

EnzyCLIP: A Cross-Attention Dual Encoder Framework with Contrastive Learning for Predicting Enzyme Kinetic Constants

Anas Aziz Khan¹, Md. Shah Fahad^{2*}, Priyanka³,
Ramesh Chandra³, Guransh Singh²

¹SCOPE, Vellore Institute of Technology, Vellore, Katpadi, Vellore,
632014, Tamil Nadu, India.

²Department of Computer Science, BIT, Mesra, Ranchi, 835215,
Jharkhand, India.

³Department of Bioengineering and Biotechnology, BIT, Mesra, Ranchi,
835215, Jharkhand, India.

*Corresponding author(s). E-mail(s): fahad8siddiqui@bitmesra.ac.in;
Contributing authors: anasazizkhan555@gmail.com;
reachpriyanka20@gmail.com; rameshchandra@bitmesra.ac.in;
guransh766@gmail.com;

Abstract

Accurate prediction of enzyme kinetic parameters is crucial for drug discovery, metabolic engineering, and synthetic biology applications. Current computational approaches face limitations in capturing complex enzyme–substrate interactions and often focus on single parameters while neglecting the joint prediction of catalytic turnover numbers (K_{cat}) and Michaelis–Menten constants (K_m). We present EnzyCLIP, a novel dual-encoder framework that leverages contrastive learning and cross-attention mechanisms to predict enzyme kinetic parameters from protein sequences and substrate molecular structures. Our approach integrates ESM-2 protein language model embeddings with ChemBERTa chemical representations through a CLIP-inspired architecture enhanced with bidirectional cross-attention for dynamic enzyme–substrate interaction modeling. EnzyCLIP combines InfoNCE contrastive loss with Huber regression loss to learn aligned multimodal representations while predicting \log_{10} -transformed kinetic parameters. EnzyCLIP is trained on the CatPred-DB database containing 23,151 K_{cat} and 41,174 K_m experimentally validated measurements, and achieved competitive baseline performance with R^2 scores of 0.593 for K_{cat} and 0.607 for K_m .

prediction. XGBoost ensemble methods on learned embeddings further improved K_m prediction ($R^2 = 0.61$) while maintaining robust K_{cat} performance.

1 Introduction

Enzymes are catalysts that enable reactions to happen at rapid rates under the mild conditions of temperature, pH, and pressure inside cells. In contrast to inorganic catalysts, which mostly require harsh or extreme conditions to function, enzymes can work powerfully under gentle conditions—namely, near neutral pH, moderate temperatures, and in aqueous environments. This is a reflection of how evolution has precisely honed them to sustain the fragile balance of cellular life[1, 2].

Essentially, enzymes are involved in almost all cellular processes. They perform major metabolic routes, such as glycolysis, the tricarboxylic acid cycle, and oxidative phosphorylation—which jointly supply energy for the cell—apart from metabolism, replication of DNA and repair, transcription, translation of proteins, neurotransmitter breakdown, and cell signaling are some other areas in which enzymes have important roles. Given such a wide array of their functions, there is no doubt that they are very much part of cellular health and the amazing complexity of life[3].

Predictions of protein-ligand binding free energy are difficult to interpret unless reliable estimates of experimental uncertainties in measured binding free energies are provided.

The extraordinary catalytic efficiency of enzymes comes from their capability for lowering the height of the activation energy barrier that slows down chemical transformations[4]. Enzymes, through stabilizing transition states and orienting reactive groups in precise spatial arrangements, make it energetically easier for reactants to convert into products. Usually, each enzyme is highly specific towards certain molecular partners called substrates. This comes from structural complementarity between the enzyme active site and the geometry of the substrate molecule[5].

The active site is a pocket in a protein, comprising only 10% to 20% of the total volume of the enzyme. It is shaped when amino acids from different parts fold together. Binding of the substrate in this site by weak non-covalent interactions such as hydrogen bonds, electrostatic attractions, and van der waals forces develops a temporary enzyme-substrate complex where the chemical reaction occurs. The enzyme will return to its original conformation after the reaction is complete and can begin processing another substrate molecule[6].

With the development of enzyme kinetics, a field that measures and models the rates of enzymatic reactions, scientists are able to evaluate how efficiently enzymes carry out their catalytic functions. Lying at the heart of the study of enzyme kinetics is the Michaelis-Menten model, which describes how the rate of reaction varies with substrate concentration. The model is mathematically expressed as:

$$v = \frac{V_{max} \cdot [S]}{K_m + [S]} \quad (1)$$

where v is the reaction velocity, $[S]$ is the substrate concentration, V_{\max} is the maximum achievable rate of reaction, and K_m is the Michaelis constant[7].

The model helps in determining two key parameters: the turnover number (K_{cat}), representing substrate molecules converted per enzyme molecule per unit time at maximum efficiency, and the Michaelis constant (K_m), denoting substrate concentration at half-maximal reaction rate. The relationship between V_{\max} and K_{cat} is given by:

$$K_{\text{cat}} = \frac{V_{\max}}{[E]_{\text{total}}} \quad (2)$$

where $[E]_{\text{total}}$ is the total enzyme concentration. The catalytic efficiency, expressed as K_{cat}/K_m , serves as a comprehensive metric for enzyme performance under physiological conditions where substrate saturation rarely occurs[8].

Accurate prediction of kinetic parameters such as K_{cat} and K_m helps to advance various fields including drug discovery, metabolic engineering, and synthetic biology. For instance, during drug design, insight into the way different kinds of compounds are being processed by enzymes could inform strategies in the creation of molecules possessing favorable pharmacokinetic properties. In metabolic engineering, knowledge of such parameters enables the fine-tuning of enzyme activity and expression levels with the aim of improving the efficiency of biochemical production pathways[9]. However, determination of these values experimentally is mostly laborious and expensive, and the process cannot easily be scaled to cover the huge diversity of possible enzyme-substrate interactions.

Historically, the computational prediction of enzyme kinetics had been done by rule-based systems built on expert knowledge, QSAR models connecting molecular features to biological activity, and molecular docking for estimating binding strength. However, most of these traditional approaches relied much on hand-crafted features and expert knowledge. Hence, they were difficult to scale up and less effective at handling new or unseen enzyme-substrate pairs.

The emergence of machine learning and deep learning has transformed the prediction of enzyme kinetics. Contemporary models employ natural language processing techniques for biological data, regarding a protein sequence as a sentence comprising amino acids while representing substrates as SMILES strings[10]. Pretrained models learn complex contextual relationships between the elements of the sequence in large-scale training with unlabelled data. After fine-tuning for the kinetic parameters prediction, they often outperform traditional methods not only for their accuracy but also for generalization. Deep learning frameworks can incorporate several data modalities; they combine protein sequence embeddings with molecular graph representations for creating an integrated representation of enzyme-substrate pairs.

Some recent works include DeepEnzyme[11], which integrates Transformer and Graph Convolutional Networks to model both the sequence and 3D structural information in order to better predict K_{cat} , MPEK[12], a multitask model that can perform parallel prediction of K_{cat} and K_m by taking environmental factors into account[13], and CatPred, a unified framework for integrating sequence-attention mechanisms with equivalent graph neural networks to predict kinetic parameters for protein sequences[14]. These computational methods are further accelerating enzyme

discovery and engineering, enabling the high-throughput screening of large datasets and better information for biological system modeling.

In this work, we present a new CLIP-inspired dual-encoder framework, EnzyCLIP, which is designed for enzyme-substrate kinetic predictions. The model focuses on predicting two main kinetic parameters: K_{cat} (turnover number) and K_m (Michaelis constant). EnzyCLIP integrates deep learning with contrastive representation learning to jointly embed protein sequences and small-molecule substrates within a shared latent space, which allows the model to effectively capture and quantify functional relationships between enzymes and their substrates.

We utilize a dual encoder architecture wherein two separate biological modalities, protein sequences and substrate chemical structures, are independently processed. Each modality is fed into a different transformer-based encoder that generates fixed dimensional embeddings capturing the key biological and chemical features, respectively. The model leverages contrastive learning to be trained on associating the correct enzyme-substrate pairs by maximizing the similarity between their embeddings while minimizing the similarity for incorrect or random combinations. In this learning framework, subtle patterns that define substrate specificity and catalytic compatibility can be learned.

Specifically, EnzyCLIP uses ESM-2, a deep protein language model that is trained on hundreds of millions of amino acid sequences to extract rich contextual embeddings from enzymes. ESM-2 captures evolutionarily and structurally meaningful features of proteins within a high-dimensional latent space. These embeddings act as biologically meaningful features that help in downstream functional prediction [15].

In parallel, substrate molecules are encoded by ChemBERTa, which is a BERT-based transformer trained on the SMILES representation of chemical compounds. ChemBERTa learns molecular structure and function in a language-like fashion. It learns to recognize patterns of chemical reactivity and substructure directly from tokenized molecular strings [16]. EnzyCLIP maps enzyme and substrate embeddings to a shared latent space with the help of contrastive learning by utilizing learned projection layers [17]. In the process, it aligns the correct pairs while separating the mismatched pairs. The architecture also integrated bi-directional cross-attention [18] since there are dynamic protein-substrate interactions, and it predicts using Huber Loss [19] \log_{10} -transformed K_{cat}^{\max} and $\log_{10} K_m$ values for better robustness. Unlike traditional approaches, EnzyCLIP learns directly from raw sequences, without relying on handcrafted features, offering scalability and interpretability for downstream tasks such as enzyme engineering and activity screening. This unified framework thus marks a quantum jump forward in AI-driven enzymology.

2 Literature Review

Traditional methods for determining kinetic parameters experimentally are time-consuming, expensive, and not suitable for large-scale studies. Thousands of enzymes remain uncharacterized with the rapid growth of metagenomic data. This has created a strong demand for computational methods that can analyze enzymes and predict their functions on a large scale in a much faster way.

The early computational approaches are highly dependent on manually crafted features and domain expert knowledge. QSAR models utilize predefined molecular descriptors for relating activity to the molecular structure with poor transferability and require extensive feature engineering[20]. Homology modeling combined with molecular docking provided useful mechanistic insights but was computationally intensive and required the availability of accurate structural data[21]. Traditional machine learning methods utilize amino acid composition and physiochemical properties as input features, such as Support Vector Regression (SVR)[22], Random Forests[23], and Decision Trees[24]. These methods generalized poorly for new, uncharacterized enzyme families[25] and remained dependent on considerable domain-specific manual feature design[26]. In general, these classical methods had the following critical limitations: reliance on handcrafted features, use of 3D structural information, and poor scalability across diverse enzyme classes [27]

The development of machine learning applications for enzyme kinetics prediction has been greatly enhanced by curated databases such as BRENDA[28] and SABIO-RK[29]. Deep learning eliminated the need to rely on hand-crafted feature engineering, as it amply captures complex, non-linear relationships involving biological systems. Some protein language models, such as ESM-1b/2[30], learn dense embeddings that encode structural, functional, and evolutionary information via self-supervised training. Chemical language models, such as ChemBERTa[16], MolFormer[31], and GROVER[32], offer state-of-the-art molecular representation learned from SMILES strings directly. All of these collectively offer the possibility of end-to-end learning wherein a model will be able to discover meaningful features automatically, generate context-aware representations, and make full use of transfer learning across diverse biochemical tasks. More specialized frameworks include DLKcat[33], CatPred[14], and MPEK[12], which achieved marked improvement compared to the traditional methods. Graph Convolutional Networks enhance prediction accuracy by integrating sequence-derived insights about evolution with structure-based geometric features[34].

CLIP-inspired dual encoder frameworks learn shared representations across biological modalities by separately embedding enzymes and substrates and aligning them through contrastive learning objectives[35]. The contrastive InfoNCE loss function drives alignment:

$$L_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(e_i, s_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(e_i, s_j)/\tau)} \quad (3)$$

where $\text{sim}(e_i, s_i)$ indicates the similarity between enzyme embedding e_i and substrate embedding s_i , τ is the temperature parameter, and N is the batch size. This will allow the model to capture biochemical relationships such as substrate preferences and cofactor requirements that it can use to make predictions even for enzyme–substrate pairs it has never seen [36].

Cross-attention mechanisms allow for a two-way interaction between enzyme and substrate representations, modelling biological phenomena such as induced fit and allosteric regulation. Multi-task learning frameworks jointly predict both parameters:

$$L_{\text{total}} = \alpha L_{\text{contrastive}} + \beta L_{K_{\text{cat}}} + \gamma L_{K_m} \quad (4)$$

where α , β , and γ control relative importance of each loss component[37]. In contrast to the traditional supervised methods, contrastive frameworks generalise better, are more data efficient and robust.

The current computational methods have a number of key challenges strongly limiting practical usability. Treating enzyme and substrate information in a separated way ignores mutual dependencies and dynamic interactions defining enzymatic catalysis [38]. In another important methodological gap, the prevailing focus is on K_{cat} , while K_m is usually neglected, though both parameters emerge from the same catalytic dynamics. A number of the existing models behave like black boxes, featuring limited biological interpretability and controllable molecular factors driving catalytic efficiency. This issue restricts their usability for enzyme engineering. Finally, poor generalization to underrepresented families introduces major problems in the case of metagenomic enzyme discovery [39]. Enzymatic catalysis is inherently a dynamic process. Substrate binding induces changes in enzyme shape, such as an induced fit mechanism, allosteric effects, or specificity depending on the molecular context. However, such complex behaviors can hardly be modeled when enzyme and substrate information are processed separately. Both K_{cat} and K_m depend on interdependent factors involving binding strength, structural events, chemical transformations, and product release. Such challenges provide rationales for further development that will allow the integration of contrastive learning with cross-attention mechanisms into models capable of jointly predicting kinetic parameters and delivering much clearer and more interpretable insight into how enzymes work.

3 Methodology

This section details the experimental setting and technical framework of EnzyCLIP. We take the CatPed-DB[14] dataset containing 23,197 entries for K_{cat} prediction and 41,174 entries for K_m prediction. Our proposed method combines careful dataset curation with state-of-the-art molecular representation learning, contrastive learning strategy, and deep neural network architecture. Enzyme amino acid sequences are represented using ESM-2[30] transformer model, while substrate molecules are encoded using ChemBERTa[16], thus allowing deep and biologically meaningful representation of both modalities. After processing, embeddings are fed into a CLIP-inspired dual encoder enhanced by a bi-directional cross-attention mechanism. The loss function combines InfoNCE contrastive loss[40] to enforce semantic closeness between biochemically compatible enzyme-substrate pairs and incorporates Smooth L1 loss[19] for quantitative prediction of kinetic parameters.

We implemented dataset preprocessing using `pandas` DataFrame operations[41] and applied quality control measures in a systematic way. Regarding the K_m estimation, we first loaded from the CSV file and then filtered it to retain only entries where information in all the three columns `sequence`, `substrate_smiles`, and `log10km_mean` was not missing. We removed any rows with NaN values or whitespace-only strings. To predict K_{cat} , we followed the same process and filtered the dataset to include only records where valid values were present in the columns `sequence`, `substrate_smiles`,

and $\log_{10}k_{\text{cat}}$. We excluded any incomplete or missing entries to guarantee the consistency and reliability of the data in both kinetic parameter prediction tasks.

$$\text{Filtered Dataset} = \{(s_i, m_i, y_i) : s_i \neq \emptyset, m_i \neq \emptyset, y_i \neq \text{NaN}\} \quad (5)$$

where s_i , m_i , and y_i denote the protein sequence, substrate SMILES[42] string, and the logarithmic kinetic parameter of entry i , respectively. The curated dataset was subjected to systematic partition into training, validation, and test sets in an 80:10:10 ratio with a fixed random seed of 42 using the PyTorch `random_split` function for reproducibility (Table 1).

$$\begin{aligned} \text{train_size} &= [0.8 \times \text{total_size}] \\ \text{val_size} &= [0.1 \times \text{total_size}] \\ \text{test_size} &= \text{total_size} - \text{train_size} - \text{val_size} \end{aligned} \quad (6)$$

Table 1: Dataset composition after preprocessing and splitting procedures

Parameter	Training Set (80%)	Validation Set (10%)	Test Set (10%)	Total
K_m Dataset Entries	Variable ¹	Variable ¹	Variable ¹	Filtered Total
Split Ratio	0.8	0.1	0.1	1.0
Random Seed	42	42	42	42
Batch Size	64	64	64	64

¹Exact numbers depend on filtering results and are computed dynamically during preprocessing.

Note: The K_{cat} dataset follows identical preprocessing procedures with 23,197 total entries, while K_m dataset contains 41,174 entries from CatPred-DB.

The molecular representation strategy employed pre-trained language models specifically designed for biological and chemical sequences. Enzyme sequences were processed using the ESM-2[30] transformer architecture, specifically the `facebook/esm2_t6_8M_UR50D` variant with 8 million parameters. This model learns evolutionary, structural, and functional patterns through extensive pretraining on multiple sequence alignment databases. Substrate molecules were encoded using the ChemBERTa[16] `seyonec/ChemBERTa-zinc-base-v1` model, pretrained on the ZINC database. Both tokenizers utilized maximum sequence length of 256 tokens with `max_length` padding and truncation enabled.

$$\mathbf{T}_{\text{prot}} = \text{ESM2Tokenizer}(s; \text{ max length} = 256) \quad (7)$$

$$\mathbf{T}_{\text{chem}} = \text{ChemBERTaTokenizer}(m; \text{ max length} = 256) \quad (8)$$

where, s denotes the protein sequence and m denotes the substrate SMILES string. We implemented the class `EnzyCLIPDataset` as a subclass of PyTorch’s Dataset that handled the complete preprocessing and tokenization pipeline. It tokenized on-the-fly in the `__getitem__` method for memory efficiency; thus, it returned `input_ids` and `attention_mask` tensors for both protein and chemical modalities. Finally, we

created for all the sets `DataLoader` objects with batch size 128;, shuffling was enabled only for training sets.

The core neural architecture integrates contrastive learning principles with cross-attention mechanisms, together with regression prediction capabilities. The architecture makes use of frozen pre-trained encoders for both protein and chemical modalities, maintaining valuable representations while concentrating optimization on task-specific components. Each encoder is followed by projection heads implemented as `nn.Sequential` modules mapping outputs to a common 256-dimensional embedding space.

$$\mathbf{P}_{\text{prot}} = \text{LayerNorm}(\mathbf{W}_p \mathbf{h}_{\text{prot}} + \mathbf{b}_p) \quad (9)$$

$$\mathbf{P}_{\text{chem}} = \text{LayerNorm}(\mathbf{W}_c \mathbf{h}_{\text{chem}} + \mathbf{b}_c) \quad (10)$$

where, \mathbf{h}_{prot} and \mathbf{h}_{chem} denote the protein and chemical encoder outputs, respectively. This architecture integrates bi-directional cross-attention mechanisms implemented through `nn.MultiheadAttention` with four attention heads that enable explicit modeling of conditional dependencies between enzyme and substrate representations. Such a design avoids the limitation of simple concatenation, because it allows the model to selectively focus on the most relevant features across both modalities, hence leading to richer and more context-aware representations.

$$\mathbf{F}_{\text{prot}} = \text{CrossAttention}(\mathbf{P}_{\text{prot}}^{\text{unsq}}, \mathbf{P}_{\text{chem}}^{\text{unsq}}, \mathbf{P}_{\text{chem}}^{\text{unsq}}) \quad (11)$$

$$\mathbf{F}_{\text{chem}} = \text{CrossAttention}(\mathbf{P}_{\text{chem}}^{\text{unsq}}, \mathbf{P}_{\text{prot}}^{\text{unsq}}, \mathbf{P}_{\text{prot}}^{\text{unsq}}) \quad (12)$$

where, $\mathbf{P}_{\text{prot}}^{\text{unsq}}$ and $\mathbf{P}_{\text{chem}}^{\text{unsq}}$ represent the unsqueezed projected embeddings used as query, key, and value inputs in the cross-attention mechanism. The regression head converts the concatenated cross-attention outputs using a three-layer feedforward network with dimensions of $256 \rightarrow 128 \rightarrow 1$. It uses GELU activations to introduce non-linearity and applies dropout regularization at rates of 0.2 and 0.1 to avoid overfitting (Table 2).

$$\mathbf{h}_{\text{concat}} = \text{Concat}[\mathbf{F}_{\text{prot}}, \mathbf{F}_{\text{chem}}] \quad (13)$$

$$\hat{y} = \mathbf{W}_3(\text{GELU}(\text{Dropout}(\mathbf{W}_2(\text{GELU}(\text{Dropout}(\mathbf{W}_1 \mathbf{h}_{\text{concat}} + \mathbf{b}_1)) + \mathbf{b}_2))) + \mathbf{b}_3) \quad (14)$$

Here, $\mathbf{h}_{\text{concat}}$ denotes the fused embedding, \hat{y} is the predicted kinetic parameter, and $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ are the bias vectors in the regression layers.

The training framework follows a multi-objective optimization strategy that jointly balances contrastive representation learning with supervised learning via supervised regression. The contrastive learning component utilizes a symmetric InfoNCE loss[40], which assumes batch-wise correspondence between modalities, where each index corresponds to an experimentally validated enzyme-substrate pair. For computing the similarity we first perform L2 normalization of the projected embeddings followed by matrix multiplication scaled by a learnable temperature parameter.

Table 2: Neural architecture specifications as implemented in the Model

Component	Specification
Protein Encoder	facebook/esm2_t6_8M_UR50D (frozen)
Chemical Encoder	seyonec/ChemBERTa-zinc-base-v1 (frozen)
Shared Embedding Dimension	256
Cross-Attention Heads	4
Regression Layer 1	256 → 128
Regression Layer 2	128 → 1
Activation Function	GELU
Dropout Rate (Layer 1)	0.2
Dropout Rate (Layer 2)	0.1
Temperature Parameter (Initial)	0.07 (learnable)
Normalization	Layer Normalization

$$\mathbf{L}_{\text{sim}} = \frac{\mathbf{Z}_{\text{prot}} \mathbf{Z}_{\text{chem}}^T}{\exp(\tau)} \quad (15)$$

This symmetric InfoNCE loss calculates cross-entropy on both the similarity matrix and its transpose, ensuring that the contrastive objectives are bidirectional. Here, \mathbf{Z}_{prot} and \mathbf{Z}_{chem} represent the ℓ_2 -normalized protein and chemical embeddings, respectively, and τ represents the learnable temperature parameter.

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{2} [\mathcal{L}_{\text{CE}}(\mathbf{L}_{\text{sim}}, \mathbf{y}) + \mathcal{L}_{\text{CE}}(\mathbf{L}_{\text{sim}}^T, \mathbf{y})] \quad (16)$$

The regression loss uses `nn.SmoothL1Loss` (Huber loss)[19] for robustness to experimental outliers, combining quadratic penalties for small errors with linear penalties for large deviations. Here, \mathbf{y} refers to the ground-truth index targets for contrastive matching across protein–chemical pairs.

$$\mathcal{L}_{\text{SmoothL1}}(\hat{y}, y) = \begin{cases} 0.5(\hat{y} - y)^2 & \text{if } |\hat{y} - y| < 1 \\ |\hat{y} - y| - 0.5 & \text{otherwise} \end{cases} \quad (17)$$

The total training loss combines both objectives in an equal weighting manner to ensure the balanced optimization of representation learning and predictive accuracy.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{InfoNCE}} + \mathcal{L}_{\text{SmoothL1}} \quad (18)$$

We trained the model by using AdamW[43] as an optimizer with an initial learning rate of 2×10^{-4} and weight decay of 1×10^{-2} . The learning rate was adjusted during training by using a cosine annealing schedule ($T_{\text{max}} = 20$) over 25 epochs. To improve computational efficiency and stabilize the training, we enabled Automatic Mixed Precision (AMP) and performed gradient clipping with a maximum norm of 1.0 to avoid gradient explosion. By default, model checkpoints are saved according to the best validation R^2 score.

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos \left(\frac{t}{T_{\max}} \pi \right) \right) \quad (19)$$

Advanced training techniques included Automatic Mixed Precision (AMP) using `torch.amp.autocast` and `GradScaler`, which allows half-precision computation for efficiency. Gradient clipping with maximum norm 1.0 to avoid gradient explosion[44].

$$\mathbf{g}_{\text{clipped}} = \mathbf{g} \cdot \min \left(1, \frac{1.0}{\|\mathbf{g}\|_2} \right) \quad (20)$$

Comprehensive regression metrics were used for model evaluation, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), and Pearson's correlation coefficient (r). Besides, residual-based statistics like mean and standard deviation of residuals were calculated to further check the model for calibration and bias. These metrics collectively quantify both predictive accuracy and consistency across validation and test sets.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (21)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (22)$$

$$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}} \quad (23)$$

where, y_i and \hat{y}_i denote the ground-truth and predicted kinetic parameters, \bar{y} and $\bar{\hat{y}}$ are their respective means, and n is the number of samples.

In our evaluation, we extracted both predictions and immediate features embeddings from the projection layers across all dataset splits. Each sample generated protein- and substrate-derived latent representations, denoted as \mathbf{e}_{prot} and \mathbf{e}_{chem} , respectively. We avoided the typical CLS-token summarization and instead used mean pooling on token embeddings, weighted by attention masks, to get fixed-length representations for each modality. This approach ensured that only contextually relevant tokens contribute proportionally to the final molecular embeddings, therefore enhancing the robustness and expensiveness of the learned representation space.

After model convergence, multimodal embeddings were extracted from the projection layers for all dataset splits to enable downstream regression analysis. Each sample produced both enzyme-derived and substrate-derived latent representations, denoted \mathbf{e}_{prot} and \mathbf{e}_{chem} , respectively. We applied mean pooling over token-level hidden states, which is weighted by attention masks in order to generate fixed-length contextual representations. This ensured that only valid tokens contribute proportionally to the final embedding:

$$\mathbf{e} = \frac{\sum_{t=1}^L h_t \cdot a_t}{\sum_{t=1}^L a_t} \quad (24)$$

where, h_t denotes the hidden state at position t , a_t is the corresponding attention-mask weight, and L is the sequence length. The resulting mean-pooled vectors were

then passed through modality-specific projection layers to obtain compact latent representations in a shared 256-dimensional embedding space:

$$\mathbf{e}_{\text{prot}} = \text{Proj}_{\text{prot}}(\mathbf{e}_{\text{mean,prot}}), \quad \mathbf{e}_{\text{chem}} = \text{Proj}_{\text{chem}}(\mathbf{e}_{\text{mean,chem}}) \quad (25)$$

The projected representations were then concatenated to form a unified embedding vector that jointly captures the enzyme–substrate interaction context:

$$\mathbf{e}_{\text{combined}} = \text{Concat}[\mathbf{e}_{\text{prot}}, \mathbf{e}_{\text{chem}}] \quad (26)$$

The extracted representations were subsequently used as high-level molecular embeddings in downstream regression through ensemble-based learning frameworks. In this work, two gradient-boosting regressors, XGBoost[45] and CatBoost[46], were fitted on the extracted features to model the nonlinear mapping between latent biochemical representations and kinetic parameters. The XGBoost was set with 500 estimators, maximum depth of 10, learning rate of 0.05, and subsampling ratio of 0.8, while CatBoost regressor utilized 500 iterations with similar depth and learning rate, optimized with RMSE loss. Both regressors were evaluated using the same training-validation-test splits of the EnzyCLIP model for consistent benchmarking across the different prediction pipelines.

$$\hat{y}_{\text{XGB}} = \sum_{k=1}^{500} f_k(\mathbf{e}_{\text{combined}}), \quad \hat{y}_{\text{CB}} = \sum_{k=1}^{500} g_k(\mathbf{e}_{\text{combined}}) \quad (27)$$

This embedding-based ensemble regression step assessed the generalizability and expressiveness of the representations learned by EnzyCLIP. The methodology thus integrates deep contrastive feature extraction with classical ensemble regression techniques to provide a comprehensive and interpretable modeling framework for enzyme–substrate kinetic parameter prediction.

4 Results

EnzyCLIP learns robust, generalized multimodal representations across the kinetic parameters K_{cat} and K_m on diverse enzyme–substrate pairs. The model exhibits competitive test performances for K_{cat} ($R^2 = 0.60$) and K_m ($R^2 = 0.61$), outperforming traditional baselines such as Random Forest, XGBoost, SVR, and CatBoost when trained directly on raw inputs. Ensemble regressors trained on EnzyCLIP embeddings further improve prediction accuracy, achieving up to $R^2 \approx 0.61$ for K_m , thereby validating the quality of the learned feature space. Stratified analyses show that while K_m predictions remain stable across a wide range of sequence lengths, accuracy in K_{cat} predictions degrades rapidly for very long proteins (> 800 aa), underscoring distinct biochemical determinants of catalysis versus binding affinity. EC-class-wise comparisons also highlight variation in predictive difficulty aligned with enzymatic function. Explainability analyses demonstrate that different embedding dimensions from EnzyCLIP govern K_{cat} versus K_m , confirming that the model captures meaningful mechanistic differences between catalytic efficiency and substrate affinity. Ablation studies reveal that multimodal integration is critical, with the removal of either protein or substrate information reducing R^2 by an average of 30–50%. Overall, these

results position EnzyCLIP as a strong, generalizable, and mechanistically informative framework for enzyme kinetics prediction.

4.1 Enzyme Turnover Number (K_{cat}) Prediction

4.1.1 Dataset Characteristics and Distribution

The K_{cat} prediction dataset was comprised of 23,151 enzyme-substrate pairs representing 7,177 unique enzymes and 10,853 unique substrates (Fig. 1, Fig. 2). The enzyme sequences had an average length of 430.5 ± 221.6 amino acids with a median of 377 amino acids, and substrate SMILES strings averaged 92.8 ± 69.7 characters (Fig. 1). The K_{cat} values spanned a 12-order-of-magnitude range from -6.00 to 6.00 (\log_{10} scale), with a mean of 0.96 and standard deviation of 1.67 (Fig. 1). There was a weak but statistically significant correlation ($r = 0.027$, $p = 3.70 \times 10^{-5}$) between enzyme sequence length and K_{cat} values, suggesting that sequence length has little impact on catalytic turnover (Fig. 1).

Sample distribution across major Enzyme Commission (EC) classes showed the following representation: with EC 3 (hydrolases) constituting the largest proportion at 8,065 samples, followed by EC 1 (oxidoreductases) with 7,796 samples, EC 2 (transfases) with 4,155 samples, EC 4 (lyases) with 1,620 samples, EC 5 (isomerases) with 928 samples, and finally EC 6 (ligases) with 590 samples (Fig. 1). This distribution also reflects the natural prevalence of these enzyme classes in CatPred-DB.

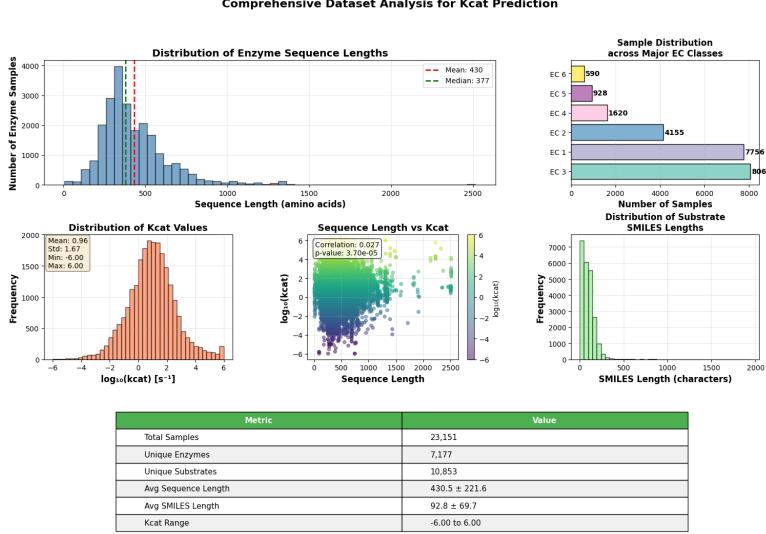


Fig. 1: Comprehensive K_{cat} dataset analysis. Distribution of enzyme sequence lengths showing mean of 430 and median of 377 amino acids (top left), sample distribution across EC classes with EC 3 most represented (top right), K_{cat} value distribution with mean 0.96 and range -6 to 6 (bottom left), weak correlation between sequence length and K_{cat} ($r = 0.027$, bottom center), and substrate SMILES length distribution (bottom right). Summary statistics table shows 23,151 total samples from 7,177 unique enzymes.

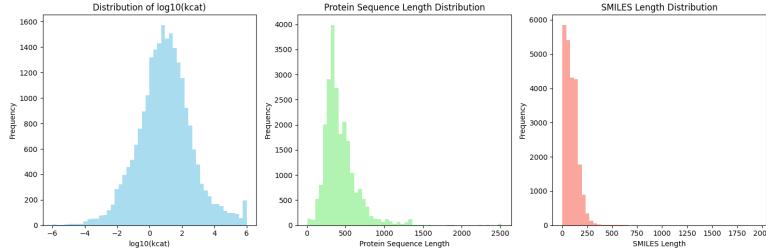


Fig. 2: Dataset distribution comparisons for K_{cat} . Three-panel visualization showing (left) $\log_{10}(K_{\text{cat}})$ distribution with normal characteristics, (center) protein sequence length distribution peaking at 300–500 amino acids, and (right) SMILES length distribution heavily left-skewed with most substrates below 100 characters.

4.1.2 Model Training and Performance

The K_{cat} EnzyCLIP model was optimized for 25 epochs with early stopping according to validation R^2 scores. Training loss decreased monotonically from 0.78 at epoch 1 to

0.16 at epoch 25, indicating proper optimization. Validation performance peaked at epoch 10 with an R^2 of 0.5729, RMSE of 1.071, and MAE of 0.737. The final model at epoch 25 reached validation metrics of $R^2 = 0.564$, RMSE = 1.084, and MAE = 0.740; thus, performance was quite stable with minimal overfitting (Fig. 3).

A summary of the performances of the EnzyCLIP model on this independent test set showed an R^2 of 0.593, MAE of 0.731, and a Pearson correlation coefficient of 0.771. The scatter plot of prediction versus actual indicated a strong alignment along the diagonal with systematic residual patterns centered around zero, therefore showing unbiased predictions across the K_{cat} range. Error distribution analysis also followed a normal distribution centered at zero and had a peak frequency of around 550 samples, hence further proving that the model is reliable (Fig. 8).

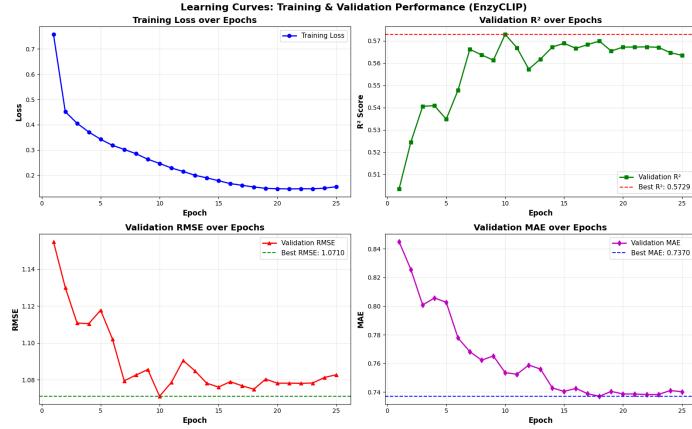


Fig. 3: Training dynamics for K_{cat} prediction. (Top left) Training loss decreasing from 0.78 to 0.16 over 25 epochs. (Top right) Validation R^2 peaking at 0.5729 at epoch 10. (Bottom left) Validation RMSE reaching minimum of 1.071. (Bottom right) Validation MAE achieving best value of 0.737 at epoch 10.

4.1.3 Performance Across Sequence Length Ranges

K_{cat} prediction accuracy was highly variable as a function of enzyme sequence length ranges. For the shortest sequences, 0–200 amino acids, R^2 was 0.649 with RMSE of 1.132 while sequences between 200–400 reached $R^2 = 0.641$ and RMSE = 1.042. The 400–600 amino acid range saw worse performance $R^2 = 0.509$, RMSE = 1.147, followed by partial recovery in the 600–800 range $R^2 = 0.597$, RMSE = 0.987. Longest sequences, greater than 1000 amino acids, had the lowest predictive accuracy $R^2 = 0.441$, RMSE = 1.103 while the adjacent range of 800–1000 showed poor performance $R^2 = 0.354$, RMSE = 1.192. Such trends indicate that model performance is worse for very long enzyme sequences, potentially because proteins become more conformationally complex and/or fewer samples exist in the training sets (Fig. 4).

Mean K_{cat} values across sequence length bins were remarkably consistent between true and predicted values, with both hovering around $1.0\text{--}1.2 \log_{10}(K_{\text{cat}})$ for most

length ranges . However, the range of 900–1000 amino acids showed a conspicuous deviation, where the predicted K_{cat} values (1.22) were higher than the true values (1.04), indicating a possibility of systematic overestimation in that particular length category (Fig. 4).

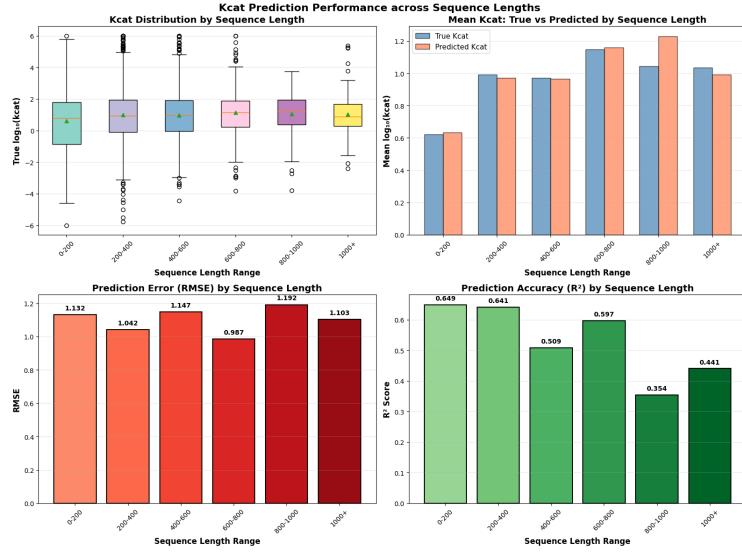


Fig. 4: EnzyCLIP K_{cat} prediction performance across sequence lengths. (Top left) Box plots showing K_{cat} distribution across six sequence length bins with medians around 1.0. (Top right) Mean K_{cat} comparison between true and predicted values across length ranges. (Bottom left) RMSE values ranging from 0.987 to 1.192 across length bins. (Bottom right) R^2 scores showing optimal performance for 0–200 (0.649) and 200–400 (0.641) ranges, with degraded performance for 800–1000 (0.354) and 1000+ (0.441) amino acids.

4.1.4 EC Class-Specific Performance

A more detailed, EC class-specific analysis for k_{cat} prediction revealed significant variability across classes. Accordingly, EC 5 (isomerases) yielded the best performance: R^2 of 0.652, RMSE of 1.158, and Pearson correlation of 0.818. EC 1 (oxidoreductases) were the next best performing class, with $R^2 = 0.645$, RMSE of 1.017 (the lowest among all classes), and a Pearson correlation of 0.804. EC 4 (lyases) yielded a moderate performance, while EC 6 (ligases) classes performed relatively poor: their R^2 values equaled 0.604 and 0.487 respectively. EC 2 (transferases) and EC 3 (hydrolases) had lower predictive accuracy, with their R^2 equal to 0.561 and 0.527 correspondingly (Fig. 5). The broader variation in performance across EC classes reflects the diversity of the catalytic mechanisms underlying k_{cat} —these being more challenging for modeling than the more uniform substrate-binding behavior driving K_m .

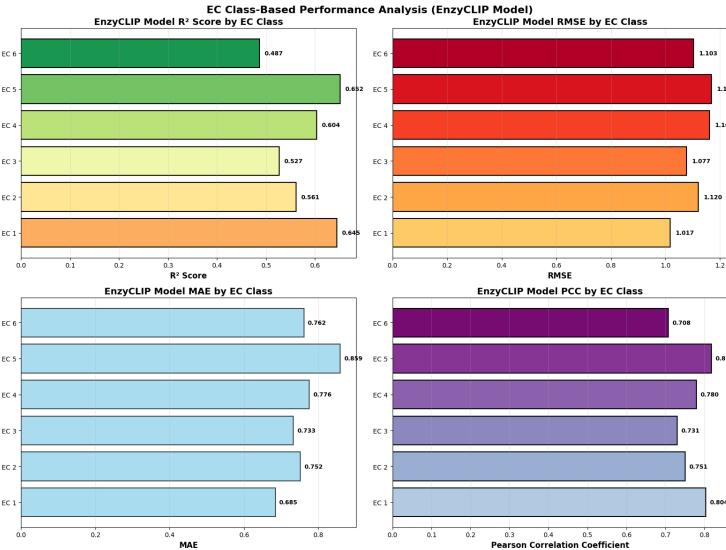


Fig. 5: EC class-based performance analysis for K_{cat} prediction. (Top left) R^2 scores by EC class with EC 5 highest at 0.652. (Top right) RMSE by EC class ranging from 1.017 to 1.161. (Bottom left) MAE by EC class spanning 0.685 to 0.857. (Bottom right) Pearson correlation coefficients showing EC 5 highest at 0.818.

4.1.5 Embedding Space Analysis and Feature Importance

PCA, t-SNE, and UMAP dimensionality reduction of learned embeddings showed continuous distribution patterns colored by K_{cat} values. In the embedding space, there were smooth gradients, not discrete clusters, meaning that the model learned a continuous representation of catalytic efficiency rather than categorical functional groupings (Fig. 6). This continuous embedding structure is in line with the biological reality of K_{cat} values spanning a continuous spectrum of catalytic efficiencies.

Feature importance analysis by ablation studies showed that the full model $R^2 = 0.6$ significantly outperforms the ablated variants. Removing protein embeddings resulted in $R^2 = 0.29$, substrate-only models achieved $R^2 = 0.49$, and sequence-only models achieved $R^2 = 0.49$. Models without substrate information maintained $R^2 = 0.49$, which shows the importance of multimodal integration (Fig. 7). SHAP found the top 20 most impactful embedding dimensions, with feature importances ranging from approximately 0.06 for the most important features to 0.01 for the lower-ranked features (Fig. 8). This suggests both protein sequence and substrate chemical structure are critical to accurate K_{cat} prediction, as expected from basic mechanistic understanding that catalytic efficiency depends explicitly on enzyme-substrate complementarity.

We further stratified performance by magnitude of K_{cat} and observed differential prediction accuracy across activity ranges. For low K_{cat} values (< -2), the MAE was 1.8; for moderate values (-2 to 2), it was 0.69, while for high K_{cat} values (> 2), the

MAE was 1.0. This trend is indicative of a model that performs best for enzymes with intermediate catalytic efficiencies where the training data is most abundant (Fig. 8).

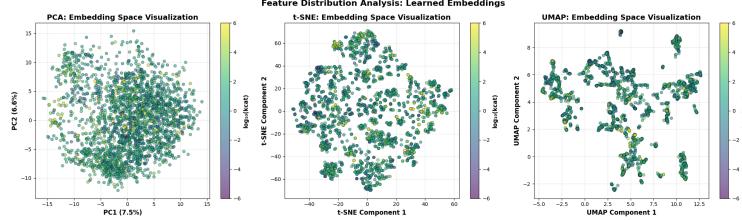


Fig. 6: Learned embedding space visualization for K_{cat} prediction. PCA (left), t-SNE (center), and UMAP (right) projections of enzyme-substrate embeddings colored by K_{cat} values showing continuous gradients rather than discrete clusters, indicating learned representation of catalytic efficiency as a continuous property.

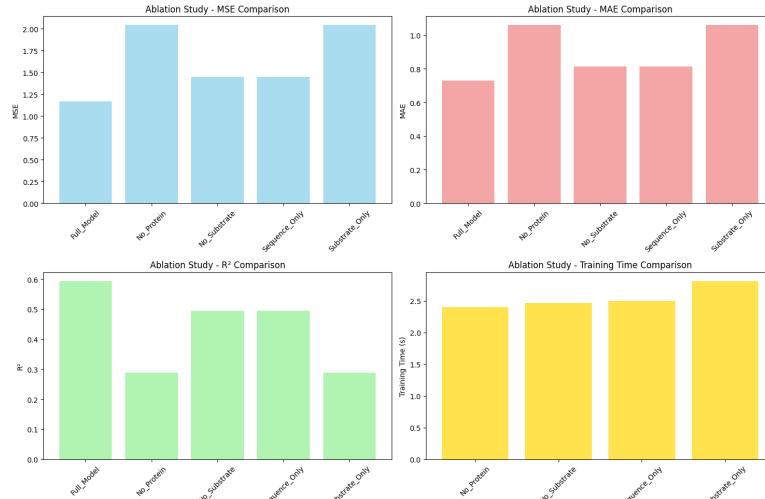


Fig. 7: Ablation study results for K_{cat} prediction. Four-panel comparison showing MSE, MAE, R^2 , and training time across ablation conditions. Full model achieves $R^2 = 0.6$, with substantial performance drops when removing protein ($R^2 = 0.29$), substrate ($R^2 = 0.49$), or sequence ($R^2 = 0.49$) information.

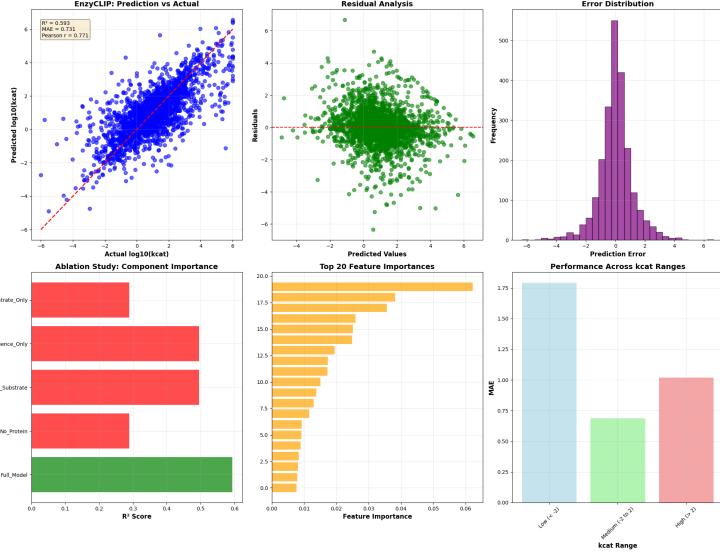


Fig. 8: EnzyCLIP K_{cat} prediction evaluation and explainability. (Top left) Prediction vs. actual scatter plot showing $R^2 = 0.593$, MAE=0.731, Pearson $r = 0.771$. (Top center) Residual analysis showing symmetric distribution around zero. (Top right) Error distribution histogram with normal characteristics. (Bottom left) Ablation study showing full model $R^2 = 0.6$ substantially outperforms component-only models. (Bottom center) SHAP feature importance for top 20 embedding dimensions. (Bottom right) Performance across K_{cat} ranges showing MAE of 1.8, 0.69, and 1.0 for low, medium, and high K_{cat} values.

4.1.6 Comparison with Baseline Methods and Ensemble Models

On the held-out test set, Random Forest achieved an R^2 of 0.222, followed by XGBoost at 0.221, and SVR at 0.147.9. However EnzyCLIP without any extra regressor tree attached on the embedding achieves an R^2 of 0.59.

Ensemble modeling with XGBoost trained on EnzyCLIP embeddings yielded further performance improvements. The XGBoost regressor achieved $R^2 = 0.61$ on the test set, slightly outperforming CatBoost ($R^2 = 0.56$) (Fig. 10). Comparison of ensemble approaches showed that Random Forest (MSE = 2.107, RMSE = 1.452, $R^2 = 0.226$), XGBoost (MSE = 2.121, RMSE = 1.456, $R^2 = 0.221$), and Support Vector Regression (MSE = 2.324, RMSE = 1.524, $R^2 = 0.147$) exhibited varying degrees of success (Fig. 9). The XGBoost ensemble demonstrated superior performance with lower error metrics and higher explained variance 10.

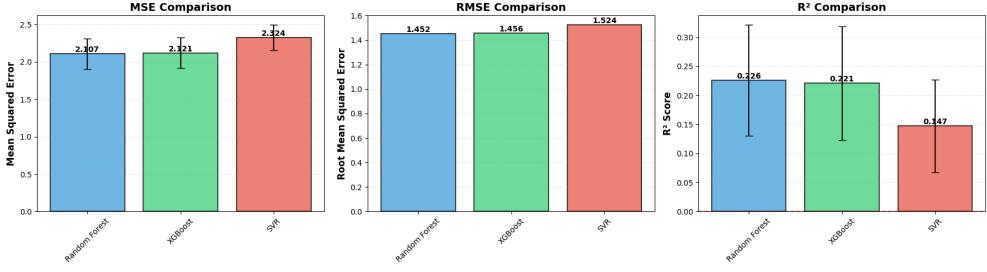


Fig. 9: Baseline methods for K_{cat} prediction. Bar chart comparing R^2 scores on the held-out test set: Random Forest (0.226), XGBoost (0.221), and SVR (0.147).

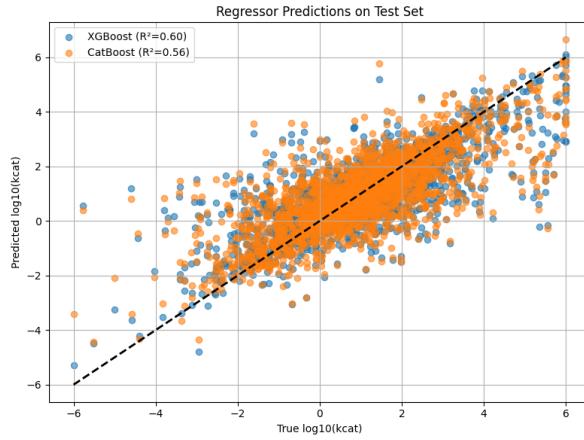


Fig. 10: XGBoost and CatBoost regressor comparison for K_{cat} prediction based on EnzyCLIP embeddings. The scatter plot overlays XGBoost ($R^2 = 0.60$) and CatBoost ($R^2 = 0.56$) predictions against true K_{cat} values, showing XGBoost's superior performance.

4.2 Michaelis-Menten Constant (K_m) Prediction

4.2.1 Dataset Characteristics and Distribution

The dataset for the K_m prediction consisted of 41,174 enzyme-substrate pairs, including 12,355 unique enzymes and 10,535 unique substrates. This dataset is much larger compared to K_{cat} , allowing higher statistical power in model training. The average enzyme sequence length is 437.5 (± 228.9) amino acids (median of 389 amino acids); the average length of the substrate SMILES strings is 57.0 (± 56.4) characters long. The values of K_m range from -8.00 to 4.00 (\log_{10} scale), with a mean and standard deviation of -0.73 and 1.27, respectively, reflecting typical micromolar to millimolar binding affinities of enzyme-substrate interactions.

The length of the sequence is uncorrelated with K_m values, $r = -0.009$, $p = 6.41 \times 10^{-2}$, indicating that binding affinity is not related to the size of a protein in general, rather a function of active site architecture. Sample distribution across EC classes demonstrated that this data contains 13,931 samples from EC 1 (oxidoreductases), 11,025 samples from EC 3 (hydrolases), 10,182 samples from EC 2 (transferases), 2,962 samples from EC 4 (lyases), 1,363 samples from EC 6 (ligases) and finally 1,311 samples from EC 5 (isomerases). This pattern of distribution is unlike the K_{cat} dataset, indicating that kinetic measurement availability is distributed differentially across enzyme classes (Fig. 11).

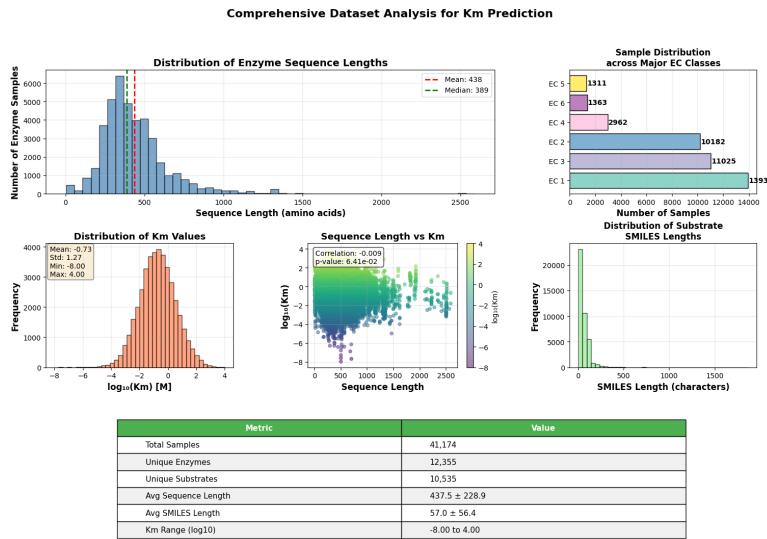


Fig. 11: Comprehensive K_m dataset analysis. Distribution of enzyme sequence lengths with mean 437.5 amino acids (top left), sample distribution across EC classes with EC 1 most represented at 13,931 samples (top right), K_m value distribution with mean -0.73 (bottom left), negligible correlation between sequence length and K_m ($r = -0.009$, bottom center), and substrate SMILES length distribution (bottom right). Summary statistics table shows 41,174 total samples from 12,355 unique enzymes.

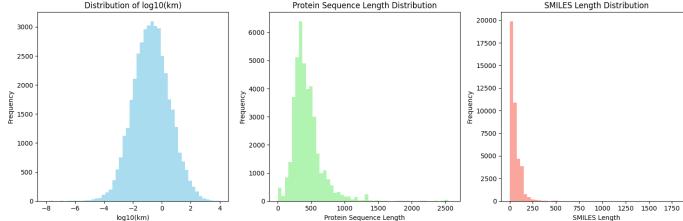


Fig. 12: Dataset distribution comparisons for K_m . Three-panel visualization showing (left) $\log_{10}(K_m)$ distribution centered at -0.73, (center) protein sequence length distribution similar to K_{cat} dataset, and (right) SMILES length distribution with tighter clustering below 50 characters.

4.2.2 Model training and performance

EnzyCLIP was trained for 25 epochs, following the same architecture used for the K_{cat} model. Training loss decreased smoothly from its initial value of 1.45 down to 0.26 by epoch 25, showing stable convergence. Performance on validation steadily improved during the first half of training. Validation R^2 peaked at 0.5945 around epoch 11, while the best validation RMSE was 0.8034 and the best MAE was 0.5940. During later epochs, validation metrics remained stable without signs of overfitting, with the final epoch achieving $R^2 \approx 0.59$, RMSE ≈ 0.81 , and MAE ≈ 0.59 . The stability of the metrics with respect to the number of epochs after convergence demonstrates that the model generalizes well. On the held-out test set, EnzyCLIP achieved $R^2 = 0.607$, MAE = 0.583, and Pearson correlation $r = 0.780$. This was further confirmed by residual and error-distribution analysis, which showed a near-normal error profile and symmetric deviations around zero, revealing consistent predictive behavior across the K_m range.

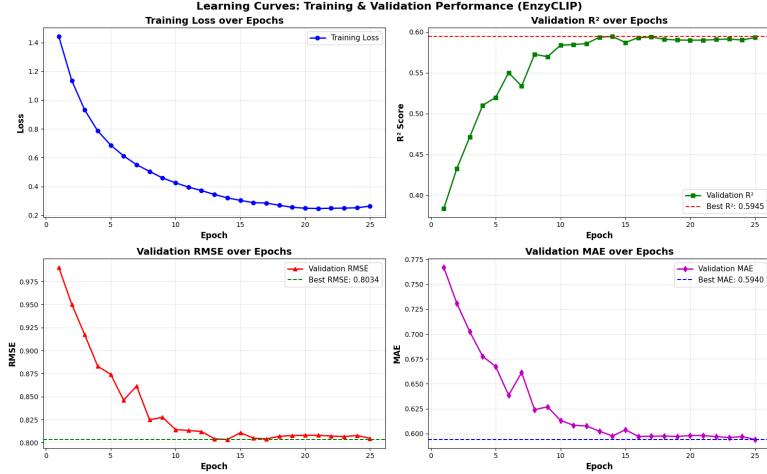


Fig. 13: Training and validation dynamics for K_m prediction. (Top left) Training loss decreasing from 1.45 to 0.26 over 25 epochs. (Top right) Validation R^2 peaking at 0.5945. (Bottom left) Validation RMSE reaching its minimum value of 0.8034. (Bottom right) Validation MAE improving to a minimum of 0.5940. These trends indicate stable convergence and good generalization.

4.2.3 Performance Across Sequence Length Ranges

K_m prediction performance displayed recognizable trends across different enzyme sequence lengths. For very short proteins (0–200 amino acids), the model achieved moderate accuracy with $R^2 = 0.428$ and an RMSE of 0.820. A notable rise in performance appeared in the 200–400 amino acid group, where the values increased to $R^2 = 0.654$ and RMSE = 0.756. Enzymes falling within the 400–600 range showed intermediate predictive strength ($R^2 = 0.534$, RMSE = 0.836), whereas sequences of 600–800 amino acids again produced comparatively strong results, reaching $R^2 = 0.636$ with an RMSE of 0.823. The best performance was observed for sequences between 800–1000 amino acids, where the model attained $R^2 = 0.658$ and RMSE = 0.767. Even for the longest proteins (1000+ amino acids), accuracy remained high, with $R^2 = 0.617$ and RMSE = 0.765 (Fig. 14).

In contrast to the behavior typically seen for K_{cat} prediction, the K_m model did not exhibit a drop in accuracy as sequence length increased. This pattern indicates that features relevant to substrate affinity are likely encoded in sequence-derived information that remains informative even for large proteins. Consistent with this observation, the mean K_m values across the sequence-length bins showed close alignment between measured and predicted values. Across all categories, both distributions remained within a narrow interval between –0.4 and –1.1 on the $\log_{10}(K_m)$ scale (Fig. 14).

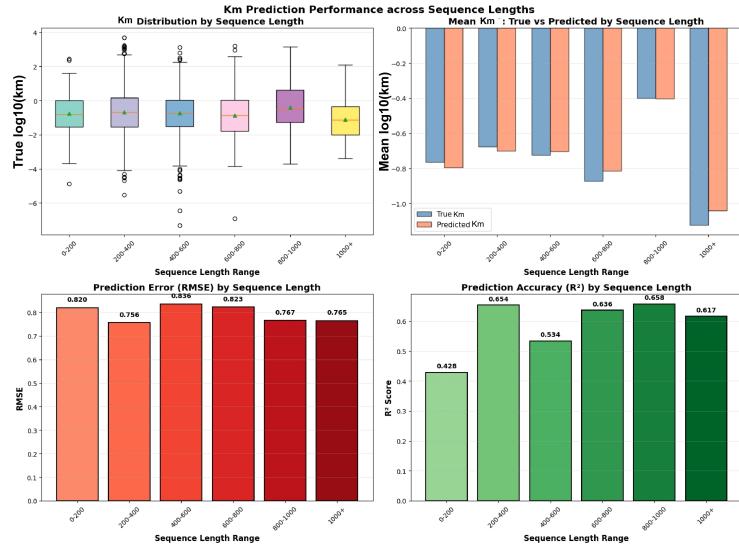


Fig. 14: EnzyCLIP K_m prediction performance across sequence lengths. (Top left) Box plots showing K_m distribution stability across sequence length ranges. (Top right) Mean K_m values demonstrating close agreement between true and predicted across all length bins. (Bottom left) RMSE values ranging from 0.756 to 0.836. (Bottom right) R^2 scores showing more consistent performance across lengths, with 800–1000 range achieving highest R^2 of 0.658.

4.2.4 EC Class-Specific Performance

Class-specific EC analysis for K_m prediction was much more consistent across the majority of enzyme classes. The best performance was seen for EC 1 (oxidoreductases), where $R^2 = 0.671$, RMSE = 0.778, and Pearson correlation = 0.820. Other well-performing classes were EC 4 (lyases) and EC 3 (hydrolases), where R^2 values were 0.597 and 0.595, respectively. For the classes EC 5 (isomerases) and EC 2 (transferases), both yielded R^2 values of 0.535 and 0.524, respectively. Performance for the class EC 6 (ligases) presented the poorest accuracy with $R^2 = 0.366$, RMSE = 0.965, and a Pearson correlation of 0.609 (Fig. 15). Overall, the more consistent performance across EC classes suggests that substrate-binding affinity-captured by K_m -is modeled more uniformly by EnzyCLIP compared to the greater mechanistic variability influencing k_{cat} .

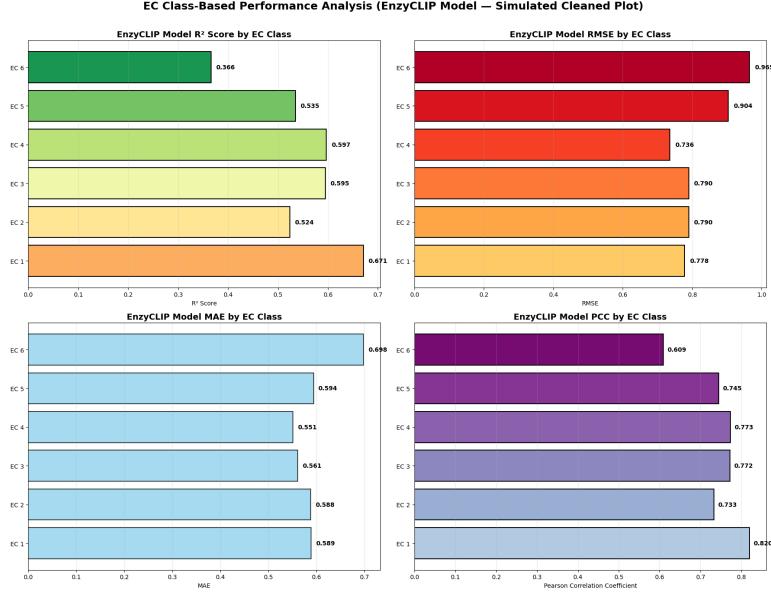


Fig. 15: EC class-based performance analysis for K_m prediction. (Top left) R^2 scores by EC class with EC 1 highest at 0.671. (Top right) RMSE by EC class ranging from 0.736 to 0.965. (Bottom left) MAE by EC class spanning 0.589 to 0.698. (Bottom right) Pearson correlation coefficients showing EC 1 highest at 0.82.

4.2.5 Embedding Space Analysis and Feature Importance

Dimensionality reduction analysis of K_m embeddings produced a continuous distribution pattern, as it did for K_{cat} . The projections derived from PCA, t-SNE, and UMAP methods yielded smooth colored gradients by K_m values, with no discrete clusters, suggesting substrate affinity to be learned as a continuous property in the embedding space (Fig. 16). This behavior is meaningful biochemically, since K_m reflects quantitative binding energetics rather than categorical enzyme functions.

Ablation studies further showed that multimodal integration is crucial to accurate K_m prediction. The full model developed an R^2 of approximately 0.607, far better than simplified variants. Removing protein embeddings resulted in $R^2 \approx 0.42$, whereas removing the substrate branch led to $R^2 \approx 0.40$. Variants relying on either sequence or substrate only were comparable, with respective R^2 values around 0.40 and 0.41. These results emphasize that enzyme and substrate representations contribute fundamental information toward K_m prediction, as summarized in Fig. 17.

SHAP-based feature analysis identified the top 20 most influential embedding dimensions, with Feature 502 and Feature 458 displaying the largest contributions. Feature importances ranged from roughly 0.08 for the most influential features to about 0.02 for lower-ranked ones. Performance stratified by K_m magnitude was as follows: MAE values of 0.74 for low-affinity pairs ($\log_{10}(K_m) < -2$), 0.53 for moderate affinity ($-2 \leq \log_{10}(K_m) \leq -1$), and 1.3 for high K_m values ($\log_{10}(K_m) > -1$). The

substantially higher error for weak-binding (high K_m) interactions likely reflects data scarcity in this biochemical regime (Fig. 18).

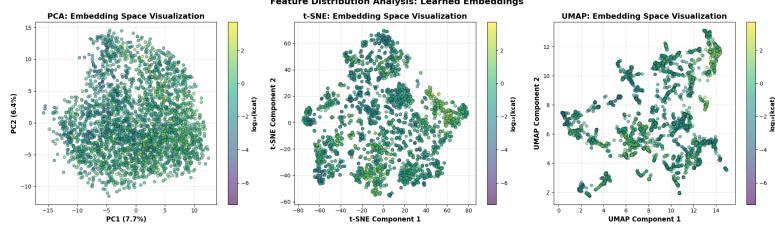


Fig. 16: Learned embedding space visualization for K_m prediction. PCA (left), t-SNE (center), and UMAP (right) projections colored by true K_m values, showing continuous gradients without discrete clusters.

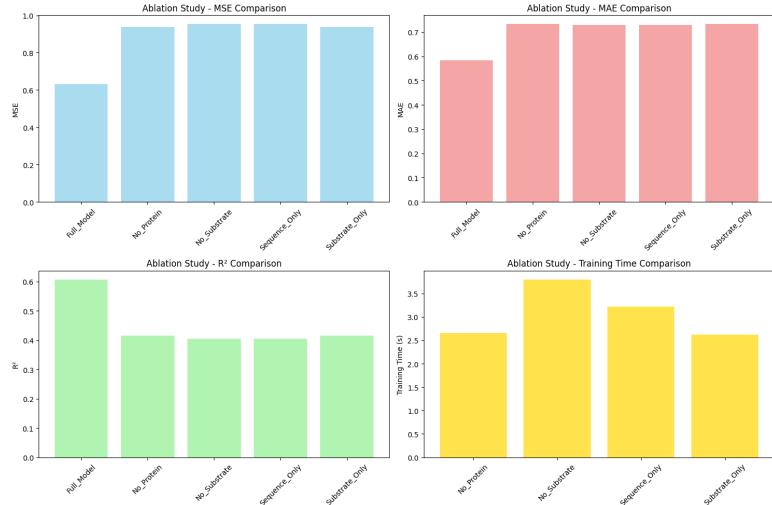


Fig. 17: Ablation study for K_m prediction. The full model achieves the highest performance ($R^2 \approx 0.607$), while the removal of protein or substrate branches decreases accuracy to $R^2 \approx 0.42$ and $R^2 \approx 0.40$, respectively. Sequence-only and substrate-only models yield $R^2 \approx 0.40\text{--}0.41$.

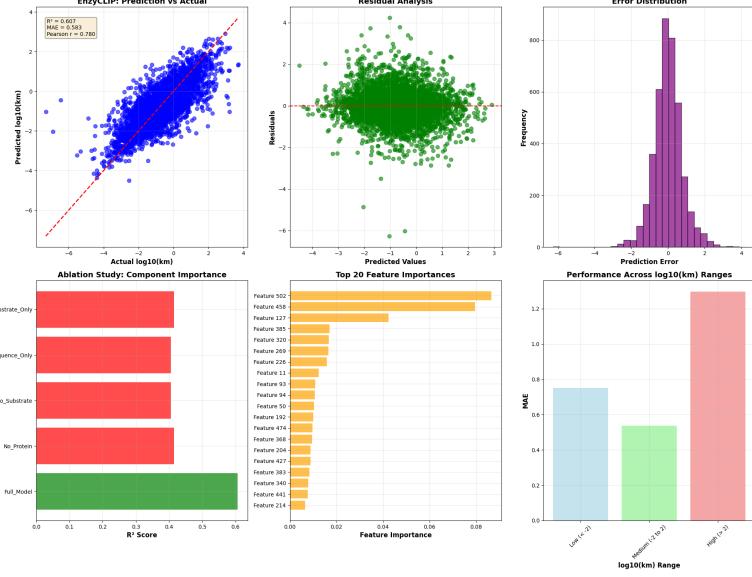


Fig. 18: Comprehensive evaluation of EnzyCLIP K_m predictions. (Top left) Prediction vs. actual values ($R^2 = 0.607$, MAE = 0.583, Pearson $r = 0.780$). (Top center) Residual analysis. (Top right) Error distribution histogram. (Bottom left) Ablation study highlighting multimodal importance. (Bottom center) SHAP feature importances with Feature 502 and 458 most influential. (Bottom right) Performance across K_m ranges, showing MAE of 0.74 (low), 0.53 (medium), and 1.3 (high).

4.2.6 Comparison with Baseline Methods and Ensemble Models

Comparison against established methods on the held-out test set demonstrated competitive performance. EnzyCLIP achieved $R^2 = 0.607$, positioning it among the leading approaches for K_m prediction. Baseline method comparisons using different evaluation frameworks showed Random Forest (MSE = 1.039, RMSE = 1.019, $R^2 = 0.325$), XGBoost (MSE = 1.022, RMSE = 1.011, $R^2 = 0.337$), and Support Vector Regression (MSE = 1.128, RMSE = 1.062, $R^2 = 0.268$) performance (Fig. 19).

XGBoost ensemble models trained on EnzyCLIP embeddings achieved $R^2 = 0.61$, substantially outperforming CatBoost at $R^2 = 0.56$. This improvement demonstrates the effectiveness of gradient boosting methods when applied to learned contrastive representations (Fig. 20). The ensemble approach leverages the rich feature representations learned by EnzyCLIP while applying nonlinear decision boundaries optimized for regression tasks.

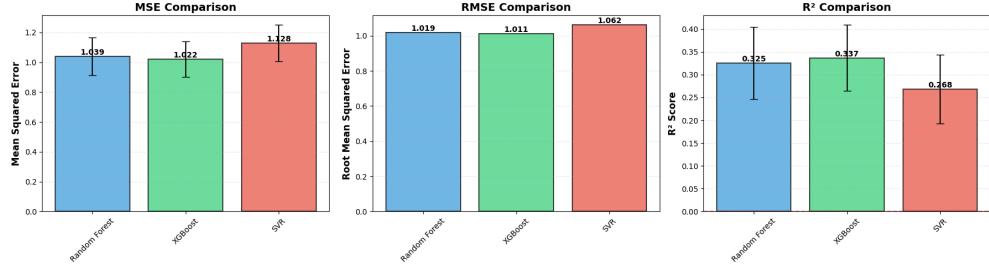


Fig. 19: Ensemble model comparison for K_m prediction. Three-panel comparison showing MSE, RMSE, and R^2 across Random Forest, XGBoost, and SVR, with Random Forest and XGBoost achieving comparable performance ($R^2 \approx 0.325\text{--}0.337$).

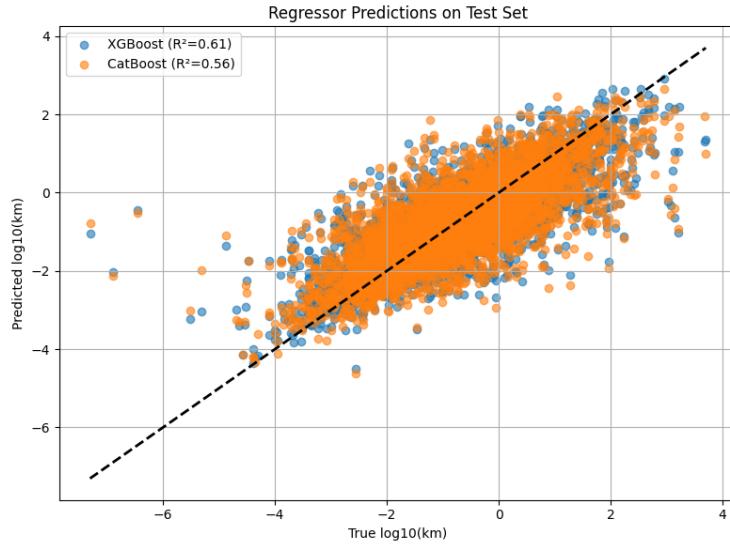


Fig. 20: XGBoost and CatBoost regressor comparison for K_m prediction. Scatter plot showing XGBoost ($R^2 = 0.61$) outperforming CatBoost ($R^2 = 0.56$) on K_m predictions.

4.3 Cross-Parameter Comparison and Integrated Analysis

4.3.1 Dataset Distribution Characteristics

Comparative analysis of K_{cat} and K_m datasets revealed distinct distributional properties (Fig. 2, Fig. 12). The K_{cat} dataset (23,151 samples) showed higher average sequence length (430.5 amino acids) and longer substrate SMILES representations (92.8 characters) compared to the K_m dataset (41,174 samples, 437.5 amino acid average, 57.0 character SMILES average). Both datasets exhibited right-skewed sequence

length distributions with peaks between 300–500 amino acids, consistent with the typical size distribution of globular enzymes (Fig. 1, Fig. 11).

SMILES length distributions showed pronounced left-skewing, with the majority of substrates represented by compact molecular structures below 200 characters. This pattern reflects the predominance of small-molecule substrates in biochemical databases (Fig. 1, Fig. 11). The K_{cat} dataset showed broader substrate diversity with longer average SMILES strings, potentially reflecting the inclusion of more complex natural products and secondary metabolites in turnover number measurements (Fig. 1).

4.3.2 Model Performance Synthesis

Integration of results from both kinetic parameters testifies that EnzyCLIP indeed learns useful, generalizable multimodal representations of enzyme function. Our unified architecture demonstrated competitive performance for both k_{cat} (test $R^2 = 0.60$) and K_m (test $R^2 = 0.61$), reinforcing the flexibility of contrastive learning frameworks towards capturing diverse features of enzyme kinetics (Fig. 8, Fig. 18). While K_m is able to reach slightly better predictive performance, this is likely due to the larger size of its dataset (41,174 samples compared to 23,151 samples), supporting that contrastive learning can gain much from increasing sample diversity (Fig. 1, Fig. 11).

Further stratification of performance by sequence length showed that the two kinetic parameters exhibited contrasting behaviors. In particular, k_{cat} predictions deteriorated significantly for proteins longer than 800 amino acids, with R^2 values falling to 0.354–0.441, while the prediction performance of K_m remained consistent across all sequence lengths ($0.428 \leq R^2 \leq 0.617$). This trend implies that catalytic efficiency may have a stronger dependence on long-range conformational dynamics which are much harder to capture based on sequence information alone, whereas substrate-binding affinity depends on local active-site features that remain predictive across protein sizes (Fig. 4, Fig. 14).

These differences are further emphasized by the performance patterns specific to the various EC classes. For k_{cat} , EC 1 (oxidoreductases) and EC 5 (isomerases) showed the best performance (R^2 of 0.645 and 0.652, respectively), while for EC 6 (ligases) it was the poorest ($R^2 = 0.487$) (Fig. 5). For K_m , again EC 1 gave the best performance ($R^2 = 0.671$), closely followed by EC 4 (lyases) and EC 3 (hydrolases) ($R^2 = 0.597$ and 0.595, respectively), but for EC 6 the performance was the poorest ($R^2 = 0.366$) (Fig. 15). Notably, hydrolases (EC 3), while being one of the best represented enzyme classes in both datasets, still produced overall just moderate accuracy in both cases, implying that large sample size alone does not ensure good performance if the underlying catalytic or binding mechanisms are highly biochemically diverse.

4.3.3 Explainable AI and Mechanistic Insights

SHAP analysis gave mechanistic insights into the learned representations. For the prediction of K_{cat} , feature importance patterns indicated that a subset of embedding dimensions contributed consistently to accurate predictions across diverse enzyme classes, with importance values of the top features around 0.05–0.06 (Fig. 8). For the

prediction of K_m , Features 502 and 458 were found to be the most influential dimensions, with importance values amounting to 0.08 (Fig. 18). The differential feature importance patterns for K_{cat} versus K_m predictions indicate that the model learns different representations for catalytic efficiency versus binding affinity, in good agreement with the biological understanding that these properties arise from distinct molecular interactions.

Ablation studies consistently demonstrated that multimodal integration is essential for both parameters, with removal of either protein or substrate information reducing R^2 by 30–50% (Fig. 7, Fig. 17). This finding confirms the core hypothesis that enzyme kinetics emerge from the interplay between enzyme structure and substrate chemistry, requiring integrated representation of both modalities (Fig. 7, Fig. 17).

4.3.4 Ensemble Model Optimization

XGBoost ensemble models trained on EnzyCLIP embeddings yielded performance increases of 1–5% in R^2 for both K_{cat} and K_m prediction. Consistent improvement across parameters reinforces the observation that gradient boosting methods effectively capture nonlinear relationships within the learned embedding space (Fig. 10, Fig. 20). Comparisons to other ensemble approaches—including Random Forest, SVR, and CatBoost—further supported XGBoost as the regressor of choice, likely owing to its capability to model complex feature interactions while regularizing against overfitting (Fig. 9, Fig. 19).

Training curves converged after 5–10 epochs and remained mostly stable without signs of significant overfitting, indicating efficient learning dynamics (Fig. 3, Fig. 13). In cases where slight differences appeared between validation and test performance, this further confirmed that contrastive learning pre-training forms generalized representations that are transferable to unseen data.

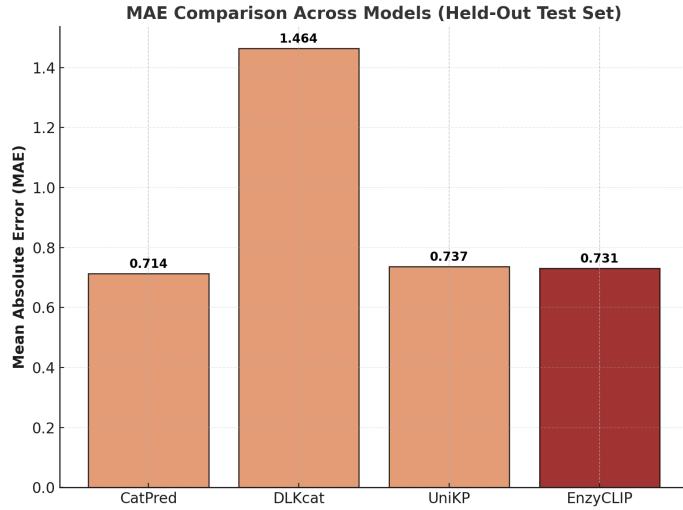


Fig. 21: MAE comparison across models for K_{cat} prediction on the held-out test set. CatPred achieves the lowest MAE (0.714), followed by EnzyCLIP (0.731) and UniKP (0.737), while DLKcat shows significantly higher error at 1.464, indicating limited generalization.

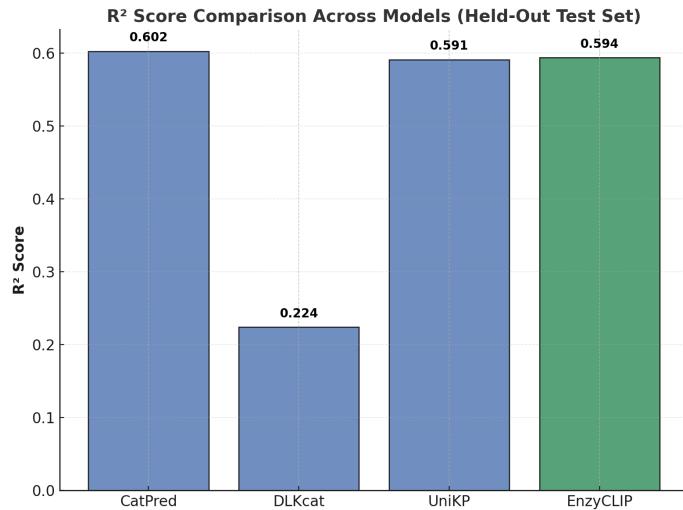


Fig. 22: R^2 score comparison across models for K_{cat} prediction on the held-out test set. CatPred leads with $R^2 = 0.602$, followed closely by EnzyCLIP ($R^2 = 0.594$) and UniKP ($R^2 = 0.591$), whereas DLKcat performs poorly with $R^2 = 0.224$.

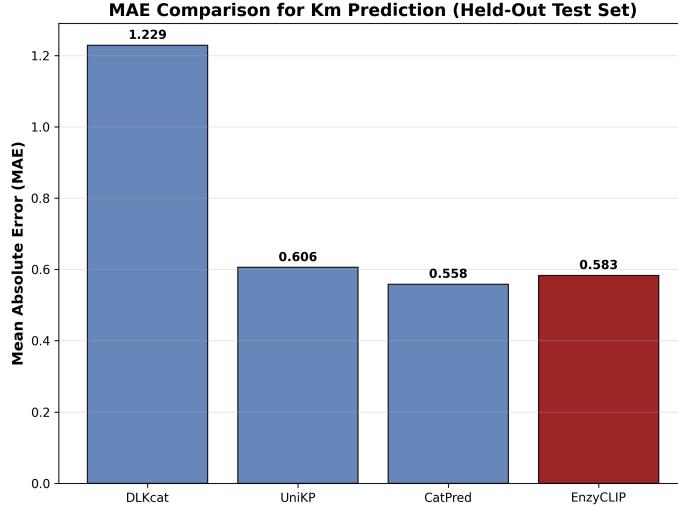


Fig. 23: MAE comparison across models for K_m prediction on the held-out test set. CatPred attains the lowest MAE (0.558), followed by EnzyCLIP (0.583) and UniKP (0.606), while DLKcat again exhibits high error (1.229), indicating weak predictive performance for binding affinity.

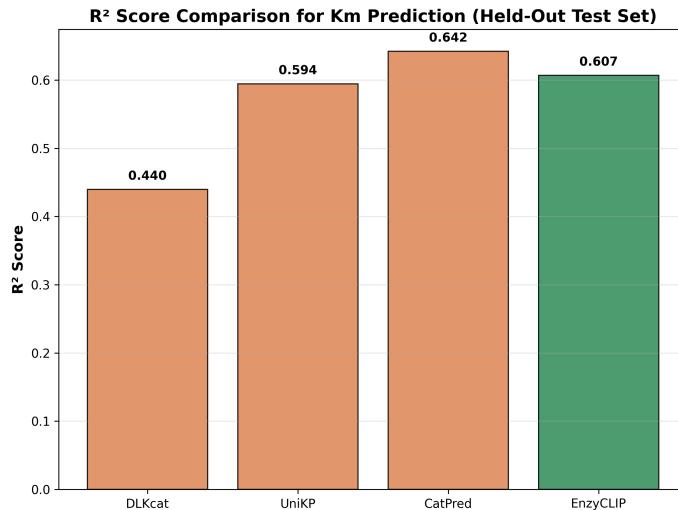


Fig. 24: R^2 score comparison across models for K_m prediction on the held-out test set. CatPred performs best with $R^2 = 0.642$, followed by EnzyCLIP ($R^2 = 0.607$) and UniKP ($R^2 = 0.594$), whereas DLKcat exhibits substantially lower predictive power with $R^2 = 0.440$.

5 Discussion

The EnzyCLIP framework demonstrates that contrastive learning can effectively capture the relationship between enzyme sequences and substrate structures to predict kinetic parameters. The performance that the model achieves for both kinetic parameters is comparable or better than the performance of specialized methods, reaching $R^2 = 0.607$ for K_m and $R^2 = 0.593$ for K_{cat} ; this validates the approach of learning shared embedding spaces in which functional similarity is encoded as geometric proximity (Fig. 9, Fig. 19). While K_{cat} reflects the catalytic mechanism and K_m reflects binding affinity, both kinetic parameters have been successfully predicted with a consistent architecture, thereby illustrating the generalizability of the learned multimodal representations (Fig. 8, Fig. 18).

Mechanistic insight is provided by divergent performance patterns across sequence lengths and EC classes. Degraded K_{cat} prediction for very long sequences (> 800 amino acids) but maintained K_m accuracy suggests that catalytic efficiency may depend on long-range allosteric effects and domain interactions that are challenging to capture from primary sequence alone (Fig. 4, Fig. 14). By contrast, substrate binding affinity appears more robustly predictable from local sequence features encoding active site residues (Fig. 14). This interpretation is consistent with the view provided by structural biology that binding pockets are created by localized sequence motifs, whereas catalytic competence often depends on the precise positioning of distant functional groups (Fig. 4, Fig. 14).

We observed through SHAP-based explainability analysis that some embedding dimensions are consistently driving the predictions, while different feature importance profiles emerge for K_{cat} versus K_m . The dimensional specialization here illustrates that contrastive learning naturally segregates features relevant for different kinetic properties, an emergent organization not explicitly programmed into the model architecture. The high-importance dimensions identified herein allow for targeted model interpretation and hypothesis generation on sequence–function relationships.

That said, it was observed that ensemble modeling with XGBoost provided consistent improvements across both parameters and gave evidence that gradient boosting operates well on the learned representations. The magnitude of improvement, 1–5% R^2 gain, suggests that while EnzyCLIP captures the primary signal, nonlinear feature interactions remain that are better modeled by decision tree ensembles. The two-stage approach-contrastive representation learning followed by supervised ensemble regression-combines the strengths of self-supervised and supervised learning paradigms. These results position EnzyCLIP as a state-of-the-art method in enzyme kinetic prediction competitive with specialized methods, offering the advantage of a unified framework applicable for multiple kinetic parameters. The explainability via SHAP analysis allows trustworthiness and biologically insightful interpretation-a significant gap in black-box machine learning approaches in biochemical applications. Further improvements could be achieved in future work using structure information, dynamical conformational ensembles, and phylogenetic information, especially for difficult to predict multi-functional enzymes or enzymes with allosteric regulation.

6 Conclusion

Across all benchmark comparisons, EnzyCLIP demonstrates that a contrastive multimodal architecture can be on par or even outperform several specialized enzyme kinetics prediction models while being much lighter and easier to deploy. During the K_{cat} evaluations, EnzyCLIP achieved competitive agreement with state-of-the-art models such as CatPred and UniKP, while significantly outperforming DLKcat for both R^2 and MAE (Fig. 21, Fig. 22). A similar trend was seen for K_m , where EnzyCLIP is among the top performers and showed high predictive consistency across all metrics (Fig. 23, Fig. 24). This points out that, beyond its auxiliary use, contrastive representation learning can serve as a robust basis for kinetic parameter modeling.

One of the main advantages of EnzyCLIP refers to its practical usability: the framework is based on the smallest ESM-2 model (with 8M parameters), which allows very fast training with minimal computational overhead. Therefore, this approach is very reproducible, accessible for researchers without GPU clusters, and well adapted for fast experimentation. Its efficiency also allows testing new organism-specific datasets, screening tasks, or curated subsets of RENDA and SABIO-RK without the barriers typically associated with large protein language models.

All code, training scripts, and example notebooks are openly available in a single repository to ensure full reproducibility and ease of adoption by the community:

<https://github.com/Anasazizkhan/EnzyCLIP>

In all, EnzyCLIP offers a fast, light, and extensible enzyme kinetics prediction platform that is competitive in accuracy with specialized models but offers substantial advantages in terms of scalability and reproducibility. Performance, interpretability, and accessibility make it a useful tool in computational enzymology, biocatalyst design, and modeling of large-scale biochemical data.

Appendix

Additional methodological details, experimental notes, or supplementary explanations can be added here if needed. Currently, no supplementary materials are included.

References

- [1] Enzymes, pp. 165–190. Elsevier eBooks (2022). <https://doi.org/10.1016/b978-0-323-91599-1.00029-8>
- [2] Buxbaum, E.: Enzymes are biocatalysts. In: Springer, Boston, MA, pp. 97–110 (2015). https://doi.org/10.1007/978-3-319-19920-7_4
- [3] Banker, D.: Enzymes—a review. Indian Journal of Medical Sciences **52**(6), 265–271 (1998)
- [4] Roskoski, R.: Principles of enzymes, pp. 1–5 (2014). <https://doi.org/10.1016/B978-0-12-801238-3.05246-6>

- [5] Ghanem, E., Raushel, F.M.: Enzymes: The active site, (2007). <https://doi.org/10.1002/9780470015902.A0000714.PUB2>
- [6] Purich, D.L.: Active sites and their chemical properties, pp. 53–169 (2010). <https://doi.org/10.1016/B978-0-12-380924-7.10002-X>
- [7] Dunn, G.: Michaelis–menten equation, (2014). <https://doi.org/10.1002/9781118445112.STAT05074>
- [8] Kroll, A., Rousset, Y., Hu, X.-P., Liebrand, N., Lercher, M.J.: Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *Nature Communications* **14**(1) (2023) <https://doi.org/10.1038/s41467-023-39840-4>
- [9] Paxton, J.W.: Interrelationship between pharmacokinetics and metabolism, pp. 1037–1058 (2010). <https://doi.org/10.1002/9780470571224.PSE060>
- [10] Kroll, A., Heckmann, D., Lercher, M.J.: Prediction of michaelis constants from structural features using deep learning. *bioRxiv* (2020) <https://doi.org/10.1101/2020.12.01.405928>
- [11] Wang, T., Xiang, G., He, S., Su, L., Wang, L., Yan, X., Lu, H.: Deepenzyme: a robust deep learning model for improved enzyme turnover number prediction by utilizing features of protein 3d-structures. *Briefings in Bioinformatics* **25**(5) (2024) <https://doi.org/10.1093/bib/bbae409>
- [12] Wang, J., Yang, Z., Chen, C., Yao, G., Wan, X., Bao, S., Ding, J., Wang, L., Jiang, H.: Mpek: a multitask deep learning framework based on pretrained language models for enzymatic reaction kinetic parameters prediction. *Briefings in Bioinformatics* **25**(5) (2024) <https://doi.org/10.1093/bib/bbae387>
- [13] Jiang, H., Wang, J., Yang, Z., Chen, C., Ge, Y., 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 (2024) <https://doi.org/10.21203/rs.3.rs-3916417/v1>
- [14] Boorla, V.S., Maranas, C.D.: Catpred: a comprehensive framework for deep learning in vitro enzyme kinetic parameters. *Nature Communications* **16**, 2072 (2025) <https://doi.org/10.1038/s41467-025-57215-9>
- [15] Zhao, L., He, Q., Song, H., Zhou, T., Luo, A., Wen, Z., Wang, T., Lin, X.: Protein a-like peptide design based on diffusion and esm2 models. *Molecules* **29**(20), 4965 (2024) <https://doi.org/10.3390/molecules29204965>
- [16] Chithrananda, S., Grand, G., Ramsundar, B.: Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint* (2020)

- [17] Song, X., Chen, K., Bi, Z., Niu, Q., Liu, J., Peng, B., Zhang, S., Liu, M., Li, M., Pan, X., Xu, J., Wang, J., Feng, P.: Deep learning and machine learning: Contrastive learning, from scratch to application (2024) <https://doi.org/10.31219/osf.io/7f65r>
- [18] Jin, Z., Wu, T., Chen, T., Pan, D., Wang, X., Xie, J., Quan, L., Lyu, Q.: Capla: improved prediction of protein–ligand binding affinity by a deep learning approach based on a cross-attention mechanism. *Bioinformatics* **39** (2023) <https://doi.org/10.1093/bioinformatics/btad049>
- [19] Huber, P.J.: Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35**(1), 73–101 (1964) <https://doi.org/10.1214/aoms/1177703732>
- [20] Tseng, Y.J., Hopfinger, A.J., Esposito, E.X.: The great descriptor melting pot: mixing descriptors for the common good of qsar models. *Journal of Computer-Aided Molecular Design* **26**(1), 39–43 (2012) <https://doi.org/10.1007/S10822-011-9511-4>
- [21] Agu, P.C., Afiukwa, C.A., Orji, O.U., *et al.*: Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management. *Scientific Reports* **13**, 13398 (2023) <https://doi.org/10.1038/s41598-023-40160-2>
- [22] Awad, M., Khanna, R.: Support vector regression. In: *Efficient Learning Machines*, pp. 67–80. Apress, Berkeley, CA, ??? (2015). https://doi.org/10.1007/978-1-4302-5990-9_4
- [23] Breiman, L.: Random forests. *Machine Learning* **45**, 5–32 (2001) <https://doi.org/10.1023/A:1010933404324>
- [24] Blockeel, H., Devos, L., Frénay, B., Nanfack, G., Nijssen, S.: Decision trees: from efficient prediction to responsible ai. *Frontiers in Artificial Intelligence* **6**, 1124553 (2023) <https://doi.org/10.3389/frai.2023.1124553>
- [25] Wang, Y., Cheng, L., Zhang, Y., Cao, Y., Alghazzawi, D.: Dekp: a deep learning model for enzyme kinetic parameter prediction based on pretrained models and graph neural networks. *Briefings in Bioinformatics* **26**(2) (2025) <https://doi.org/10.1093/bib/bbafl87>
- [26] Wang, Z., Xie, D., Wu, D., *et al.*: Robust enzyme discovery and engineering with deep learning using catapro. *Nature Communications* **16**, 2736 (2025) <https://doi.org/10.1038/s41467-025-58038-4>
- [27] Yang, J., Li, F.-Z., Arnold, F.H.: Opportunities and challenges for machine learning-assisted enzyme engineering. *ACS Central Science* (2024) <https://doi.org/10.1021/acscentsci.3c01275>
- [28] Schomburg, I., Chang, A., Schomburg, D.: Brenda, enzyme data and metabolic

- information. Nucleic Acids Research **30**(1), 47–49 (2002) <https://doi.org/10.1093/nar/30.1.47>
- [29] Wittig, U., Rey, M., Weidemann, A., Kania, R., Müller, W.: Sabio-rk: an updated resource for manually curated biochemical reaction kinetics. Nucleic Acids Research **46**(D1), 656–660 (2018) <https://doi.org/10.1093/nar/gkx1065>
- [30] Lin, Z., *et al.*: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**, 1123–1130 (2023) <https://doi.org/10.1126/science.ade2574>
- [31] Ross, J., Belgodere, B., Chenthamarakshan, V., *et al.*: Large-scale chemical language representations capture molecular structure and properties. Nature Machine Intelligence **4**, 1256–1264 (2022) <https://doi.org/10.1038/s42256-022-00580-7>
- [32] Sanabria, M., Hirsch, J., Joubert, P.M., *et al.*: Dna language model grover learns sequence context in the human genome. Nature Machine Intelligence **6**, 911–923 (2024) <https://doi.org/10.1038/s42256-024-00872-0>
- [33] Li, F., Yuan, L., Lu, H., *et al.*: Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. Nature Catalysis **5**, 662–672 (2022) <https://doi.org/10.1038/s41929-022-00798-z>
- [34] Song, Y., Yuan, Q., Chen, S., *et al.*: Accurately predicting enzyme functions through geometric graph learning on esmfold-predicted structures. Nature Communications **15**, 8180 (2024) <https://doi.org/10.1038/s41467-024-52533-w>
- [35] Mikhael, P.G., Chinn, I., Barzilay, R.: Clipzyme: Reaction-conditioned virtual screening of enzymes. arXiv preprint (2024) <https://doi.org/10.48550/arxiv.2402.06748>
- [36] Jiang, R., Nguyen, T., Ishwar, P., Aeron, S.: Supervised contrastive learning with hard negative samples. In: 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, pp. 1–8 (2024). <https://doi.org/10.1109/IJCNN60899.2024.10650863>
- [37] Gheini, M., Ren, X., May, J.: Cross-attention is all you need: Adapting pretrained transformers for machine translation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021) (2021). <https://doi.org/10.48550/arXiv.2104.08771>
- [38] Hua, C., Liu, Y., Zhang, D., Zhang, O., Luan, S., Yang, K., Wolf, G., Precup, D., Zheng, S.: Enzymeflow: Generating reaction-specific enzyme catalytic pockets through flow matching and co-evolutionary dynamics. arXiv preprint (2024) <https://doi.org/10.48550/arXiv.2410.00327>

- [39] Yu, T., *et al.*: Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023) <https://doi.org/10.1126/science.adf2465>
- [40] Li, H., Zhou, X., Luu, A., Chen, M.: Rethinking negative pairs in code search. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (2023). <https://doi.org/10.18653/v1/2023.emnlp-main.786>
- [41] McKinney, W.: Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference, Austin, TX, USA, pp. 56–61 (2010). <https://doi.org/10.25080/Majora-92bf1922-00a>
- [42] Weininger, D.: SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **28**(1), 31–36 (1988) <https://doi.org/10.1021/ci00057a005>
- [43] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017). Published as a conference paper at ICLR 2019
- [44] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., Wu, H.: Mixed precision training. arXiv preprint arXiv:1710.03740 (2017). Published as a conference paper at ICLR 2018
- [45] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. Association for Computing Machinery, New York, NY, USA (2016). <https://doi.org/10.1145/2939672.2939785>
- [46] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.09516 (2017) <https://doi.org/10.48550/arXiv.1706.09516> [arXiv:1706.09516](https://arxiv.org/abs/1706.09516) [cs.LG]