Multimodal Regression for Enzyme Turnover Rates Prediction

Bozhen Hu[1,2] , **Cheng Tan**[1] , **Siyuan Li**[1] , **Jiangbin Zheng**[1] , **Jun Xia**[1*] , **Stan Z. Li**[1*]

[1]AI Division, School of Engineering, Westlake University

[2]Zhejiang University

{hubozhen,tancheng,lisiyuan,zhengjiangbin,xiajun,stan.zq.li}@westlake.edu.cn

Abstract

The enzyme turnover rate is a fundamental parameter in enzyme kinetics, reflecting the catalytic efficiency of enzymes. However, enzyme turnover rates remain scarce across most organisms due to the high cost and complexity of experimental measurements. To address this gap, we propose a multimodal framework for predicting the enzyme turnover rate by integrating enzyme sequences, substrate structures, and environmental factors. Our model combines a pre-trained language model and a convolutional neural network to extract features from protein sequences, while a graph neural network captures informative representations from substrate molecules. An attention mechanism is incorporated to enhance interactions between enzyme and substrate representations. Furthermore, we leverage symbolic regression via Kolmogorov-Arnold Networks to explicitly learn mathematical formulas that govern the enzyme turnover rate, enabling interpretable and accurate predictions. Extensive experiments demonstrate that our framework outperforms both traditional and state-of-the-art deep learning approaches. This work provides a robust tool for studying enzyme kinetics and holds promise for applications in enzyme engineering, biotechnology, and industrial biocatalysis.

1 Introduction

Enzymes, a type of protein found in cells, facilitate and accelerate chemical reactions crucial for various bodily functions, including muscle building, detoxification, and digestion. They often work in conjunction with other substances like stomach acid and bile [Kosal, 2023], making the quantitative study of enzyme kinetics an important topic. The enzyme turnover rate (number) k_{cat}, which defines the maximum chemical conversion rate of a reaction, is a critical parameter for understanding the metabolism, proteome allocation, growth, and physiology of an organism [Li *et al.*,

2022a]. Currently, the determination of enzyme kinetic parameters heavily relies on laboratory experimentation [Nilsson *et al.*, 2017]. These procedures are time-consuming, costly, and labor-intensive, resulting in a restricted database of experimentally determined kinetic parameters due to the absence of high-throughput techniques. Consequently, the repositories of k_{cat} values in enzyme databases such as BRENDA [Schomburg *et al.*, 2017] and SABIO-RK [Wittig *et al.*, 2018] remain sparse when compared to the vast array of organisms and metabolic enzymes. While the sequence database UniProtKB [Magrane and Consortium, 2011] now boasts over 248M (million) protein sequences, the enzyme databases BRENDA and SABIO-RK contain only 17K (thousand) of experimentally measured kinetic parameters [Yu *et al.*, 2023]. This scarcity of k_{cat} data in databases underscores the need for the development of computational methods to predict k_{cat}.

With the rapid advancement of deep learning (DL) models, their applications have expanded to various fields, such as drug design, and enzyme reaction prediction [Hu *et al.*, 2024]. The prediction of kinetic parameters using DL-based techniques can be framed as a compound-protein interaction (CPI) prediction problem. This approach has been employed to predict various enzyme-related parameters, such as binding affinities (k_d), Michaelis–Menten constants (k_m), and enzyme turnover rates (k_{cat}). For example, DLKcat [Li *et al.*, 2022a] is a famous DL approach for k_{cat} prediction for metabolic enzymes from any organism using protein sequences and substrate structures. Based on this, the subsequent work is UniKP models [Yu *et al.*, 2023], which considers environmental factors such as pH and temperature that can impact enzyme kinetics significantly. Moreover, DLTKcat [Qiu *et al.*, 2024] is proposed to quantify the influence of temperature on the k_{cat} prediction, which demonstrates the feature importance of temperature under different cases [Arroyo *et al.*, 2022]. However, these works do not explicitly learn relationships between k_{cat} and environmental factors. Additionally, they employ different models to independently extract representations from enzyme sequences and substrates, lacking deeper interactions and associations. This is despite the profound connection between enzymes and substrates in enzymatic reactions, akin to the classic lock-and-key (or template) theory of enzyme specificity [Prokop *et al.*, 2012], as shown in Figure 1.

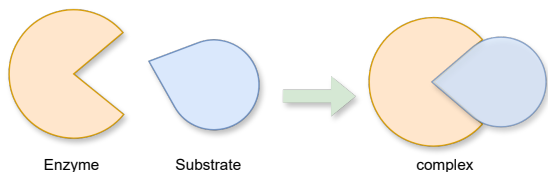---

*Corresponding authors.

Figure 1: An illustration of the enzyme and substrate lock-and-key model.

To address the limited availability of k_{cat} values, we propose ProKcat, a multimodal deep learning framework that integrates enzyme sequences, substrate structures, and environmental factors. Enzyme sequences are processed using a pre-trained protein language model (LM) [Lin *et al.*, 2023] and a convolutional neural network (CNN), while substrate molecules, represented in SMILES format [Honda *et al.*, 2019], are encoded with a graph neural network (GNN) [Zhou *et al.*, 2020; Liu *et al.*, 2023]. An attention module enhances interactions between enzyme and substrate representations through soft alignment. To improve interpretability, we adopt Kolmogorov–Arnold Networks [Liu *et al.*, 2024] for symbolic regression, enabling explicit modeling of relationships among inputs such as temperature and k_{cat}.

The contributions of this paper are summarized as follows: (1) A comprehensive multimodal framework for predicting enzyme turnover rates, which fuses various types of information and generates meaningful representations. (2) A specially designed interaction attention module that enhances the interactions derived from enzyme sequences and substrate structures. (3) The first DL model to employ symbolic regression for k_{cat} prediction with high efficiency, establishing clearer relationships between input factors and k_{cat}. (4) Extensive experiments demonstrating that our proposed framework outperforms existing k_{cat} prediction models, offering a promising approach for understanding enzyme kinetics with significant potential impact on biochemistry.

2 Related Works

2.1 Deep Learning-based Enzyme Turnover Rates Prediction

In the initial stage, CNNs [LeCun *et al.*, 1995], recurrent neural networks (RNNs) [Grossberg, 2013], and GNNs are employed to process enzyme and substrate features, followed by linear regression to predict enzyme kinetic parameters [Lim *et al.*, 2021]. For instance, DLKcat [Li *et al.*, 2022a] integrates a GNN for substrates and a CNN for proteins, facilitating high-throughput prediction of k_{cat} and identifying critical amino acid residues influencing these predictions. Subsequently, EF-UniKP and Revised UniKP [Yu *et al.*, 2023] are developed to predict temperature-dependent k_{cat} values using an Extra Tree [Sharaff and Gupta, 2019] model, which processes concatenated representation vectors derived from a pre-trained LM, ProtT5 [Elnaggar *et al.*, 2021], and a SMILES transformer model [Honda *et al.*, 2019]. TurNuP [Kroll *et al.*, 2023] characterizes chemical reactions through differential reaction fingerprints and repre-

sents enzymes using a modified Transformer model [Vaswani *et al.*, 2017]. GELKcat [Du *et al.*, 2023] assigns weights to substrate and enzyme features using an adaptive gate network. In order to tackle the out-of-distribution problem, Cat-Pred [Boorla and Maranas, 2024] outputs predictions as gaussian distributions (including a mean and a variance). Recently, DLTKcat [Qiu *et al.*, 2024] has shown potential in enzyme sequence design by predicting the impact of amino acid substitutions on k_{cat} across various temperatures. However, these methods typically extract feature vectors from protein sequences and substrates independently, followed by a simple predictor, with limited interaction between the two modules. The evident relationships between temperature and k_{cat} remain underexplored. To address these issues, we propose the incorporation of an additional attention module and the use of Kolmogorov-Arnold Networks (KANs) for symbolic regression of k_{cat}.

2.2 Compound–Protein Interaction Prediction

Traditional methods for CPI prediction typically involve screening candidates from a vast chemical space using various experimental assays [Trott and Olson, 2010] or employing molecular dynamics simulations [Hollingsworth and Dror, 2018], both of which are inefficient. Recent advances in DL, however, have revolutionized CPI prediction by offering new approaches. Most DL-based techniques represent compounds as one-dimensional (1D) sequences or molecular graphs and proteins as 1D sequences, performing joint representation learning and interaction prediction within a unified framework. For instance, DeepConvDTI [Lee *et al.*, 2019] employs CNNs to extract low-dimensional representations of compounds and proteins, concatenates these representations, and then feeds them into fully connected (FC) layers to predict interactions. HyperattentionDTI [Zhao *et al.*, 2022] models complex non-covalent inter-molecular interactions among atoms and amino acids based on a CNN. More recently, PerceiverCPI [Nguyen *et al.*, 2023] has utilized a cross-attention mechanism to enhance the learning capabilities for compound and protein interaction representations. PSC-CPI [Wu *et al.*, 2024] captures the dependencies between protein sequences and structures through multi-scale contrastive learning.

3 Methodology

3.1 Preliminaries

Problem Definitions and Notations. Let $\mathcal{S} = \{s_i\}_{i=1,...,L_p}$ be a protein sequence with the length L_p, s_i is the i-th amino acid. A SMILES format [Honda *et al.*, 2019] of compound is transformed by RDKit [Bento *et al.*, 2020] as a graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_i\}_{i=1,...,N_v}$ and $\mathcal{E} = \{\varepsilon_{ij}\}_{i,j=1,...,N_v}$ denote the sets of vertices and edges with N_v atoms, ε_{ij} represents there is a chemical bond between atom v_i and v_j. The environmental factor, temperature, is denoted as T.

Extended Connectivity Fingerprints (ECFPs) [Rogers and Hahn, 2010] are a type of topological fingerprint used to characterize molecular substructures. They describe the features of substructures by considering each atom and its surrounding circular neighborhoods within a specified diameter range,
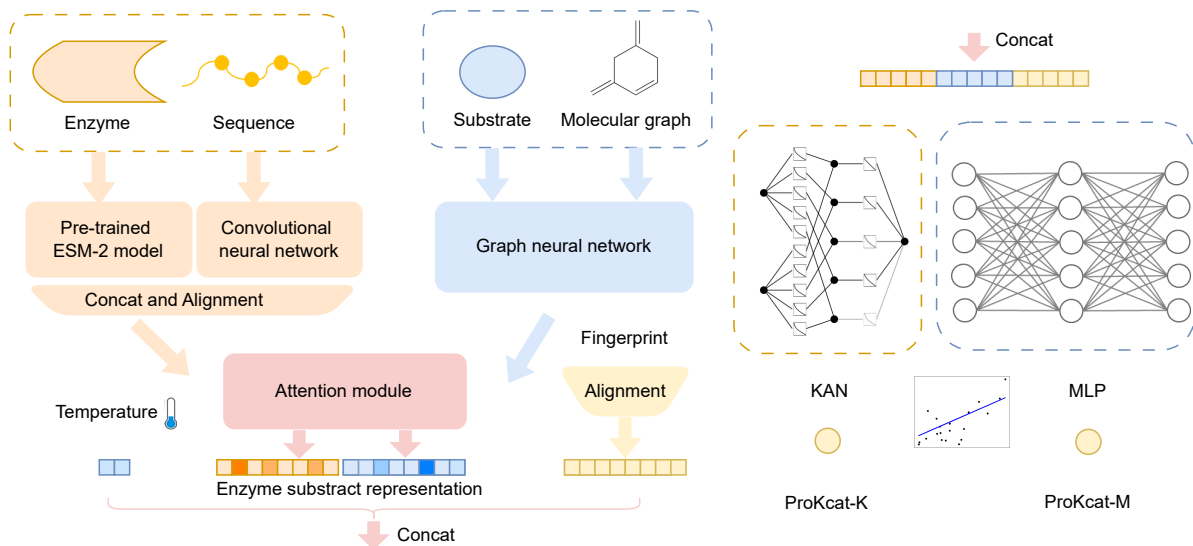
Figure 2: The overall framework of ProKcat integrates a pre-trained ESM-2 model and a CNN to analyze enzyme sequences, while a GNN processes substrate structures. To enhance the interaction between enzyme and substrate representations, an attention module is introduced. The attention-weighted outputs are then combined with molecular fingerprint features and temperature values. These fused features are fed into either a Kolmogorov–Arnold Network (ProKcat-K) or a multilayer perceptron (ProKcat-M) to predict $\log_{10} k_{\text{cat}}$ values.

which has been demonstrated effective in small molecular characterization, similarity searching, and compound-protein representation learning [Kroll *et al.*, 2023]. We use the open-source toolkit RDKit [Bento *et al.*, 2020] in cheminformatics to calculate the fingerprint of compounds to obtain the feature vector $\mathbf{h}_f$ of length 1024 as another part of the input. Our purpose is for the regression of the $\log_{10} k_{\text{cat}}$ values.

Enzyme Turnover Rates. Based on the Michaelis-Menten kinetics equation [Michaelis *et al.*, 1913], k_{cat} represents the number of substrate molecules converted to product per unit time by a single enzyme molecule when the enzyme is fully saturated with substrate, having units of time^{-1}, it is also called the enzyme turnover number.

Arrhenius Equation. Several models have been proposed to elucidate the relationship between temperature and biological processes [Brown *et al.*, 2004], with the Arrhenius equation emerging as the predominant choice in this field [Arroyo *et al.*, 2022]. This equation describes how the rate constant of a chemical reaction varies with the absolute temperature according to the formula:

$$k = Ae^{\frac{-E}{k_B T}} \tag{1}$$

where k represents a biological parameter (such as enzyme reaction rate), k_B denotes Boltzmann's constant, T stands for the absolute temperature, E signifies the effective activation energy for a specific process, and A serves as a normalization constant that characterizes the process as a whole. It can be rearranged as $\ln k = \frac{-E}{k_B}\frac{1}{T} + \ln A$, meaning a linear relationship between $\ln k$ and $\frac{1}{T}$.

Kolmogorov-Arnold Networks. KAN [Liu *et al.*, 2024] is based on the Kolmogorov-Arnold representation theorem [Tikhomirov, 1991; Braun and Griebel, 2009], which

states that any multivariate continuous function $f : [0,1]^n \to \mathbb{R}$ on a bounded domain can be written as a finite composition of continuous functions of a single variable x_i and the binary operation of addition,

$$f(x) = \sum_{j=1}^{2n+1} \Phi_j \left(\sum_{i=1}^{n} \phi_{j,i}(x_i) \right) \tag{2}$$

where $\phi_{j,i} : [0,1]^n \to \mathbb{R}$ and $\Phi_j : \mathbb{R} \to \mathbb{R}$. This 2-layer width Kolmogorov-Arnold representation may not be smooth. Thus, Liu *et al.* extend this theorem by proposing a generalized architecture with wider and deeper KANs. Given a vector $\mathbf{x} \in \mathbb{R}^{n_0}$, and $\mathbf{\Phi}_l$ represents the l-th KAN layer, a KAN with L_K layers can be expressed as

$$\text{KAN}(\mathbf{x}) = \mathbf{\Phi}_{L_K-1} \circ \cdots \circ \mathbf{\Phi}_1 \circ \mathbf{\Phi}_0 \circ \mathbf{x} \tag{3}$$

where $\circ$ means function composition. The function $\phi(x)$ is composed of the sum of the SiLU basis function [Paul *et al.*, 2022] and a linear combination of B-splines during its implementation. B-splines are commonly employed for interpolating or approximating data points in a seamless fashion. The definition of a spline involves specifying its order, typically set at 3, and the number of intervals, refering to the number of segments or subintervals between adjacent control points [Vaca-Rubio *et al.*, 2024].

3.2 Overall Framework

The framework of ProKcat, as depicted in Figure 2, leverages multimodal features of enzymes and substrates. We use a currently prevalent pre-trained protein LM, ESM-2 (650M) [Lin *et al.*, 2022] to generate protein sequence embeddings and adopt a CNN to learn from enzyme sequence on this task; then, the learned embeddings are concatenated. A feature

alignment network with MLP layers standardizes latent features to a unified dimension d, yielding the feature vector of enzyme sequences $\mathbf{h}_p \in \mathbb{R}^{L_p \times d}$. Similarly, another feature alignment network transforms ECFPs feature vectors $\mathbf{h}_f \in \mathbb{R}^{1024}$ into the compound latent space $\mathbf{h}'_f \in \mathbb{R}^d$. For substrates represented as a graph $G = (\mathcal{V}, \mathcal{E})$, a graph attention network (GAT) is employed to learn from the compound graph structure. GNNs, known for their effectiveness in the enzyme kinetics parameters prediction tasks [Li *et al.*, 2022a; Qiu *et al.*, 2024], update atom vectors and their neighboring vectors through neural network transformations. The GAT outputs real-valued molecular vector representations for substrates, denoted as $\mathbf{h}_c \in \mathbb{R}^{N_v \times d}$.

Enzyme-Substrate Attention Module. In contrast to previous k_{cat} prediction methods like UniKP [Yu *et al.*, 2023] and DLKcat [Li *et al.*, 2022a], our approach involves the design of an enzyme-substrate attention module to amplify interactions between enzymes and substrates. The enzyme vectors $\mathbf{h}_p \in \mathbb{R}^{L_p \times d}$ and compound vectors $\mathbf{h}_c \in \mathbb{R}^{N_v \times d}$ are derived in latent spaces. A soft alignment matrix $\mathcal{A} \in \mathbb{R}^{N_v \times L_p}$ is computed as follows:

$$\mathcal{A} = \sigma(\mathbf{h}_c W \mathbf{h}_p^\top) \quad (4)$$

$\mathcal{A}_{ij}$ represents the interaction strength between the i-th atom of compounds and j-th residue of proteins [Li *et al.*, 2022b]. The parameter matrix W is trainable, and $\sigma(\cdot)$ denotes an activation function, such as the Tanh function [Fan, 2000], $^\top$ denotes transposition. Then, we compute the intermediate compound-to-protein and protein-to-compound features using FC layers and the soft alignment matrix $\mathcal{A}$,

$$\begin{aligned} \mathbf{h}_{p2c} &= \mathcal{A} \cdot \text{FC}(\mathbf{h}_p) \\ \mathbf{h}_{c2p} &= \mathcal{A}^\top \cdot \text{FC}(\mathbf{h}_c) \end{aligned} \quad (5)$$

where $\mathbf{h}_{p2c} \in \mathbb{R}^{N_v \times d}$ and $\mathbf{h}_{c2p} \in \mathbb{R}^{L_p \times d}$ represent the learned compound-to-protein and protein-to-compound features. The weights of protein features $\mathbf{h}_p$ should be derived from the compound-to-protein features $\mathbf{h}_{c2p}$ and the features $\mathbf{h}_p$ themselves, with similar operations for the weights of compound features. Therefore, the attention weights can be computed using the Softmax [Joulin *et al.*, 2017] as

$$\begin{aligned} \alpha_p &= \text{Softmax}(\text{FC}(\mathbf{h}_{c2p} \parallel \mathbf{h}_p)) \\ \alpha_c &= \text{Softmax}(\text{FC}(\mathbf{h}_{p2c} \parallel \mathbf{h}_c)) \end{aligned} \quad (6)$$

where $\alpha_p \in \mathbb{R}^{L_p \times 1}$ and $\alpha_c \in \mathbb{R}^{N_v \times 1}$ represent the normalized attention weights of protein and compound features, respectively, and $\parallel$ denotes the concatenation operation. Subsequently, the attention-weighted protein and compound features are computed through matrix multiplication as $\mathbf{h}'_p = \alpha_p \cdot \mathbf{h}_p$ and $\mathbf{h}'_c = \alpha_c \cdot \mathbf{h}_c$, resulting in the weighted features $\mathbf{h}'_p \in \mathbb{R}^{L_p \times d}$ and $\mathbf{h}'_c \in \mathbb{R}^{N_v \times d}$.

This enzyme-substrate interaction process can be performed within a multi-head attention framework, where the distinct weighted protein and compound features from various heads are combined to derive the final enzyme-substrate representations. This mechanism is expressed as

$$\begin{aligned} \mathbf{h}_p^{'multi} &= \parallel_i^{L_h} \text{Softmax}(\text{FC}^i(\mathbf{h}_{c2p}^i \parallel \mathbf{h}_p)) \cdot \mathbf{h}_p \\ \mathbf{h}_c^{'multi} &= \parallel_i^{L_h} \text{Softmax}(\text{FC}^i(\mathbf{h}_{p2c}^i \parallel \mathbf{h}_c)) \cdot \mathbf{h}_c \end{aligned} \quad (7)$$

where there are L_h heads, $\mathbf{h}_p^{'multi} \in \mathbb{R}^{L_p \times L_h d}$, $\mathbf{h}_c^{'multi} \in \mathbb{R}^{N_v \times L_h d}$, then, linear projections are conducted to make the generated representations have a unified latent dimension.

Multivariable Regression. In latent spaces, we have acquired the weighted protein and compound representations $\mathbf{h}'_p \in \mathbb{R}^{L_p \times d}$ and $\mathbf{h}'_c \in \mathbb{R}^{N_v \times d}$, along with aligned fingerprint feature vectors $\mathbf{h}'_f \in \mathbb{R}^d$. A global average pooling operation is performed on these enzyme and substrate feature vectors to derive the comprehensive protein and compound level features, denoted as $\mathbf{h}''_p \in \mathbb{R}^d$ and $\mathbf{h}''_c \in \mathbb{R}^d$. Subsequently, these feature vectors are concatenated, resulting in $\mathbf{h} = [\mathbf{h}''_p \parallel \mathbf{h}''_c \parallel \mathbf{h}'_f]$, where $\mathbf{h} \in \mathbb{R}^{3d}$. Referring to the Arrhenius equation depicted in Eq. 1, the environmental factor, temperature, T, along with its reciprocal $\frac{1}{T}$, are introduced as additional controllable variables within the latent space, which are characterized by the learned features $\mathbf{h} \in \mathbb{R}^{3d}$.

In this paper, we are presented with a choice between two approaches: one entails the utilization of MLPs for regression to predict the $\log_{10} k_{\text{cat}}$ values, designated as the ProKcat-M; while the alternative option involves constructing a KAN for the prediction of these values, known as the ProKcat-K. Referring to Eq. 3, the KAN module with L_K layers in the ProKcat-K model can be mathematically formulated as

$$\log_{10} k_{\text{cat}} = \mathbf{\Phi}_{L_K-1} \circ \cdots \circ \mathbf{\Phi}_1 \circ \mathbf{\Phi}_0 \circ (\mathbf{h} \parallel T \parallel \frac{1}{T}) \quad (8)$$

here, the input feature vectors comprise two components: one being the learned enzyme-substrate weighted representations derived from a deep neural network, and the other component consisting of the environmental variables, specifically temperature, T and $\frac{1}{T}$. The former component is acquired from a black-box model, implying the inability to explicitly define a formula expressing the representations and their associations with the input enzyme sequences and substrate compounds. Nevertheless, a clear linear relationship exists between $\log_{10} k_{\text{cat}}$ and $\frac{1}{T}$ based on the Arrhenius equation, indicating that the input variables T and $\frac{1}{T}$ exhibit explicit relationships with the output $\log_{10} k_{\text{cat}}$.

Conventional symbolic regression techniques are commonly intricate and arduous to diagnose. Their outcomes frequently lack clear intermediate insights. In contrast, KANs engage in continuous exploration, yielding smoother and more robust results, which have been demonstrated superior performance in function representation compared to MLPs across various tasks such as regression and partial differential equation (PDE) solving [Liu *et al.*, 2024]. As a result, the ProKcat-K model can establish direct associations between input variables and the output, offering partial interpretability. Notably, with the given pre-trained sequence-substrate representations and temperature inputs, the direct calculation of $\log_{10} k_{\text{cat}}$ based on the learned formula becomes feasible. This capability holds significant implications for the fields of bioinformatics and biochemistry.

4 Experiments

4.1 Experimental Setup

Datasets. The fundamental details pertaining to enzymatic reactions are sourced from the databases BRENDA [Schom-
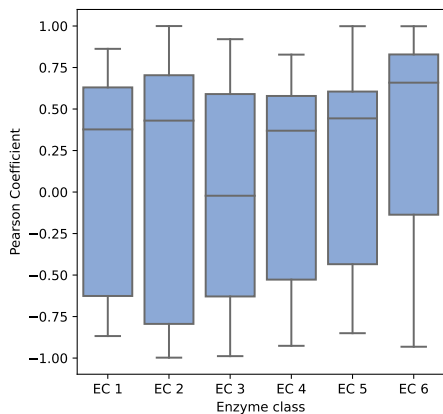
Figure 3: Pearson correlations between temperature and k_{cat} (with a significance level of $p < 0.05$) across diverse enzyme classes.

burg *et al.*, 2017] and SABIO-RK [Wittig *et al.*, 2018], encompassing key attributes such as the enzyme commission (EC) number, enzyme type, operating temperature, k_{cat} values, substrate name, Uniprot ID, etc. Leveraging the UniProt ID and substrate designation, the enzyme sequences and substrate SMILES strings are retrieved from the UniProtKB repository [Magrane and Consortium, 2011] and the PubChem compound database [Kim *et al.*, 2023]. The process mirrors the procedures adopted by other DL-based models for k_{cat} prediction [Li *et al.*, 2022a; Qiu *et al.*, 2024]. Following the operations in DLKcat [Li *et al.*, 2022a], redundant entries sharing identical SMILE strings, amino acid sequences, operational temperatures, and k_{cat} values are eliminated from the raw dataset, retaining solely the highest k_{cat} value when other attributes are equivalent. The resulting dataset comprises over 10,000 entries from BRENDA and 4,000 entries from SABIO-RK. To address the uneven distribution of temperature values, an oversampling technique is employed to augment the dataset by duplicating entries at lower and higher temperature ranges [Qiu *et al.*, 2024].

Enzymes exhibit sensitivity to temperature variations. A gradual increment in temperature typically enhances enzymatic reaction rates; however, beyond a certain threshold, further temperature values lead to a decline in reaction speed. This phenomenon arises due to temperature's dual impact on enzymatic reactions, while elevated temperatures can expedite reaction kinetics, they also accelerate enzyme denaturation, thereby diminishing active enzyme concentrations and catalytic efficiency [Arcus and Mulholland, 2020]. Considering the intricate interplay between temperature and k_{cat} in biological reactions, the analysis of Pearson correlations between temperature and k_{cat} (p-value < 0.05) across six enzyme classes, encompassing top-level EC numbers 1-6 including oxidoreductase, transferase, hydrolase, lyases, isomerases, and ligases, is conducted. The correlation analysis, depicted in Figure 3, reveals a positive relationship between temperature and k_{cat} for all enzyme classes except hydrolases, as indicated by the median values.

Following previous works UniKP [Yu *et al.*, 2023] and DLTKcat [Qiu *et al.*, 2024], the resulted dataset is partitioned

into training and testing subsets at a ratio of 90% and 10%, with the training set further segmented into a validation subset comprising one-tenth of the training data. Other experimental details are presented in the appendix.

Baselines. The baselines selected for comparison with our proposed model in predicting k_{cat} can be categorized into two primary groups. The first group comprises machine learning (ML) models, such as Linear Regression [Groß, 2003], Decision Tree [Song and Ying, 2015], AdaBoost [Hastie *et al.*, 2009], Support Vector Regressor [Awad *et al.*, 2015]. The second group consists of DL-based approaches, including CNN [LeCun *et al.*, 1995], RNN [Grossberg, 2013], MLP regressor [Dutt and Saadeh, 2022], these three models are the basic DL networks. Additionally, the specific k_{cat} prediction methods based on DL techniques are considered, including DLKcat [Li *et al.*, 2022a], UniKP [Yu *et al.*, 2023], TurNuP [Kroll *et al.*, 2023], GELKcat [Du *et al.*, 2023], and DLTKcat [Qiu *et al.*, 2024]. Five different initializations are conducted to evaluate these methods. The mean and standard deviation values are reported.

Metrics. To evaluate the efficacy of our proposed model, a comprehensive set of metrics is employed, encompassing the coefficient of determination (R^2), the Pearson correlation coefficient (PCC), the root mean square error (RMSE), and the mean absolute error (MAE) [Yu *et al.*, 2023]. Eq. 10 provided in the appendix elucidates that R^2 signifies the proportion of variance explained, with values ranging between 0 and 1, where a value of 1 denotes a perfect model fit. The PCC quantifies the linear relationship between predicted and actual values, varying from -1 to 1, where 1 indicates a perfect positive linear correlation, -1 represents a perfect negative linear correlation, and 0 signifies no linear association. RMSE serves as a metric to assess the disparities between predicted and observed values. MAE offers an alternative measure of the deviations between predicted and actual outcomes.

4.2 Results of Enzyme Turnover Rates Prediction

Table 1 presents a comparative analysis of model performance in predicting k_{cat} values. The results indicate that ML-based approaches demonstrate competitive performance in terms of RMSE, PCC, MAE, and R^2 when compared to DL-based models, such as CNN and RNN. The limitation observed in the DL models is attributed to their complex network architecture requirements and the challenge posed by the relatively small dataset size, hindering the complete training of DL models. These DL-based methods also have higher standard deviation values.

In contrast, models like UniKP, TurNup, DLTKcat, and our proposed ProKcat-M leverage large-scale pre-trained models' embeddings, leading to improved predictive performance. For instance, UniKP utilizes ProtT5-XL for encoding enzyme sequences and employs a pre-trained LM, SMILES Transformer model, to represent substrate structures, resulting in the second-highest ranking across all performance metrics. Our proposed model, ProKcat-M, incorporates a state-of-the-art protein sequence encoder and features an enzyme-substrate attention module, surpassing all other models with the lowest RMSE, MAE, highest PCC, and R^2 val-

Category	Method	RMSE↓	PCC↑	MAE↓	R^2↑
ML-based	Linear Regression [Groß, 2003]	$1.18_{\pm0.05}$	$0.64_{\pm0.02}$	$0.88_{\pm0.06}$	$0.38_{\pm0.02}$
	Support Vector [Awad *et al.*, 2015]	$1.35_{\pm0.05}$	$0.44_{\pm0.03}$	$1.04_{\pm0.09}$	$0.19_{\pm0.03}$
	Decision Tree [Song and Ying, 2015]	$1.27_{\pm0.07}$	$0.65_{\pm0.03}$	$0.83_{\pm0.01}$	$0.29_{\pm0.01}$
	AdaBoost [Hastie *et al.*, 2009]	$1.34_{\pm0.01}$	$0.48_{\pm0.02}$	$1.07_{\pm0.01}$	$0.21_{\pm0.02}$
DL-based	CNN [LeCun *et al.*, 1995]	$1.42_{\pm0.16}$	$0.34_{\pm0.06}$	$1.11_{\pm0.05}$	$0.10_{\pm0.05}$
	RNN [Grossberg, 2013]	$1.35_{\pm0.12}$	$0.44_{\pm0.05}$	$1.05_{\pm0.06}$	$0.19_{\pm0.10}$
	MLP regressor [Dutt and Saadeh, 2022]	$1.08_{\pm0.08}$	$0.72_{\pm0.05}$	$0.81_{\pm0.04}$	$0.48_{\pm0.04}$
k_{cat} Prediction Models	DLKcat [Li *et al.*, 2022a]	$1.13_{\pm0.15}$	$0.75_{\pm0.06}$	$0.73_{\pm0.08}$	$0.47_{\pm0.11}$
	UniKP [Yu *et al.*, 2023]	$0.82_{\pm0.01}$	$0.85_{\pm0.02}$	$0.58_{\pm0.01}$	$0.67_{\pm0.02}$
	TurNuP [Kroll *et al.*, 2023]	$0.89_{\pm0.01}$	$0.62_{\pm0.04}$	$0.91_{\pm0.03}$	$0.38_{\pm0.04}$
	GELKcat [Du *et al.*, 2023]	$1.00_{\pm0.04}$	$0.78_{\pm0.03}$	$0.69_{\pm0.05}$	$0.58_{\pm0.03}$
	DLTKcat [Qiu *et al.*, 2024]	$0.91_{\pm0.02}$	$0.80_{\pm0.02}$	$0.63_{\pm0.03}$	$0.65_{\pm0.02}$
	ProKcat-M (Proposed)	$\mathbf{0.71}_{\pm0.01}$	$\mathbf{0.88}_{\pm0.02}$	$\mathbf{0.48}_{\pm0.02}$	$\mathbf{0.74}_{\pm0.01}$

Table 1: Performance comparison of different models. The best results are shown in bold.

Method	RMSE↓	PCC↑	MAE↓	R^2 ↑
ProKcat-M	0.71	0.88	0.48	0.74
w/o attention	0.89	0.82	0.56	0.67
w/o CNN	1.04	0.77	0.70	0.56
w/o ESM-2	0.90	0.81	0.65	0.66
w/o enzyme	1.26	0.59	0.96	0.32
w/o substrate	1.17	0.65	0.85	0.41
w/o fingerprint	1.11	0.68	0.78	0.47

Table 2: Ablation of ProKcat-M, we compare it with the models removing the attention module (w/o attention) and the models removing the CNN, ESM-2 embeddings, enzyme sequences (CNN and ESM-2 embeddings), substrate structures (GNN), the fingerprint feature vectors ($\mathbf{h}_f$).

Figure 4: The test performance of ProKcat-M when choosing different latent dimensions d.

ues. These results underscore the superior predictive capacity of ProKcat-M in estimating k_{cat} values.

Ablation Study. Table 2 presents the results of the ablation study on ProKcat-M, comparing it with various models where specific components are removed. ProKcat-M, the full model, achieves promising performance, the subsequent rows in the table represent the performance of models with specific components removed:

- Removing the attention module (w/o attention) leads to a moderate decrease in performance across all metrics.

- Omitting CNN or the ESM-2 embeddings results in a higher RMSE and MAE, lower PCC and R^2, indicating the importance of CNN and EMS-2 embeddings in the model. Both CNN and EMS-2 embeddings are extracted from enzyme sequences, the former is learned on this task, but the latter is trained on large-scale protein sequences in an unsupervised way.

- Removing enzyme sequences (w/o enzyme) significantly deteriorates the model's predictive ability, as evidenced by the substantial increase in RMSE and MAE.

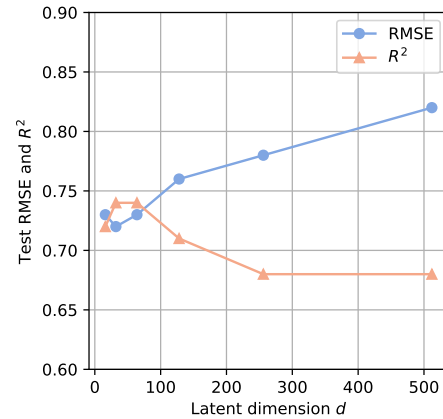- Excluding substrate structures (w/o substrate) and fingerprint feature vectors (w/o fingerprint) also lead to de-

creased performance.

The latent dimension, denoted as d, serves as a critical network hyperparameter. Leveraging the AutoML toolkit NNI [Microsoft, 2021], we conducted a search for the optimal value with search space $\{16, 32, 64, 128, 256\}$. The results are shown in Figure 4, determining that a lower value of $d = 32$ surpasses larger alternatives.

4.3 Results of Symbolic Regression

As depicted in Eq. 8, symbolic regression is performed utilizing a KAN model, where the input to the KAN comprises the concatenated representations $\mathbf{h} \in \mathbb{R}^{3d}$, temperature T, and $\frac{1}{T}$. On this task, many KAN models appear to suffer from overfitting, there exists a substantial disparity between the training and testing metrics. For instance, a 5-depth KAN model with 28K parameters achieves an RMSE of 0.66 on the training set but 0.99 on the test set. Adjustments such as reducing network depth or training duration have been demonstrated to enhance KAN performance. KANs are renowned for their proficiency in regression or PDE solving in mathematical and physical domains. Consequently, addressing the
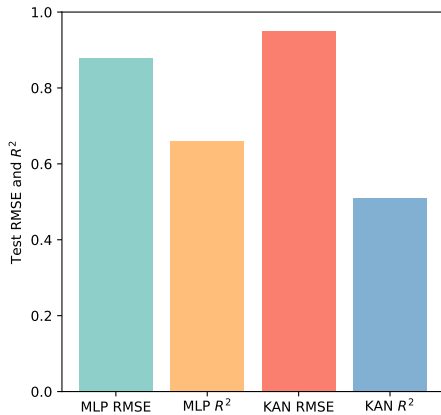
Figure 5: Comparisons of ProKcat-M and ProKcat-K on the same test set in terms of RMSE and R^2.

challenge of overfitting becomes crucial in tackling this intricate, high-dimensional problem using KANs. This outcome underscores the effectiveness of KANs in achieving competitive performance with a relatively modest parameter count. For example, a 5-depth KAN model with 28K parameters attains an R^2 of 0.41, whereas a 3-depth KAN model with 25K parameters achieves an R^2 of 0.50.

Beyond mitigating the issue of overfitting in KANs, our main aim is to derive a comprehensive formula to establish the relationships between input and output variables. Given that the embeddings $\mathbf{h}$ originate from a deep neural network, our focus centers on explicitly modeling the correlations between T, $\frac{1}{T}$, and k_{cat}. To derive an explicit and concise equation, we perform a linear projection on the feature vectors $\mathbf{h}_p'' : \mathbb{R}^d \to \mathbb{R}^1, \mathbf{h}_c'' : \mathbb{R}^d \to \mathbb{R}^1, \mathbf{h}_f' : \mathbb{R}^d \to \mathbb{R}^1$, resulting in the concatenated feature vectors $\mathbf{h} : \mathbb{R}^{3d} \to \mathbb{R}^3$.

In our experiments, we observe that by reducing the dimensionality of the combined vector ($\mathbf{h}$) to 3, a streamlined 2-layer KAN model can be developed. Specifically, utilizing B-spline details with 5 intervals, order 3, and steps 5. To ensure a fair comparison between KAN and MLP, we implement a 2-layer MLP with the latent dimension ranging from $3d + 2$ to 1. Both the 2-layer KAN model and the 2-layer MLP model possess nearly equivalent trainable parameters, approximately 0.1K. From Figure 5, we observe that the 2-layer KAN model yields results comparable to a 2-layer MLP. Notably, the 2-layer KAN model demonstrates higher efficiency, with an inference time of 1 ms (millisecond) per sample, in contrast to 3.6 ms per sample for the 2-layer MLP.

According to the Arrhenius equation (Eq. 1), a significant linear correlation is observed between the natural logarithm of the rate constant ($\ln k$) and the reciprocal of temperature ($\frac{1}{T}$), indicating a predominantly empirical relationship. It is noted that in practical scenarios, the relationships are more complicated with data noise existed. The symbolic regression-derived function of the 2-layer KAN model (ProKcat-K) is expressed as:

$$0.02 * |9.92 * \mathbf{h}_p'' - 1.29| - 0.09 * |3.57 * \mathbf{h}_c'' - 0.21| +$$
$$0.01 * e^{3.57 * \mathbf{h}_f'} - 0.03 * |9.22 * \frac{1}{T} - 3.71| -$$
$$0.02 * e^{-9.22 * T} + 0.21$$

$$(9)$$

The equation reveals a linear association between $\frac{1}{T}$ and $\log_{10} k_{\text{cat}}$, aligning with the principles of Eq. 1. This underscores the reliability of our learned regression function to a certain degree. When given the pre-trained representations and temperature inputs, the direct calculation of $\log_{10} k_{\text{cat}}$ can be achieved. It is the first time to derive such an equation in the DL-based k_{cat} prediction field, which is promising.

5 Conclusion

This paper presents a novel multimodal framework that integrates enzyme sequences, substrate compound structures, and additional features using a pre-trained language model LM, a convolutional neural network, and a graph neural network to predict k_{cat} values. The proposed enzyme-substrate attention module effectively learns attention weights by capturing relationships between sequence and atomic-level features. Recognizing the critical importance of enzyme–compound interactions, the ProKcat-M model achieves superior predictive performance compared to existing baselines. Furthermore, the ProKcat-K variant employs the Kolmogorov–Arnold Network architecture to perform accurate k_{cat} predictions with lower inference time, while also yielding an explicit equation that relates input variables to the output. However, a key limitation of the KAN model is its sensitivity to overfitting, which necessitates careful architectural and regularization design.

Contribution Statement

Bozhen Hu and Cheng Tan contributed equally to this work. They jointly designed the model, conducted the experiments, and co-authored the manuscript. Siyuan Li and Jiangbin Zheng provided conceptual guidance. Jun Xia and Stan Z. Li reviewed and revised the manuscript. We also thank Sizhe Qiu, a Ph.D. student in the Department of Engineering Science at the University of Oxford, for his valuable insights and biological expertise.

References

[Arcus and Mulholland, 2020] Vickery L Arcus and Adrian J Mulholland. Temperature, dynamics, and enzyme-catalyzed reaction rates. *Annual review of biophysics*, 49:163–180, 2020.

[Arroyo *et al.*, 2022] José Ignacio Arroyo, Beatriz Díez, Christopher P Kempes, Geoffrey B West, and Pablo A Marquet. A general theory for temperature dependence in biology. *Proceedings of the National Academy of Sciences*, 119(30):e2119872119, 2022.

[Awad *et al.*, 2015] Mariette Awad, Rahul Khanna, Mariette Awad, and Rahul Khanna. Support vector regression. *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, pages 67–80, 2015.

[Bento *et al.*, 2020] A Patrícia Bento, Anne Hersey, Eloy Félix, Greg Landrum, Anna Gaulton, Francis Atkinson, Louisa J Bellis, Marleen De Veij, and Andrew R Leach. An open source chemical structure curation pipeline using rdkit. *Journal of Cheminformatics*, 12:1–16, 2020.

[Boorla and Maranas, 2024] Veda Sheersh Boorla and Costas D Maranas. Catpred: A comprehensive framework for deep learning in vitro enzyme kinetic parameters kcat, km and ki. *bioRxiv*, pages 2024–03, 2024.

[Braun and Griebel, 2009] Jürgen Braun and Michael Griebel. On a constructive proof of kolmogorov's superposition theorem. *Constructive approximation*, 30:653–675, 2009.

[Brown *et al.*, 2004] James H Brown, James F Gillooly, Andrew P Allen, Van M Savage, and Geoffrey B West. Toward a metabolic theory of ecology. *Ecology*, 85(7):1771–1789, 2004.

[Du *et al.*, 2023] Bing-Xue Du, Haoyang Yu, Bei Zhu, Yahui Long, Min Wu, and Jian-Yu Shi. Gelkcat: An integration learning of substrate graph with enzyme embedding for kcat prediction. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 408–411. IEEE, 2023.

[Dutt and Saadeh, 2022] Muhammad Ibrahim Dutt and Wala Saadeh. A multilayer perceptron (mlp) regressor network for monitoring the depth of anesthesia. In *2022 20th IEEE Interregional NEWCAS Conference (NEW-CAS)*, pages 251–255. IEEE, 2022.

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

[Fan, 2000] Engui Fan. Extended tanh-function method and its applications to nonlinear equations. *Physics Letters A*, 277(4-5):212–218, 2000.

[Groß, 2003] Jürgen Groß. *Linear regression*, volume 175. Springer Science & Business Media, 2003.

[Grossberg, 2013] Stephen Grossberg. Recurrent neural networks. *Scholarpedia*, 8(2):1888, 2013.

[Hastie *et al.*, 2009] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009.

[Hollingsworth and Dror, 2018] Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, 2018.

[Honda *et al.*, 2019] Shion Honda, Shoi Shi, and Hiroki R Ueda. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*, 2019.

[Hu *et al.*, 2024] Bozhen Hu, Cheng Tan, Yongjie Xu, Zhangyang Gao, Jun Xia, Lirong Wu, and Stan Z Li. Protgo: Function-guided protein modeling for unified representation learning. *Advances in Neural Information Processing Systems*, 37:88581–88604, 2024.

[Joulin *et al.*, 2017] Armand Joulin, Moustapha Cissé, David Grangier, Hervé Jégou, et al. Efficient softmax approximation for gpus. In *International conference on machine learning*, pages 1302–1310. PMLR, 2017.

[Kim *et al.*, 2023] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2023 update. *Nucleic acids research*, 51(D1):D1373–D1380, 2023.

[Kosal, 2023] Erica Kosal. Digestion. *Introductory Biology: Ecology, Evolution, and Biodiversity*, 2023.

[Kroll *et al.*, 2023] Alexander Kroll, Yvan Rousset, Xiao-Pan Hu, Nina A Liebrand, and Martin J Lercher. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *Nature Communications*, 14(1):4139, 2023.

[LeCun *et al.*, 1995] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[Lee *et al.*, 2019] Ingoo Lee, Jongsoo Keum, and Hojung Nam. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15(6):e1007129, 2019.

[Li *et al.*, 2022a] Feiran Li, Le Yuan, Hongzhong Lu, Gang Li, Yu Chen, Martin KM Engqvist, Eduard J Kerkhoven, and Jens Nielsen. Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis*, 5(8):662–672, 2022.

[Li *et al.*, 2022b] Min Li, Zhangli Lu, Yifan Wu, and Yao-Hang Li. Bacpi: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction. *Bioinformatics*, 38(7):1995–2002, 2022.

[Lim *et al.*, 2021] Sangsoo Lim, Yijingxiu Lu, Chang Yun Cho, Inyoung Sung, Jungwoo Kim, Youngkuk Kim, Sungjoon Park, and Sun Kim. A review on compound-protein interaction prediction methods: data, format, rep-

resentation and model. *Computational and Structural Biotechnology Journal*, 19:1541–1556, 2021.

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

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

[Liu *et al.*, 2023] Yue Liu, Xihong Yang, Sihang Zhou, Xinwang Liu, Zhen Wang, Ke Liang, Wenxuan Tu, Liang Li, Jingcan Duan, and Cancan Chen. Hard sample aware network for contrastive deep graph clustering. In *Proc. of AAAI*, 2023.

[Liu *et al.*, 2024] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.

[Magrane and Consortium, 2011] Michele Magrane and UniProt Consortium. Uniprot knowledgebase: a hub of integrated protein data. *Database*, 2011:bar009, 2011.

[Michaelis *et al.*, 1913] Leonor Michaelis, Maud L Menten, et al. Die kinetik der invertinwirkung. *Biochem. z*, 49(333-369):352, 1913.

[Microsoft, 2021] Microsoft. Neural Network Intelligence, 1 2021.

[Nguyen *et al.*, 2023] Ngoc-Quang Nguyen, Gwanghoon Jang, Hajung Kim, and Jaewoo Kang. Perceiver cpi: a nested cross-attention network for compound–protein interaction prediction. *Bioinformatics*, 39(1):btac731, 2023.

[Nilsson *et al.*, 2017] Avlant Nilsson, Jens Nielsen, and Bernhard O Palsson. Metabolic models of protein allocation call for the kinetome. *Cell Systems*, 5(6):538–541, 2017.

[Paul *et al.*, 2022] Ashis Paul, Rajarshi Bandyopadhyay, Jin Hee Yoon, Zong Woo Geem, and Ram Sarkar. Sinlu: Sinu-sigmoidal linear unit. *Mathematics*, 10(3):337, 2022.

[Prokop *et al.*, 2012] Zbynek Prokop, Artur Gora, Jan Brezovsky, Radka Chaloupkova, Veronika Stepankova, and Jiri Damborsky. Engineering of protein tunnels: Keyhole-lock-key model for catalysis by the enzymes with buried active sites. *Protein engineering handbook*, 3:421–464, 2012.

[Qiu *et al.*, 2024] Sizhe Qiu, Simiao Zhao, and Aidong Yang. Dltkcat: deep learning-based prediction of temperature-dependent enzyme turnover rates. *Briefings in Bioinformatics*, 25(1):bbad506, 2024.

[Rogers and Hahn, 2010] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[Schomburg *et al.*, 2017] I Schomburg, L Jeske, M Ulbrich, S Placzek, A Chang, and D Schomburg. The brenda enzyme information system–from a database to an expert system. *Journal of biotechnology*, 261:194–206, 2017.

[Sharaff and Gupta, 2019] Aakanksha Sharaff and Harshil Gupta. Extra-tree classifier with metaheuristics approach for email classification. In *Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2018*, pages 189–197. Springer, 2019.

[Song and Ying, 2015] Yan-Yan Song and LU Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[Tikhomirov, 1991] VM Tikhomirov. On the representation of continuous functions of several variables as superpositions of continuous functions of one variable and addition. In *Selected Works of AN Kolmogorov*, pages 383–387. Springer, 1991.

[Trott and Olson, 2010] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

[Vaca-Rubio *et al.*, 2024] Cristian J Vaca-Rubio, Luis Blanco, Roberto Pereira, and Màrius Caus. Kolmogorov-arnold networks (kans) for time series analysis. *arXiv preprint arXiv:2405.08790*, 2024.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wittig *et al.*, 2018] Ulrike Wittig, Maja Rey, Andreas Weidemann, Renate Kania, and Wolfgang Müller. Sabio-rk: an updated resource for manually curated biochemical reaction kinetics. *Nucleic acids research*, 46(D1):D656–D660, 2018.

[Wu *et al.*, 2024] Lirong Wu, Yufei Huang, Cheng Tan, Zhangyang Gao, Bozhen Hu, Haitao Lin, Zicheng Liu, and Stan Z Li. Psc-cpi: Multi-scale protein sequence-structure contrasting for efficient and generalizable compound-protein interaction prediction. In *The 38th Annual AAAI Conference on Artificial Intelligence*, 2024.

[Yu *et al.*, 2023] Han Yu, Huaxiang Deng, Jiahui He, Jay D Keasling, and Xiaozhou Luo. Unikp: a unified framework for the prediction of enzyme kinetic parameters. *Nature Communications*, 14(1):8211, 2023.

[Zhao *et al.*, 2022] Qichang Zhao, Haochen Zhao, Kai Zheng, and Jianxin Wang. Hyperattentiondti: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics*, 38(3):655–662, 2022.

[Zhou *et al.*, 2020] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.