MAE, RMSE, and R^2, achieving the lowest prediction errors and highest explained variance. Notably, DeepPH is the only model that achieves a positive R^2 on this test set, whereas both VENUS-DREAM-based baselines yield negative R^2 scores, indicating poor variance explanation and weak generalization. Additionally, DeepPH attains the best MAE values for

acidic and alkaline enzyme subsets, with scores of 1.450 and 2.002, respectively. Overall, DeepPH outperforms existing models.

Table 2: Performance comparison on the EH and VD test sets under average-based evaluation.

Model	MAE	RMSE	R^2	Test Set
EpHod-SVR	0.785	0.987	-0.114	EH
EpHod-RLAT	0.670	0.873	0.129	
EpHod-Ensemble	0.665	0.864	0.147	
DeepPH (Ours)	**0.592**	**0.799**	**0.271**	
VENUS-DREAM-MAML	0.952	1.199	-0.358	VD
VENUS-DREAM-Reptile	4.907	5.152	-24.080	
DeepPH (Ours)	**0.648**	**0.868**	**0.289**	

Best score in each block is in **bold.*

4.2 Subgroup Analysis by Sequence Properties

To further examine the robustness of DeepPH in diverse enzyme characteristics, we performed a stratified evaluation based on sequence length. Specifically, each test set is divided into two subgroups: (1) moderate-length sequences ($[32, 1022]$ residues), representing the typical size range of structured enzymes, and (2) extreme length sequences, including both short (<32 residues) and long (>1022 residues) proteins. The results in Table 3 and Table 4 demonstrate that DeepPH consistently achieves strong performance across both subgroups. On the moderate-length set, it attains the best overall accuracy. In contrast, on the extreme-length set—where sequence properties deviate substantially from typical enzymes—all models, including baseline methods, exhibit noticeable performance degradation. This observation confirms that proteins with very short or very long sequences are more challenging for pH prediction, likely due to reduced contextual information or increased structural variability. These results show the model's adaptability to variable-length inputs and its broad applicability in real-world enzyme analysis, including edge cases that are often underrepresented in curated datasets.

4.3 Attention–Function Correlation

We analyzed the multi-head attention weights across residue properties, amino acid types and structural exposure to study the model's internal focus. As shown in the Fig. 4A, the distribution of attention across different amino acid types exhibits a clear divergence among the four attention heads. Acidic residues receive less overall attention, yet with some degree of differentiation. Head4 shows the highest attention to glutamate (E), followed by Head3 and Head2, while Head1 shows minimal attention to acidic residues. In contrast, basic residues emerge as focal points across all heads, but with distinct preference patterns—Head2 is highly specialized in recognizing arginine (R), Head3 preferentially attends to lysine (K), and Head1 and Head4 exhibit a more balanced distribution between K and R. This diversity in attention suggests that the model has learned to distinguish between the functional roles of basic residues in pH regulation. For polar residues, although the overall attention is moderate to low, a noticeable variation exists among heads. Head3 shows the highest sensitivity to polar residues such as glutamine (Q) and serine (S), followed by Head4, while Head1 and Head2 contribute less. In contrast, non-polar residues receive consistently low attention across all heads with minimal inter-head variation, suggesting a relatively limited contribution to the model's task or the possibility that relevant signals from these residues are captured through other features. Collectively, the observed head-specific attention patterns reflect a degree of biochemical awareness and functional modularity within the model, particularly emphasizing its ability to recognize residue-specific roles in pH-dependent mechanisms.

Fig. 4B presents the mean attention values assigned to four categories of amino acid residues—acidic, basic, polar, and non-polar—further stratified by spatial location (surface and core, defined by a relative solvent-accessible surface area (RSA) threshold of 0.2 [26, 30]) and across the four attention heads. Overall, surface residues consistently receive higher attention scores than core residues, with this trend being particularly prominent for basic and acidic residues, highlighting the importance of spatial structure in the model's attention mechanism.

Among all attention heads, Head4 assigns the highest attention to surface-exposed basic residues. Head3 and Head2 also exhibit clear surface-core disparities in both basic and acidic residue types, whereas Head1 displays a more balanced distribution, assigning relatively similar attention across residue categories and structural locations. These patterns suggest that the model not only demonstrates biochemical selectivity but also demonstrates a clear degree of spatial awareness, preferentially focusing on surface-exposed residues with potential functional relevance, especially those involved in pH-dependent regulation.

To evaluate whether the model's attention mechanism reflects the structural and functional properties of proteins, we conducted a case study on two representative proteins with distinct optimal pH values. Fig. 5 highlights the top 10% attention-scoring residues that are also located on the surface (defined as RSA > 0.2), shown in red. Fig. 5A shows the alkaline enzyme Protein Nef (UniProt ID: O57064), where high-attention residues are widely distributed in solvent-accessible regions, including helices, loops, and parts of β-sheet structures, indicating that the model reaches much attention to surface exposed structural elements. Fig.5B shows the acidic enzyme CelB (UniProt ID: Q97VS7), a putative cellulase from Saccharolobus solfataricus with an optimal activity. In this case, high-attention residues appear more clustered, primarily located along one side of exposed α-helices and β-sheet edges, forming a localized structural hotspot.

In both cases, highly attended residues are located exclusively on the protein surface and are not buried in the structural core. These observations suggest that the model is capable not only of recognizing biochemical preferences in residue types but also of demonstrating spatial awareness by prioritizing surface regions with potential pH-dependent functional relevance.

5 Ablation Study

In order to assess the contribution of each critical component in our proposed model, we conducted a comprehensive ablation study. Specifically, we individually removed the attention mechanism,

Table 3: Performance comparison on the moderate-length sequence subgroup ($32 \leq L \leq 1022$), under average-based and interval-aware evaluation.

Model	MAE	RMSE	R^2	MAE_{int}	$RMSE_{int}$	Test Set
EpHod-SVR	0.783	0.988	-0.073	0.582	0.849	EH
EpHod-RLAT	0.659	0.859	0.189	0.482	0.737	
EpHod-Ensemble	0.656	0.858	0.190	0.485	0.741	
DeepPH (Ours)	**0.599**	**0.800**	**0.296**	**0.418**	**0.651**	
VENUS-DREAM-MAML	0.947	1.196	-0.300	0.689	0.979	VD
VENUS-DREAM-Reptile	4.889	5.136	-22.981	4.421	4.754	
DeepPH (Ours)	**0.662**	**0.882**	**0.293**	**0.448**	**0.709**	

Best score in each block is in **bold.*

Table 4: Performance comparison on extreme-length sequences ($L < 32$ or $L > 1022$), under average-based and interval-aware evaluation.

Model	MAE	RMSE	R^2	MAE_{int}	$RMSE_{int}$	Test Set
EpHod-SVR	0.796	0.983	-0.405	0.648	0.875	EH
EpHod-RLAT	0.726	0.940	-0.287	0.603	0.807	
EpHod-Ensemble	0.706	0.890	-0.151	0.570	0.766	
DeepPH (Ours)	**0.554**	**0.791**	**0.090**	**0.435**	**0.658**	
VENUS-DREAM-MAML	0.984	1.216	-0.818	0.785	1.068	VD
VENUS-DREAM-Reptile	5.008	5.240	-32.811	4.766	5.036	
DeepPH (Ours)	**0.568**	**0.782**	**0.246**	**0.416**	**0.622**	

Best score in each block is in **bold.*

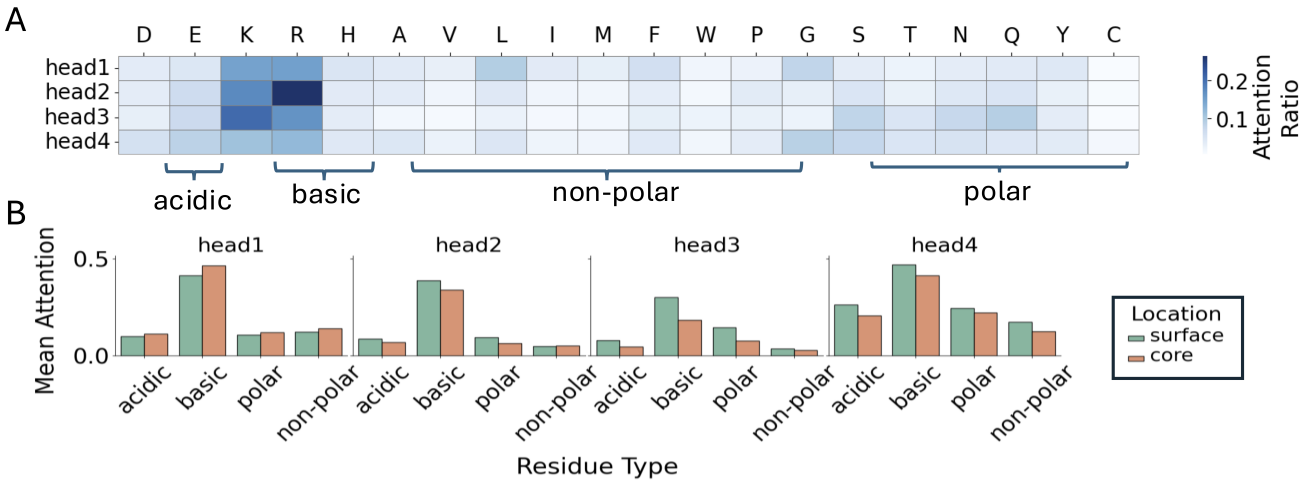

Figure 4: Attention-based analysis of the four attention heads in DeepPH. (A) Mean attention weights for each of the 20 amino acid types across the four attention heads. (B) Attention distribution between surface (RSA > 0.2) and core residues across residue categories and attention heads.

ProtT5 embedding module, and ESMFold structure prediction module to evaluate their effects on model performance. Comparison of the two evaluation settings on the VD test set (803 samples) is provided in Table 5 for average-based evaluation and Table 6 for interval-aware evaluation.

We observed a consistent trend regarding the contribution of each model component. The removal of the attention mechanism consistently led to the most significant performance degradation in both settings, with MAE and RMSE increasing dramatically by approximately 4 to 5 times compared to the full model. Removing

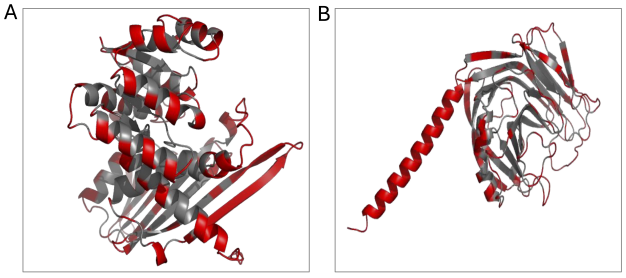

Figure 5: Structural visualization of high-attention surface residues (red) for two representative proteins with different pH preferences. Surface residues are defined as those with RSA > 0.2, and only the top 10% of attention scores are shown. (A) O57064, a protein with an alkaline pH optimum, displays high-attention sites mainly on surface helices and loops. (B) Q97VS7, an acidic pH protein, shows concentrated high-attention residues along exposed helices and β-sheet edges.

Table 5: Ablation study of DeepPH under under Average-based Evaluation.

Model Variant	MAE	RMSE	R^2
w/o Attention	2.714	4.176	-15.480
w/o ProtT5	0.684	0.951	0.145
w/o ESMFold	0.649	0.872	0.282
DeepPH	0.648	0.868	0.289

Table 6: Ablation study of DeepPH under interval-aware evaluation.

Model Variant	MAE_{int}	$RMSE_{int}$
w/o Attention	2.374	3.946
w/o ProtT5	0.469	0.761
w/o ESMFold	0.448	0.698
DeepPH	0.443	0.696

ProtT5 embeddings or ESMFold structural predictions also consistently resulted in moderate but meaningful performance drops. This indicates that both sequence and structural information play complementary roles, and the attention mechanism substantially enhances the integration and predictive capacity of DeepPH in evaluation frameworks.

6 Discussion

In this study, we proposed DeepPH, a framework that predicts the optimal pH range of enzymes using only sequence input, while incorporating 3D structural features derived from the sequence. Unlike existing approaches that often treat pH prediction as a classification or single-point regression task, DeepPH models the problem as a range prediction task, aligning more closely with the uncertainty inherent in experimental annotations. This is achieved through a custom-designed loss function that accommodates both precise and interval-based labels, enabling robust learning across diverse annotation types.

A key advantage of DeepPH lies in its incorporation of 3D structural information through our proposed graph based structural encoder, which infers residue-level spatial context directly from the sequence without requiring experimentally resolved structures. Furthermore, the cross-modal attention pooling mechanism dynamically aligns features of the sequence and structure, enhancing the model's ability to capture localized pH-relevant patterns. In particular, DeepPH maintains stable performance across enzymes with extreme sequence lengths, indicating its generalizability. Ablation studies also demonstrate that removing the attention module results in the most significant performance degradation, underscoring its central role in integrating sequence and structure information.

In addition to precision, DeepPH offers biological interpretability. Attention analysis reveals that the model consistently highlights surface-exposed basic residues known to influence pH activity. This suggests that the model not only performs well in predictive tasks, but also has the potential to uncover mechanistic insights into pH sensitivity.

Despite the strengths of the current framework, some limitations remain. Although many enzymes exhibit activity over a broad pH range, their catalytic efficiency often varies across different pH values. However, such pH-activity profiles are rarely reported in a standardized or machine-readable format, limiting their utility for model training and evaluation. In addition, a small subset of enzymes, referred to as ribozymes, is composed of RNA rather than protein. As our model is explicitly designed to process protein sequences and structures, it is not applicable to these RNA-based catalysts. Accurate prediction of optimal pH for ribozymes would require a different modeling approach that incorporates RNA-specific structural and functional features.

References

[1] Mario Barroca, Gustavo Santos, Björn Johansson, Florian Gillotin, Georges Feller, and Tony Collins. 2017. Deciphering the factors defining the pH-dependence of a commercial glycoside hydrolase family 8 enzyme. *Enzyme and Microbial Technology* 96 (2017), 163–169.

[2] Ana Beloqui, Andrei Yu Kobitski, Gerd Ulrich Nienhaus, and Guillaume Delaittre. 2018. A simple route to highly active single-enzyme nanogels. *Chemical science* 9, 4 (2018), 1006–1013.

[3] Anže Lošdorfer Božič and Rudolf Podgornik. 2017. pH dependence of charge multipole moments in proteins. *Biophysical journal* 113, 7 (2017), 1454–1465.

[4] R Buller, S Lutz, RJ Kazlauskas, R Snajdrova, JC Moore, and UT Bornscheuer. 2023. From nature to industry: Harnessing enzymes for biocatalysis. *Science* 382, 6673 (2023), eadh8615.

[5] Nicki Skafte Detlefsen, Søren Hauberg, and Wouter Boomsma. 2022. Learning meaningful representations of protein sequences. *Nature communications* 13, 1 (2022), 1914.

[6] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. 2021. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* 44, 10 (2021), 7112–7127.

[7] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. 2021. ProtTrans: towards cracking the language of life's code through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2021), 7112–7127.

[8] Manuel Ferrer, Olga Golyshina, Ana Beloqui, and Peter N Golyshin. 2007. Mining enzymes from extreme environments. *Current opinion in microbiology* 10, 3 (2007), 207–214.

[9] Japheth E Gado, Matthew Knotts, Ada Y Shaw, Debora Marks, Nicholas P Gauthier, Chris Sander, and Gregg T Beckham. 2023. Deep learning prediction of enzyme optimum pH. *bioRxiv* (2023), 2023–06.

[10] Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. 2021. Structure-based protein function prediction using graph convolutional networks. *Nature communications* 12, 1 (2021), 3168.

[11] Benjamin Giovanni Iovino and Yuzhen Ye. 2024. Protein embedding based alignment. *BMC bioinformatics* 25, 1 (2024), 85.

[12] Wolfgang Kabsch and Christian Sander. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules* 22, 12 (1983), 2577–2637.

[13] Laura J Kingsley and Markus A Lill. 2015. Substrate tunnels in enzymes: structure–function relationships and computational methodology. *Proteins: Structure, Function, and Bioinformatics* 83, 4 (2015), 599–611.

[14] Anu Kumar, Sunny Dhiman, Bhanu Krishan, Mrinal Samtiya, Ankita Kumari, Nishit Pathak, Archana Kumari, Rotimi E Aluko, and Tejpal Dhewa. 2024. Microbial enzymes and major applications in the food industry: a concise review. *Food Production, Processing and Nutrition* 6, 1 (2024), 85.

[15] David R. Lide. 2005. *"Properties of Amino Acids", in CRC Handbook of Chemistry and Physics, Internet Version.* CRC press.

[16] Hye Jin Lim, Yu Jin Park, Yeon Jae Jang, Ji Eun Choi, Joon Young Oh, Ji Hyun Park, Jae Kwang Song, and Dong-Myung Kim. 2016. Cell-free synthesis of functional phospholipase A1 from Serratia sp. *Biotechnology for Biofuels* 9 (2016), 1–7.

[17] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv* 2022 (2022), 500902.

[18] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 6637 (2023), 1123–1130.

[19] Matteo Manfredi, Gabriele Vazzana, Castrense Savojardo, Pier Luigi Martelli, and Rita Casadio. 2025. AlphaFold2 and ESMFold: A large-scale pairwise model comparison of human enzymes upon Pfam functional annotation. *Computational and Structural Biotechnology Journal* 27 (2025), 461–466.

[20] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. 2021. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems* 34 (2021), 29287–29303.

[21] Milot Mirdita, Martin Steinegger, and Johannes Söding. 2019. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* 35, 16 (2019), 2856–2858.

[22] Erez Persi, Miquel Duran-Frigola, Mehdi Damaghi, William R Roush, Patrick Aloy, John L Cleveland, Robert J Gillies, and Eytan Ruppin. 2018. Systems analysis of intracellular pH vulnerabilities for cancer therapy. *Nature communications* 9, 1 (2018), 2997.

[23] GC Pradeep, Yun Hee Choi, Yoon Seok Choi, Se Eun Suh, Jeong Heon Seong, Seung Sik Cho, Min-Suk Bae, and Jin Cheol Yoo. 2014. An extremely alkaline novel chitinase from Streptomyces sp. CS495. *Process Biochemistry* 49, 2 (2014), 223–229.

[24] Jonas Protze, Fabian Müller, Karin Lauber, Bastian Naß, Reinhard Mentele, Friedrich Lottspeich, and Arnulf Kletzin. 2011. An extracellular tetrathionate hydrolase from the thermoacidophilic archaeon Acidianus ambivalens with an activity optimum at pH 1. *Frontiers in microbiology* 2 (2011), 68.

[25] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* 118, 15 (2021), e2016239118.

[26] Castrense Savojardo, Matteo Manfredi, Pier Luigi Martelli, and Rita Casadio. 2021. Solvent accessibility of residues undergoing pathogenic variations in humans: from protein structures to protein sequences. *Frontiers in molecular biosciences* 7 (2021), 626363.

[27] I Schomburg, L Jeske, M Ulbrich, S Placzek, A Chang, and D Schomburg. 2017. The BRENDA enzyme information system–From a database to an expert system. *Journal of biotechnology* 261 (2017), 194–206.

[28] Lise Schoonen, Sjors Maassen, Roeland JM Nolte, and Jan CM van Hest. 2017. Stabilization of a virus-like particle and its application as a nanoreactor at physiological conditions. *Biomacromolecules* 18, 11 (2017), 3492–3497.

[29] Yidong Song, Qianmu Yuan, Sheng Chen, Yuansong Zeng, Huiying Zhao, and Yuedong Yang. 2024. Accurately predicting enzyme functions through geometric graph learning on ESMFold-predicted structures. *Nature Communications* 15, 1 (2024), 8180.

[30] Matthew Z Tien, Austin G Meyer, Dariya K Sydykova, Stephanie J Spielman, and Claus O Wilke. 2013. Maximum allowed solvent accessibilites of residues in proteins. *PloS one* 8, 11 (2013), e80635.

[31] Karel J van der Weg and Holger Gohlke. 2023. TopEnzyme: A framework and database for structural coverage of the functional enzyme space. *Bioinformatics* 39, 3 (2023), btad116.

[32] Sergey D Varfolomeev, Alexander A Panin, Valeriy I Bykov, Svetlana B Tsybenova, and Alexander G Chuchalin. 2020. Chemical kinetics of the development of coronaviral infection in the human body: Critical conditions, toxicity mechanisms,"thermoheliox", and "thermovaccination". *Chemico-biological interactions* 329 (2020), 109209.

[33] Qianmu Yuan, Chong Tian, Yidong Song, Peihua Ou, Mingming Zhu, Huiying Zhao, and Yuedong Yang. 2024. GPSFun: geometry-aware protein sequence function predictions with language models. *Nucleic Acids Research* 52, W1 (2024), W248–W255.

[34] Liang Zhang, Kuan Luo, Ziyi Zhou, Yuanxi Yu, Fan Jiang, Banghao Wu, Mingchen Li, and Liang Hong. 2025. A Deep Retrieval-Enhanced Meta-Learning Framework for Enzyme Optimum pH Prediction. *Journal of Chemical Information and Modeling* 65, 7 (2025), 3761–3770.