DeepEnzyme: a robust deep learning model for improved enzyme turnover number prediction by utilizing features of protein 3D Structures

Tong Wang[1,2], Guangming Xiang[1], Siwei He[1], Liyun Su[2], Xuefeng Yan[3,4], Hongzhong Lu*[1]

1. State Key Laboratory of Microbial Metabolism, School of Life Science and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China
2. College of Science, Chongqing University of Technology, Chongqing, 400054, China
3. Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai, China
4. State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai, China

*To whom correspondence should be addressed.

[#] Correspondence to: hongzhonglu@sjtu.edu.cn

Abstract

Turnover numbers (k_{cat}), which indicate an enzyme's catalytic efficiency, have a wide range of applications in fields including protein engineering and synthetic biology. Experimentally measuring the enzymes' k_{cat} is always time-consuming. Recently, the prediction of k_{cat} using deep learning models has mitigated this problem. However, the accuracy and robustness in k_{cat} prediction still needs to be improved significantly, particularly when dealing with enzymes with low sequence similarity compared to those within the training dataset. Herein, we present DeepEnzyme, a cutting-edge deep learning model that combines the most recent Transformer and Graph Convolutional Network (GCN) architectures. To improve the prediction accuracy, DeepEnzyme was trained by leveraging the integrated features from both sequences and 3D structures. Consequently, our model exhibits remarkable robustness when processing enzymes with low sequence similarity compared to those in the training dataset by utilizing additional features from high-quality protein 3D structures. DeepEnzyme also makes it possible to evaluate how point mutations affect the catalytic activity of the enzyme, which helps identify residue sites that are crucial for the catalytic function. In summary, DeepEnzyme represents a pioneering effort in predicting enzymes' k_{cat} values with superior accuracy and robustness compared to previous algorithms. This advancement will significantly contribute to our comprehension of enzyme function and its evolutionary patterns across species.

Key words:

Enzyme Turnover Number Prediction, Deep Learning, Protein 3D Structure, Graph

Convolutional Network

Introduction

The enzyme turnover number (k_{cat}) represents the maximum number of substrate molecules that an enzyme can convert into a product per unit time under saturating conditions[1,2]. Presently, accurately predicting k_{cat} has become imperative for various applications, including protein engineering and enzyme design[3]. Additionally, estimating enzyme k_{cat} values at genome-scale is crucial for constructing sophisticated metabolic models to capture the correlations between genotypes and phenotypes[4,5]. Despite the abundant k_{cat} values available in the BRENDA[6] and SABIO-RK[7] databases, the total number of enzymes with experimentally determined k_{cat} remains substantially smaller than the vast number of all sequenced proteins. For instance, less than 1% of the enzymes listed in the UniProt database possess experimentally ascertained k_{cat} values[8].

Recently, it witnessed the breakthroughs in development of various advanced deep learning models, which make it possible to predict protein structure[9], protein function[10], and gene expression level[11] only from primary sequence features. As for genome-scale k_{cat} prediction, Heckmann et al.[12] firstly utilized machine learning models, i.e. the linear elastic net, the decision-tree-based random forest model, to infer the enzyme k_{cat} for the model organism *Escherichia coli*. They discovered a correlation between k_{cat} and specific enzyme structural features. While those classical machine learning models could successfully predict k_{cat} parameters at the genome-scale[12], its application is only suitable for *E. coli*. Subsequently, the deep learning model-DLKcat, proposed by Li et al.[13], significantly broadens the scope of k_{cat} prediction for nearly all sequenced enzymes with catalyzed substrates. By incorporating features from both the protein sequence and its substrate, DLKcat facilitated high-throughput prediction of enzyme k_{cat} and the automatic generation of genome-scale k_{cat} profiles. To further enhance the performances of DLkcat, TurNuP integrated comprehensive biochemical reaction information, encompassing fingerprint features from substrates and products, and effectively predicted the k_{cat} of wild-type enzymes[14]. It outperforms DLKcat in predicting k_{cat} for enzymes with low sequence similarity compared to those in the training dataset. However, there remains room for improvement in the prediction accuracy of TurNuP, given that its R^2 in test datasets is approximately 0.42, indicating that only a small portion of the variance in k_{cat} prediction can be explained by this model.

To larger extent, the function of a protein is determined by its 3D structure[15]. The structural data provide crucial insights into the spatial arrangement of pivotal functional residues and the interactions between the enzyme and its substrate[16,17]. The composition of amino acids and the distinct folding pattern within the protein structure can significantly influence the accessibility of active sites, substrate binding sites, and the overall stability of the enzyme[18]. These factors collectively impact enzyme function and its catalytic efficiency. Consequently, features derived from protein 3D structures have garnered special attention in the development of advanced deep learning models. These

sophisticated models can now accurately predict how amino acid mutations affect the thermal stability of enzymes[19], as well as identifying the binding sites of ligands on the 3D structure of proteins[20]. Recent breakthroughs in protein structure prediction, exemplified by the release of AlphaFold2[9], RoseTTAFold[21], and ColabFold[22], have substantially reduced the cost in acquiring high-quality protein structures. Thus, it is feasible to build large-scale structure datasets for the training of advanced deep learning models. However, despite this progress, the invaluable protein 3D structural information spanning various species has not been systematically utilized in enzyme k_{cat} prediction.

Here, we present a novel deep learning model, named as DeepEnzyme, to enhance the prediction accuracy of enzyme k_{cat} based on protein 3D structures. DeepEnzyme combines Graph Convolutional Networks (GCN)[23], a proven approach for extracting structural properties from proteins[24,25], with Transformers[26], which have demonstrated exceptional performance in protein language modeling[27,28]. This fusion allows DeepEnzyme to extract both structure and sequence features from enzymes and substrates. Compared with existing models, DeepEnzyme exhibits superior accuracy and robustness in predicting k_{cat} values. Therefore, DeepEnzyme has the potential to expedite the functional analysis of previously unstudied enzymes and provide opportunities for the rational optimization of enzyme activities in conjunction with advanced molecular technologies.

Results

Framework of a novel deep learning model for predicting k_{cat}

The novel deep learning model for k_{cat} prediction, DeepEnzyme, was developed by integrating three layers of features extracted respectively from protein 1D sequence, 3D structure, and the related substrate, utilizing distinct computation modules (Fig. 1). Specifically, an excellent Transformer model was employed to extract features from the protein 1D sequence. Simultaneously, the 3D structure underwent transformation into a contact map employing the $C_\alpha - C_\alpha < 10\text{Å}$ criteria. Subsequently, GCN was applied to extract structural features from the protein based on the contact map. Concurrently, the RDKit software[29] was employed to extract crucial substrate features, including fingerprints and adjacency matrices, based on its SMILES formula. To delve deeper into substrate features, the fingerprints and adjacency matrices generated by RDKit served as the input for a dedicated GCN module. Following this, all feature vectors obtained from both the enzyme and substrate were merged in the subsequent phase. This process resulted in the creation of a comprehensive embedding vector, encapsulating vital features from both the enzyme and substrate. Finally, the k_{cat} value was predicted employing a neural attention method based on the representation vectors. This attention mechanism enabled the model to highlight the most crucial hidden features, ensuring accurate predictions by integrating information from both enzymes and substrates.

Since a significant portion of enzymes in our dataset lacked experimentally

determined 3D structures, ColabFold[22] was utilized to predict the structures for all these proteins. The average predicted local distance difference test (pLDDT) score for all predicted protein 3D structures in this study was 92.67 (Fig. 1), guaranteeing that the structures used for our model training are of high-quality.

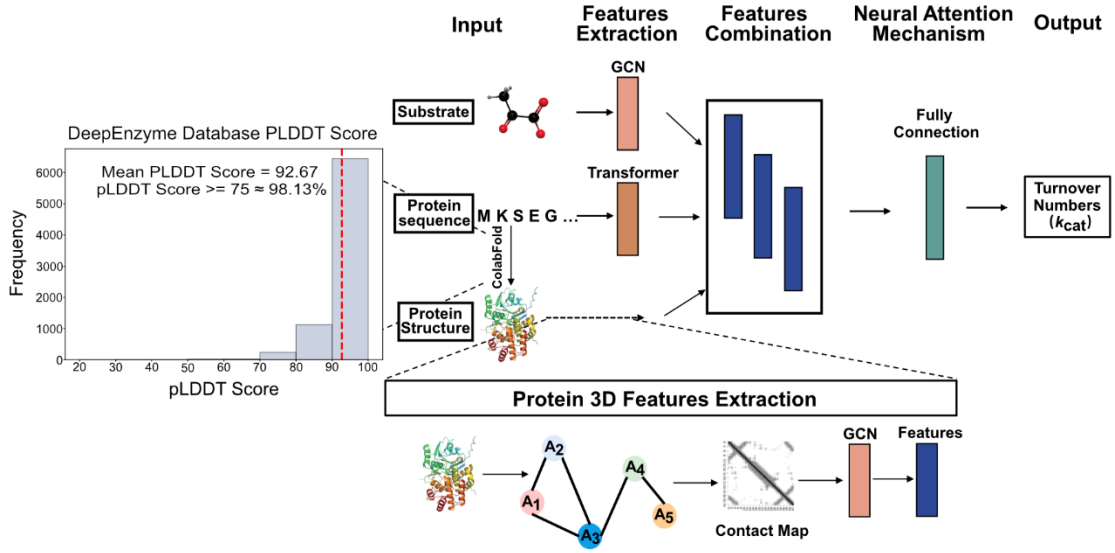


Fig. 1 The framework of DeepEnzyme for k_{cat} prediction. DeepEnzyme integrates Transformer and GCN models to distill features from both the enzyme and substrate for predicting k_{cat}. Here, GCN is employed to extract structural features based on protein 3D structures and substrate adjacency matrixes; Transformer is utilized to extract sequence features from protein sequences. ColabFold, a protein structure prediction tool, is employed to predict protein 3D structure. The prediction results indicate a mean pLDDT score of 92.67, with 98% of the predicted structures achieving a pLDDT score greater than or equal to 75.

DeepEnzyme exhibits excellent predictive accuracy for k_{cat} values.

Previously, it is shown that the high similarity in protein sequences between the training, validation and testing datasets can potentially result into overfitting and poor generalization of k_{cat} value prediction[14]. To mitigate this challenge, we preprocessed the data (Methods) to exclude highly similar sequences existing in the original DLKcat dataset[14,30]. Consequently, out of the 16,838 distinct enzyme-substrate pairs, we carefully selected and retained 11,923 of all pairs, enabling us to reconstruct a dataset that could, to some extent, avoid the too high similarity in sequences. This dataset was then randomly partitioned into training, validation, and testing datasets with a ratio at 80%, 10%, and 10%, respectively. This preprocess of enzymatic dataset allowed us to train and evaluate DeepEnzyme effectively. By mitigating the issue of sequence similarity and incorporating additional protein structure information, DeepEnzyme achieved an R^2 of 0.60 (Fig. 2a) on the test dataset after dedicated training phase, indicating that the model can explain 60% of the variance in the target variable. Furthermore, the Root Mean Square Error (RMSE) of 0.95 suggests that the model has relatively low prediction errors.

To evaluate the performance of DeepEnzyme, the collected k_{cat} values for enzymes from diverse sources and classifications were utilized, respectively. Initially, enzymes were categorized into two main types: wild-type and mutant, based on their records in the public databases. The results indicated that, for mutant enzymes, the R^2 value was 0.70 (Fig. 2c), whereas for wild-type enzymes, it was only 0.45 (Fig. 2b). Notably, mutant enzymes constituted approximately 59% of all enzymes, suggesting that the prediction accuracy might be influenced by the quantity of training dataset. Next, all enzymes were classified based on the first digit encompassing the corresponding EC number. We found that DeepEnzyme exhibits different predictive capabilities for enzymes across EC numbers. For example, the R^2 for enzymes from the EC 1 group and the EC 2 group was notably higher (Fig. 2d). This observation is consistent with the fact that the percentages of enzymes from these two groups in the dataset (37%, 26%) are significantly higher than the other enzyme groups, underscoring how the number of enzymes in each group can affect the model prediction capability. Furthermore, to illustrate the enhancement in k_{cat} prediction achieved by incorporating protein structure, we compared the performance of DeepEnzyme using different types of datasets as input (Fig. 2e). The average coefficient of determination (R^2) value on the testing dataset from 5 rounds of training was 0.58 (Fig. 2f). The results unequivocally demonstrated that the inclusion of structural data significantly improved prediction accuracy. This suggests that the features extracted from protein 3D structures by DeepEnzyme play a crucial role in characterizing enzyme function evolution.

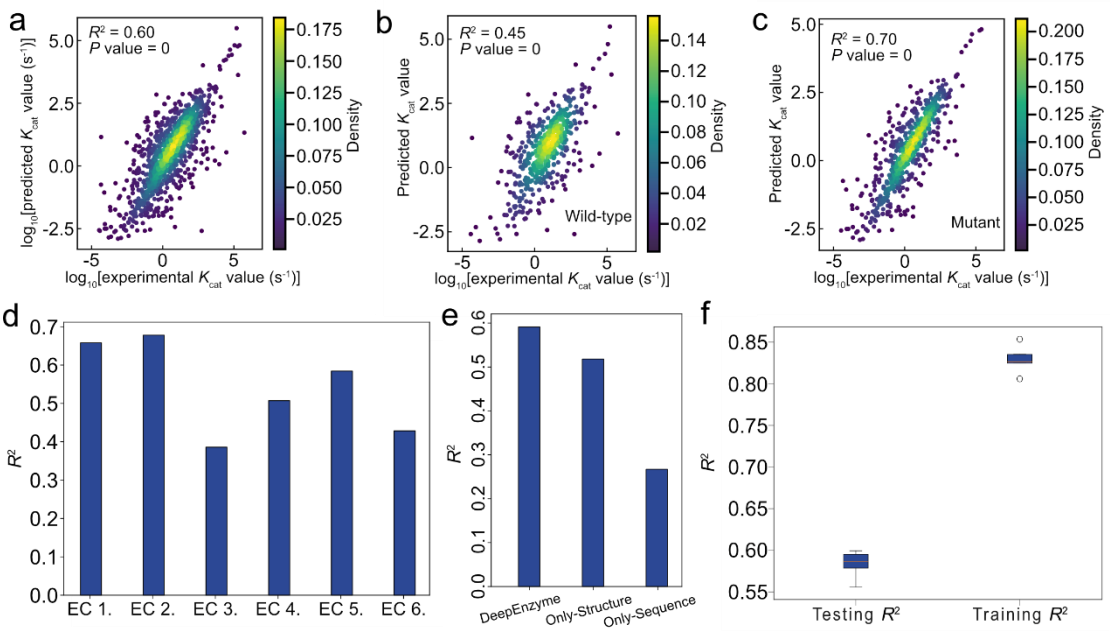


Fig. 2 DeepEnzyme performance for k_{cat} prediction. (a) The performance of DeepEnzyme on the test dataset as evaluated by the R^2 and P values between predicted and experimental k_{cat} values. (b-c) DeepEnzyme prediction performance on wild-type (b) and mutant (c) enzymes in the test dataset. (d) DeepEnzyme prediction performance for enzymes classified by different EC numbers in the test dataset. (e) The performance of the predictive ability of the model under different inputs. DeepEnzyme: enzyme sequence, enzyme structure, and substrate information are used as inputs for deep learning model training and testing; Only-structure: substrate

information and enzyme structure information as inputs; Only-sequence: substrate and enzyme sequence as inputs. (a-e) are from specific seed. (f) Comparison in average coefficient of determination (R^2) values for testing dataset and training dataset from 5 rounds of training.

DeepEnzyme has better performance k_{cat} prediction compared with existing models

To comprehensively compare DeepEnzyme with existing k_{cat} prediction tools, two of the latest publicly available models with accessible scripts, TurNuP[14] and DLKcat[13], were selected and evaluated together with DeepEnzyme. Utilizing only protein structure and catalytic substrate data as inputs, DeepEnzyme achieved an R^2 value of 0.52 on the test dataset, representing a 30% and 21% improvement compared to TurNuP and DLKcat, respectively (Fig. 3a). Moreover, when utilizing both protein sequence and structure as input, DeepEnzyme exhibited significantly superior prediction accuracy compared to TurNuP and DLKcat. In addition, the RMSE value was 0.95, which was lower than that of DLKcat but slightly higher than TurNuP (Fig. 3b). This difference in RMSE value could be attributed to the training approach of TurNuP, where the authors excluded enzyme reaction combinations with extremely low or high k_{cat} values, contributing to a reduction in the final RMSE value of the model[14]. Overall, these results underscore that DeepEnzyme has strong performance in k_{cat} prediction when utilizing protein structure as input, emphasizing the crucial role of protein 3D structure data in accurate k_{cat} prediction.

Sequence similarity exerts a notable influence on the performance of deep learning models in predicting enzyme function, particularly in predicting the substrates catalyzed by specific enzymes. This impact has been thoroughly investigated, as previously discussed by Alexander Kroll et al[30]. Notably, the prediction accuracy of a deep learning model can significantly diminish when the protein sequences in the test dataset exhibit lower similarity to those in the training dataset. To assess the performance of DeepEnzyme at various levels of sequence similarity, we employed the MMseqs2[31] method to compute sequence similarity between the test and training datasets. Subsequently, based on the calculated similarity values, the test dataset was divided into three distinct groups: 0–40%, 40–80%, and 80–100%. We then calculated the R^2 value for each group (Fig. 3c). Remarkably, DeepEnzyme demonstrated a noteworthy R^2 value of 0.55 even when the sequence similarity was only within the 0–40% range, underscoring its robustness in k_{cat} prediction under conditions of lower sequence similarity.

In a detailed comparison of the predicted performance of TurNuP, DLKcat, and DeepEnzyme on enzymes with varying levels of sequence similarity (ranging from 80–100% to 0–40%), it was observed that TurNuP experienced a substantial 47% decrease in R^2, whereas the corresponding R^2 remained relatively stable for DeepEnzyme (Fig. 3c). This disparity can be attributed to DeepEnzyme effectively incorporating enzyme structure information, which is strongly correlated with enzyme function. In reality, homologous enzymes can exhibit higher conservation in 3D structures while displaying

divergent evolutionary paths in terms of sequences. For instance, an examination of the entire k_{cat} dataset highlighted this phenomenon with the enzyme having EC number 1.3.3.4, the final common enzyme in the chlorophyll and heme biosynthesis, from different species (*Myxococcus xanthus* and *Bacillus subtilis*). Calculated by Foldseek and US-align[32,33], it revealed that the structural similarity between those two enzymes is as high as 0.88 (Fig. 3d), although their sequence similarity is only 0.27 (Fig. 3e). These observations collectively confirm that DeepEnzyme maintains robust prediction accuracy across a broad spectrum of sequence similarities, showcasing its ability to effectively leverage 3D structural information for improved predictions.

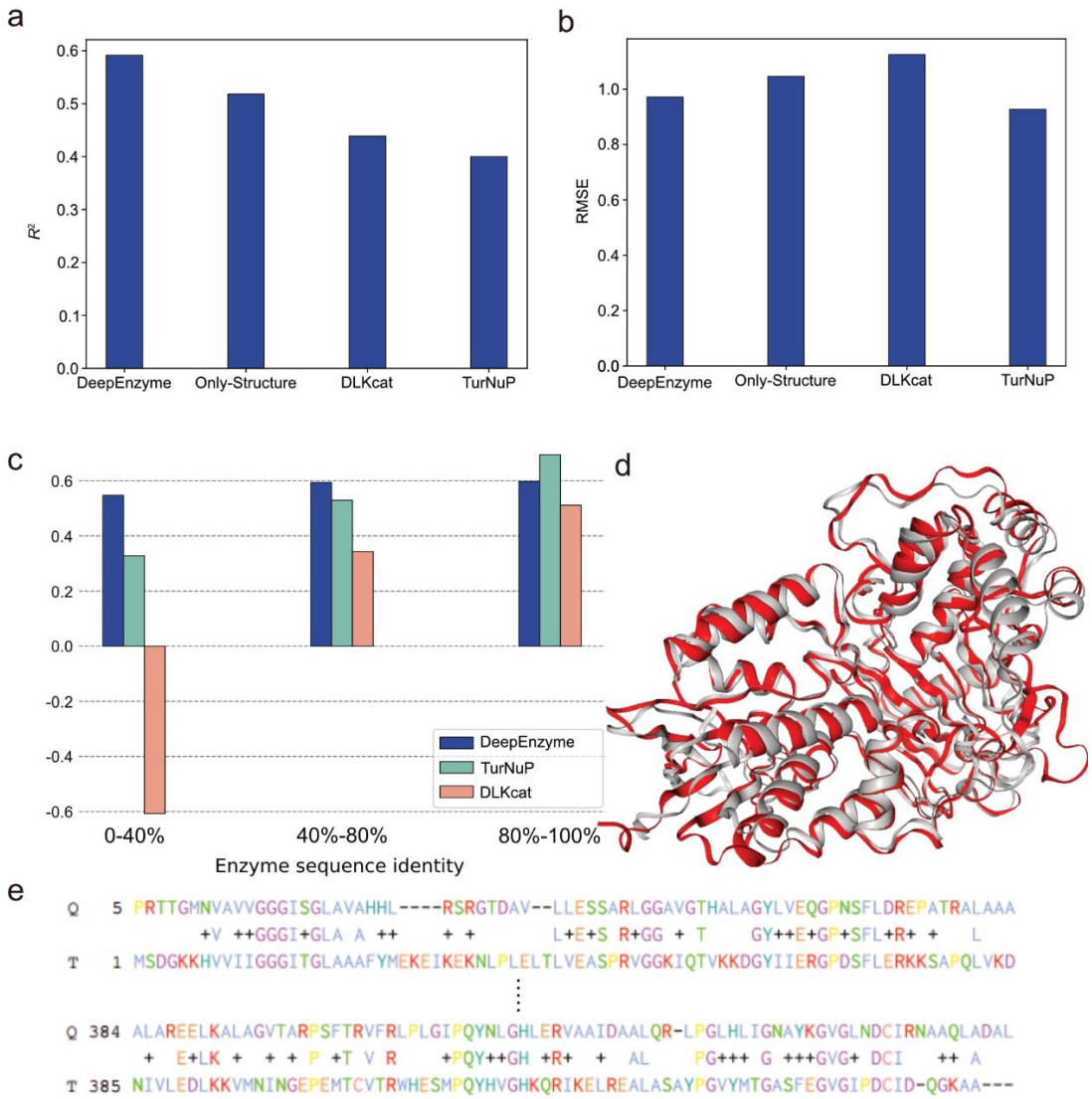


Fig. 3 DeepEnzyme has demonstrated superior predictive capabilities for k_{cat} values compared to existing models, even under lower sequence identification. (a) Comparison of R^2 values on the test dataset for different models. (b) Comparison of RMSE values on the test dataset for different models. (c) The determination coefficient R^2 of k_{cat} value predictions for enzymes in the test dataset at different levels of sequence similarity is compared among the DeepEnzyme, TurNuP, and DLKcat models. (d) Two enzymes from *Myxococcus xanthus* and *Bacillus subtilis*, both with EC numbers 1.3.3.4, are highly similar in structure (TM-Score = 0.8762), gray for enzyme from *Myxococcus xanthus* and red for enzyme from *Bacillus subtilis*.

(e) The similarity of the amino acid sequences of the two enzymes is 27% (Q for enzyme from *Myxococcus xanthus*, T for enzyme from *Bacillus subtilis*).

Test the prediction performance of DeepEnzyme using enzyme saturation mutation datasets

To validate the capability of DeepEnzyme to discern and identify missense and nonsense variants, we conducted prediction experiments using specific variants of CYP2C9, a critical enzyme in drug metabolism and personalized therapy[34-37]. Based on the enzyme activity, the variants were categorized into three groups: missense variants, synonymous variants, and nonsense variants. We utilized the well-trained DeepEnzyme to predict and compare the k_{cat} values of missense variants and nonsense variants (no comparison was conducted for synonymous variants as they do not undergo sequence alterations). The results clearly demonstrated that the predicted k_{cat} values for nonsense variants (median: k_{cat} =1.60s^{-1}) were lower than those of missense variants (median: k_{cat} = 1.97s^{-1}) (Fig. 4a). Additionally, a significant difference was observed between the two groups (P value = 1.6×10^{-94}). Interestingly, such a result is consistent with experimental evidences represented by the activity score (Fig. 4b). These findings affirm a distinct disparity in predicted k_{cat} values between missense and nonsense variants, underscoring DeepEnzyme's potential to effectively distinguish and identify these types of variants.

To evaluate the generalization ability of DeepEnzyme in predicting k_{cat} values for enzyme saturation point mutations, we utilized the model to predict k_{cat} values for a comprehensive set of phosphate-irrepressible alkaline phosphatase of Flavobacterium (PafA) mutantions[38] and rigorously examined the results. We categorized the PafA mutations into two distinct categories: high k_{cat} mutations and low k_{cat} mutations, representing mutations with higher and lower k_{cat} values compared to the wild-type enzyme (Fig. 4d), respectively. Our findings (Fig. 4c) provided compelling insights. The median predicted k_{cat} values for the high k_{cat} mutations were observed to be 15% higher than those for the low k_{cat} mutations, with a mean increase of 11% (P value = 0.0033). These results underscore the potential of DeepEnzyme in predicting the turnover numbers for enzymes with large-scale saturation point mutations.

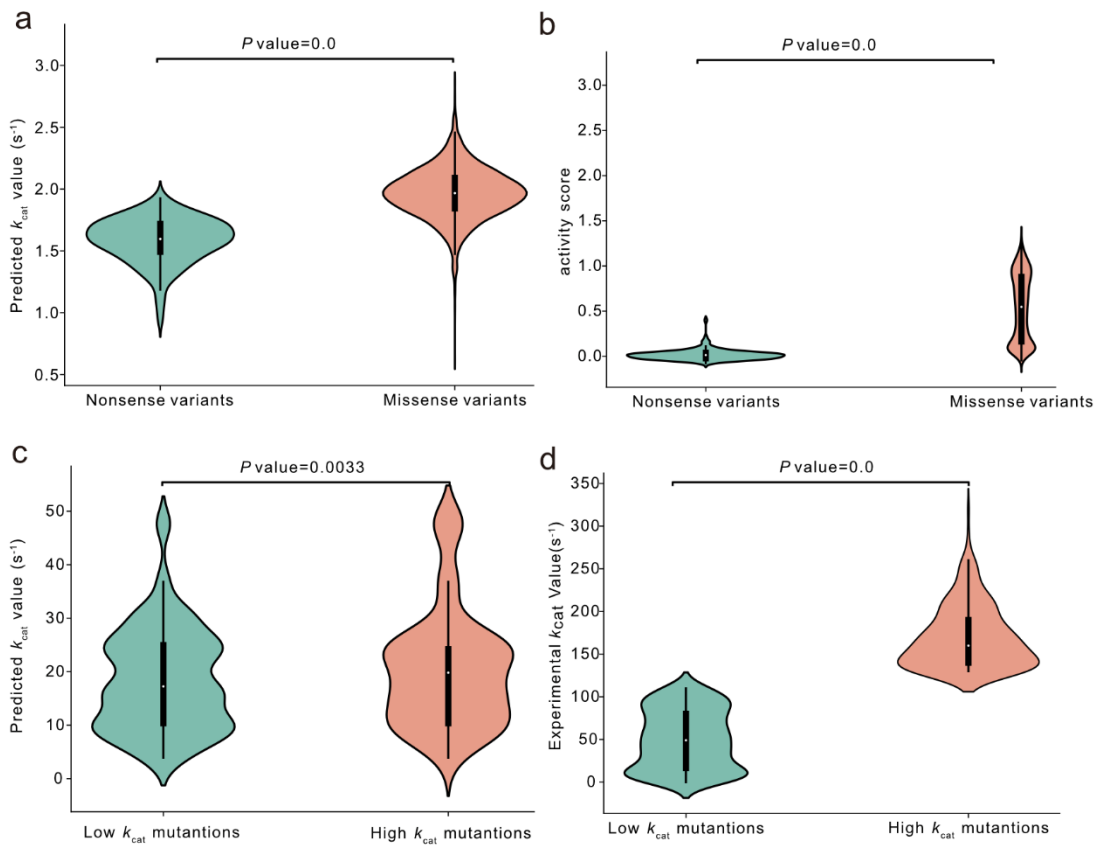


Fig. 4 Analysis of the prediction ability of DeepEnzyme for enzyme saturation point mutations. (a) Comparison of predicted results for different CYP2C9 variants: red for missense variants, green for nonsense variants, indicating inactive variants and variants that are still active. (b) Comparison of experimental activity score for different CYP2C9 variants: red for missense variants, green for nonsense variants, indicating inactive variants and variants that are still active. (c) Comparison of predicted results for different PafA mutations: green for low k_{cat} mutations, red for high k_{cat} mutations. (d) Comparison of experimental results for different PafA mutations: green for low k_{cat} mutations, red for high k_{cat} mutations.

DeepEnzyme could identify key residue sites impacting enzyme activity

The active site of an enzyme, characterized by specific amino acid residues in the protein structure, is pivotal for enzyme catalysis. Likewise, the creation of enzyme-substrate complexes requires particular binding sites in the protein structure to effectively interact with the substrate[39]. In this context, we sought to investigate whether DeepEnzyme possesses the capability to identify the importance of active sites and binding sites within enzymes. As the first example, we examined the case of PafA mentioned earlier. Leveraging site annotation data from the UniProt database, all residue sites in PafA were categorized into "Binding/Active sites" and "General sites." The structure vectors computed by DeepEnzyme were derived, followed by normalization using min-max normalization to obtain the weights for each residue. In addition to PafA, P00558 in the human EMP pathway was also used to validate DeepEnzyme performance. The results clearly demonstrated that the Binding/Active

sites in PafA and P00558 were notably higher compared to those of the General sites (Fig. 5a, Fig. 5b, Fig. 5d, and Fig. 5e). At the same time, high weighted sites (These points are weighted at the highest 2%) and binding/active sites in proteins are adjacent in structural space (Fig. 5c) or even have overlapping parts (Fig. 5f). This highlighted DeepEnzyme's ability to accurately identify and prioritize critical active sites and binding sites that significantly impact the normal function of an enzyme. Thus, DeepEnzyme showcased its potential in distinguishing and highlighting the biologically significant regions within enzyme structures, which may contribute to a deeper understanding of enzyme functionality.

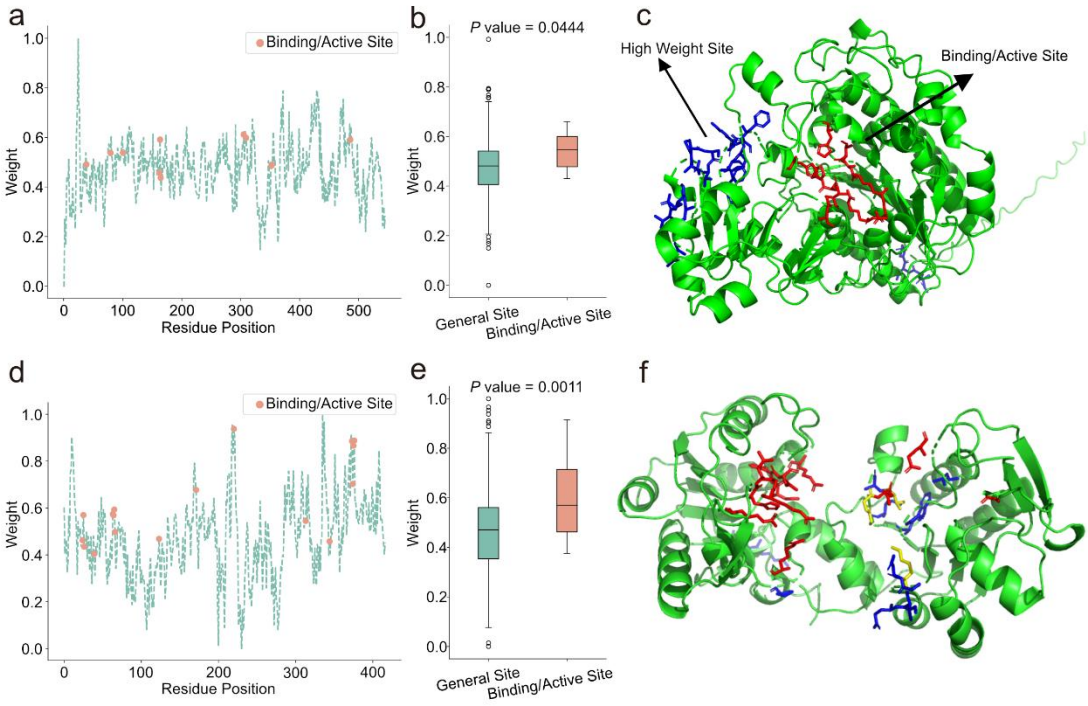


Fig. 5 Provides a graphic representation of the binding/active site and high weight site (These points are weighted at the highest 2%) areas, as well as weight analyses of residue positions in PafA and P00558. (a) The weights of different residue sites in PafA; the red points are Binding/Active sites. (b) A comparison of the weights between the Binding/Active sites and General sites in PafA; the green box line indicates General sites, and the red box line indicates Binding/Active sites. (c) The schematic representation of the Binding/Active sites and High Weight sites regions in PafA, the red region for Binding/Active sites, the blue region for High Weight sites. (d) The weights of different residue sites in P00558; the red points are Binding/Active sites. (e) A comparison of the weights between the Binding/Active sites and General sites in P00558, where the green box line indicates General sites, and the red box line indicates Binding/Active sites. (f) The schematic representation of the Binding/Active sites and High Weight sites regions in P00558, the red region for Binding/Active sites, the blue region for High Weight sites, the yellow region for the overlap between the two.

Application cases: predict the k_{cat} for enzymes from genome-scale metabolic models

Genome-scale metabolic models (GEMs) are invaluable tools for characterizing the cellular metabolism, often encompassing thousands of enzymes that catalyze various reactions. However, determining experimentally k_{cat} for a significant portion of enzymes in GEMs remains a challenging task. In this regard, DeepEnzyme presents itself as a convenient toolbox for predicting enzyme k_{cat} at the genome-scale. To showcase the potential applications of our model, we gathered and processed GEMs from diverse organisms, including *E. coli*, *S. cerevisiae*, and *H. sapiens*[40], before k_{cat} prediction. Leveraging substrate and enzyme information, the genome-scale k_{cat} values could be predicted automatically (Fig. 6a and Fig. 6b). It shows that, at a holistic level, all the k_{cat} values from multiple organisms or a single one exhibited a normal distribution, reflecting the obvious variance in predicted k_{cat} from different organisms.

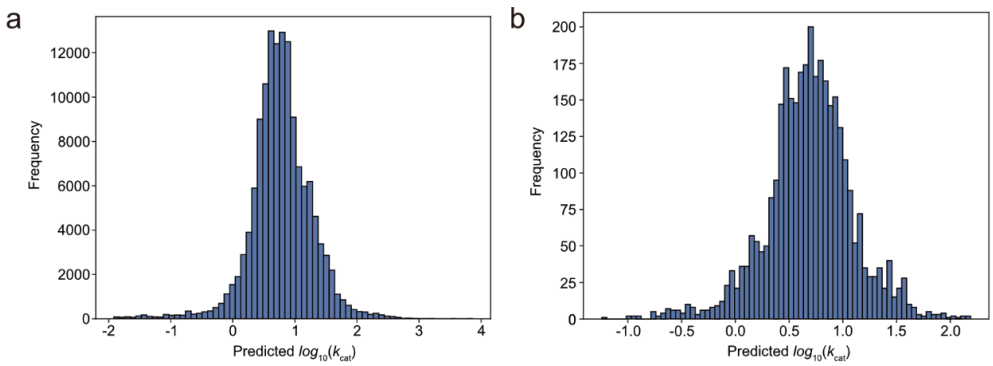


Fig. 6 Predicted k_{cat} values of enzyme-catalyzed reactions in genome-scale metabolic models and 10 model organisms in the EMP pathway. (a) Distribution of k_{cat} values predicted by our model for enzyme-catalyzed reactions in metabolic models. (b) Distribution of k_{cat} values predicted by our model for enzyme-catalyzed reactions (BiGG ID: iAF987).

Discussion

The turnover number is a fundamental parameter for understanding metabolism, growth, and resource allocation[4,5]. However, experimentally determining turnover numbers can be a time-consuming and labor-intensive process. Herein, we present DeepEnzyme, an enhanced model to improve enzyme k_{cat} prediction by utilizing advanced deep learning architectures such as Transformer and GCN. DeepEnzyme skillfully extracts and aggregates essential features from substrate fingerprints, enzyme sequences, and protein 3D structures. With significantly enhanced prediction accuracy, DeepEnzyme surpasses previously reported models in k_{cat} prediction. Specifically, it achieves an impressive R^2 value of 0.60 and a lower RMSE of 0.95 on the test dataset, outperforming both TurNuP (R^2=0.40)[14] and DLKcat (R^2=0.43)[13]. PreKcat[41], which utilizes the latest pre-trained large model for k_{cat} prediction, was released when we prepared this manuscript. However, it is difficult to compare the prediction performances of DeepEnzyme with PreKcat as the source code for training PreKcat was not accessible. According to Yu[41], PreKcat could outperform DeepEnzyme in

prediction accuracy due to its usage of the SMILES Transformer model and the pre-trained protein language model. However, with fewer parameters and smaller model size, DeepEnzyme is more than 2,000 times faster than PreKcat when completing the same prediction task (testing with the PreKcat code example on DELL Precision 5820 Tower).

DeepEnzyme demonstrates robustness in predicting turnover numbers across different levels of sequence similarity. The degree of similarity between the test and the training sequences significantly impacts the predictive performance of the model, and the prediction accuracy is notably improving as sequence similarity increases[13,14]. By comparison, even under circumstances with minimal sequence similarity (0–40%), DeepEnzyme exhibits exceptional prediction accuracy with an R^2 value of 0.54, thus outperforming previous models[13,14]. This improvement can be attributed to the inclusion of structural information about the enzyme. Even if amino acid sequence similarity is low, the structures of the enzymes are likely to exhibit similarity. Through careful preprocessing of the DLKcat dataset, we significantly reduce the tendency of DeepEnzyme to overfit to similar information during training, thereby mitigating systematic errors and enhancing the dataset's suitability for successful model training. The meticulous curation and preprocessing of the dataset contribute to the ability of the model to generalize effectively, ensuring credible predictions across diverse levels of sequence similarity.

Furthermore, DeepEnzyme empowers the evaluation of the influence of saturation mutations on enzyme activities. We demonstrate this capability using two enzymes, PafA[38] and CYP2C9[37], for which abundant experimental data on point mutations are available. The results indicate that our model can, to a considerable extent, reflect how single mutations, as observed in high-throughput experiments, affect enzyme activity. Moreover, DeepEnzyme assigns higher weight scores to active sites and binding sites within a protein structure, as these sites are closely linked to enzyme function and activity[14]. Consequently, the predictions generated by DeepEnzyme offer valuable insights that may guide the rational design of more efficient enzymes for experimental testing and validation.

While DeepEnzyme achieves superior k_{cat} prediction accuracy (R^2 = 0.6) compared to existing models, there is room for enhancement in the following aspects. Firstly, the dataset used in this work encompassed less than 12,000 enzyme-substrate pairs, which is relatively smaller compared to the larger training datasets widely used in state-of-the-art models. Expanding the dataset to encompass a more extensive and diverse range of enzyme-substrate pairs could further enhance the predictive performance of DeepEnzyme. Secondly, the pivotal experimental parameters such as pH and temperature were not taken into account during the model training and prediction, which may contribute to discrepancies between predicted and measured k_{cat} values for certain enzymes. To further enhance the performance of DeepEnzyme, more standardized experimental datasets under diverse environmental conditions should be collected. Additionally, multiple advanced pre-training models can be further merged into the current framework of DeepEnzyme to improve the prediction accuracy. It is

envisioned that, together with other deep learning models, DeepEnzyme will undoubtedly provide new chances to characterize the internal correlations among 1D sequence, 3D structures and enzyme catalytic efficiency, thus accelerating the protein engineering in the coming years.

Methods

Data preprocessing

In the reconstruction of DeepEnzyme, the DLKcat dataset was first downloaded[13], which is openly accessible and contains details about the enzyme and substrate, including the EC number, organism, enzyme sequence, simplified molecular-input line-entry system (SMILES), and k_{cat} value. In order to mitigate the impact of high sequence similarity among enzyme-substrate pairs in the dataset during model training, a preprocessing step was applied to retain only the enzyme-substrate pairs with the highest enzyme sequence length among those with identical substrates and sequence similarity greater than 90%. In order to assess how comparable the enzyme sequence was across the dataset, the MMseqs2[31] approach was employed. Only the enzyme-substrate pairs with the longest enzyme sequences were retained, with the aim of reducing bias and overfitting that may result from highly similar sequences. This preprocessing strategy allows the model to focus on more diverse and representative enzyme-substrate pairs, ultimately enhancing its generalization ability and predictive performance. After data preprocessing, the dataset includes 6,496 unique protein sequences and a total of 11,927 distinct combinations of enzymes and substrates. Additionally, the dataset is randomly divided into groups of training, validation, and testing with ratio at 80%, 10%, and 10% for the development of DeepEnzyme.

Enzyme structure prediction and contact map generation

In order to predict the enzyme structure and gather the structural data of enzymes, ColabFold[22] was used to predict protein structure, which is approximately 40-60 times faster than AlphaFold2. As a whole, the predicted structures have an average predicted local distance difference test (pLDDT) score of 92.67, with over 98% of the enzymes having a pLDDT score higher than 75 (Fig. 1), indicating that the structures used in this work are of high quality. Next, the protein 3D structure was converted into a contact map, which could transform the protein structures into a reasonable format suitable for deep learning algorithms. For creating the contact map, the residues within the protein 3D structures were defined as nodes, while two nodes were connected by a linker if the distance between the respective C_α atoms of two residues was smaller than 10Å. It has been demonstrated that this procedure could successfully capture crucial structural features used in previous deep learning models for protein function prediction[10,42].

Deep learning pipeline for turnover number prediction in DeepEnzyme

To accurately predict enzyme k_{cat}, an end-to-end deep learning model that

combines the Transformer and Graph Convolutional Network (GCN) architectures has been conceived and developed. The Transformer and GCN are particularly designed to extract hidden features from both protein 1D-sequence and 3D structures. Here, the thorough description of the pipeline for training and testing DeepEnzme is given as following.

The protein sequence baseline. In order to extract protein sequence features, the Transformer model, which has excellent performances in natural language processing, is employed in this work. Similar to DLKcat, the protein sequence is split into an overlapping sequence composed by n-grams of amino acids. Before the Transformer, an embedding layer is employed to convert the protein sequence into a word tensor. The hidden vectors of the protein sequence are then determined using word embedding as input, followed by a Transformer encoder with position encodings (Eq.1), 4 multi-head attention layers (Eq.2,3), and a hidden dimension (d_{model}) of 64.

$$PE_{pos,2i} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$
$$PE_{pos,2i+1} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{1}$$

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)W^o$$
$$where \quad head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

Where pos is the position of the amino acid, d_{model} is the dimension of the vector, Q, K, V stand for Query vector, Key vector and Value vector respectively.

The protein and substrate structure baseline. The contact map is used to represent the protein structure, and the substrate graph from the DLKcat database is used to characterize substrate information. GCN (Eq.4) generates substrate vectors and protein structure vectors as its final output via a convolutional neural network.

$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{4}$$

The neural attention mechanism baseline. To obtain the final k_{cat} value output, the neural attention mechanism is incorporated, which enables us to capture the attention weights of various residue sites in the enzyme[43]. This attention mechanism pays more attention into the important fragments of the protein sequence and structure. This attention mechanism (Eq.5-7) works in conjunction with the Transformer and GCN components, which makes it feasible to extract features from the protein sequence and 3D structures.

$$\mathrm{C} = \left\{c_1^{(t)}, c_2^{(t)}, \ldots, c_n^{(t)}\right\} \tag{5}$$

$$h_{substrate} = f(W_{inter} y_{substrate} + b) \tag{6}$$

$$h_i = f(W_{inter} c_i + b) \tag{7}$$

$$a_i = \sigma(h_{substrate}^T h_i) \tag{8}$$

where C is a set of hidden vectors representing the protein sequence, $c_1^{(t)}$ to $c_n^{(t)}$ are sub-hidden vectors for split subsequences within the protein sequence, $y_{substrate}$ is the molecular vector of the substrate, W_{inter} and b are the weight matrix and the bias vector, respectively, used in the neural network, f is a nonlinear activation function, for example, ReLU (Rectified Linear Unit)[44], α_i is the final attention weight value, which is calculated using the attention mechanism, σ is the element-wise sigmoid function, used for specific calculations within the model, and T represents the transpose function.

Evaluation metrics. In this study, three performance metrics are used to evaluate the prediction performances of the model: the coefficient of determination (R^2, Eq.9) the root mean square error (RMSE, Eq.10), and the Pearson correlation coefficient (PCC, Eq.11). These metrics are widely used in regression analysis to examine the correlation and accuracy of predictions against experimental values.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_{ie} - y_{ip})^2}{\sum_{i=1}^{n} (y_{ie} - \bar{y}_{ie})^2} \tag{9}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_{ie} - y_{ip})^2}{n}} \tag{10}$$

$$PCC = \frac{\sum_{i=1}^{n} (y_{ie} - \bar{y}_{ie})(y_{ip} - \bar{y}_{ip})}{\sqrt{\sum_{i=1}^{n} (y_{ie} - \bar{y}_{ie})^2} \sqrt{\sum_{i=1}^{n} (y_{ip} - \bar{y}_{ip})^2}} \tag{11}$$

Where y_{ie} represents the experimental value of k_{cat}, y_{ip} represents a predicted value for k_{cat}, and $\bar{y}$ represents the corresponding averages.

Comparison of different deep learning models in k_{cat} prediction

Utilizing the code and datasets stored in GitHub by Li et al.[13] and Kroll et al.[14], the prediction performance of DLKcat and TurNuP was evaluated respectively. Based on the similarity of the protein sequences, the DLKcat test dataset, the TurNuP test dataset, and our test dataset were separated into three distinct subsets. MMseqs2[31] was utilized to determine the degree of similarity between the protein sequences in the training dataset and the test dataset, FoldSeek[32] and US-align[33] were used to determine the degree of partial protein structural similarity. The R^2 value calculated from test dataset was used to evaluated the prediction accuracy of above different deep learning models.

Evaluative the prediction performance of DeepEnzyme based on large-scale enzyme saturation mutation experiments

Sequence data for CYP2C9 variants and PafA mutations were obtained from the studies of Amorosi, C.J., et al.[37], and Markin, C.J., et al.[38], respectively, and ColabFold was used to predict the 3D structures for all variants of CYP2C9 and PafA in order to generate the corresponding independent datasets containing substrate, protein sequence, and protein structures. The k_{cat} values were predicted directly on the two datasets with DeepEnzyme, and the predictions were compared in groups, which were defined by Amorosi, C.J., et al.[37] or the k_{cat} values in the study by Markin, C.J., et al.[38] .

Interpretation analysis of DeepEnzyme in ranking the key residue sites related to enzyme activities

Wild-type PafA and P00558 were used as input for the k_{cat} prediction by DeepEnzyme, and the data in each column of the protein structural feature matrix extracted by GCN were averaged, while min-max normalization was performed to obtain the weight of each residue site, and the sites were grouped according to the site classification (binding/active sites and general sites) in the UniProt database[8]. Then, the weight score of sites was compared in groups.

Statistical analysis

In all statistical analysis, the two-sided t-test in the Python package SciPy was used[45].

Data availability

The large file used to train the model could be found in: https://figshare.com/authors/Leo_wang/12865928.

Code availability

The codes are accessible at https://GitHub.com/hongzhonglu/DeepEnzyme.

Author contributions

HZL designed the research. TW performed the research. HZL and TW analyzed the data. All authors interpreted the results, discussed, drafted and approved the final manuscript.

Acknowledgements

This work is supported by grant 2022YFA0913000 from the National Key R&D Program of China, Shanghai Pujiang Program, grant 22208211 and 22378263 from the National Natural Science Foundation of China (NSFC).

Conflict of Interests

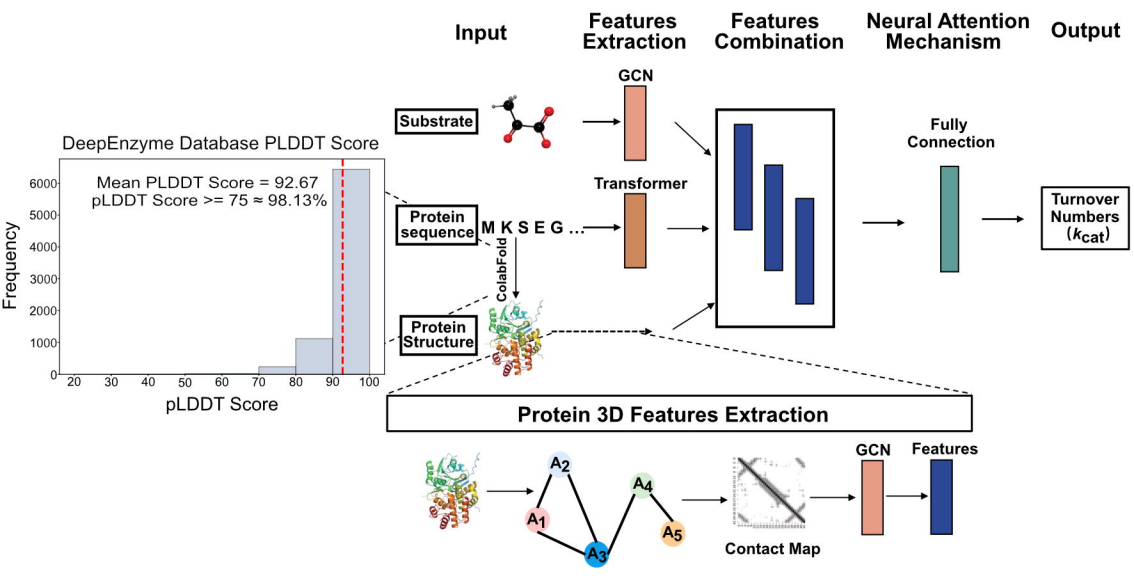
The authors declare that they have no conflicts of interest.

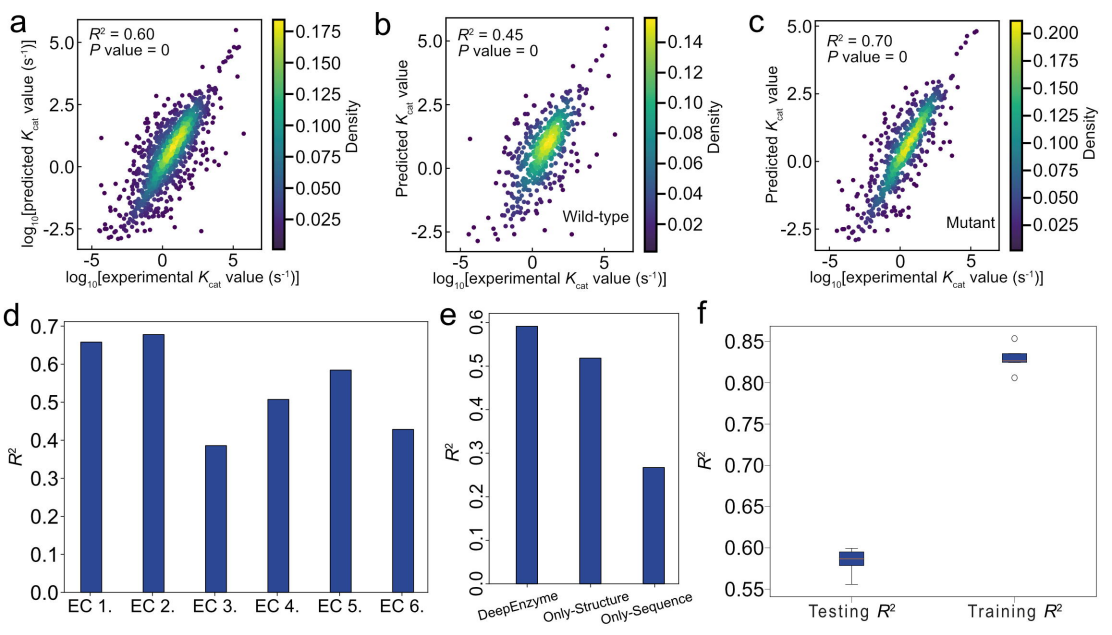
References

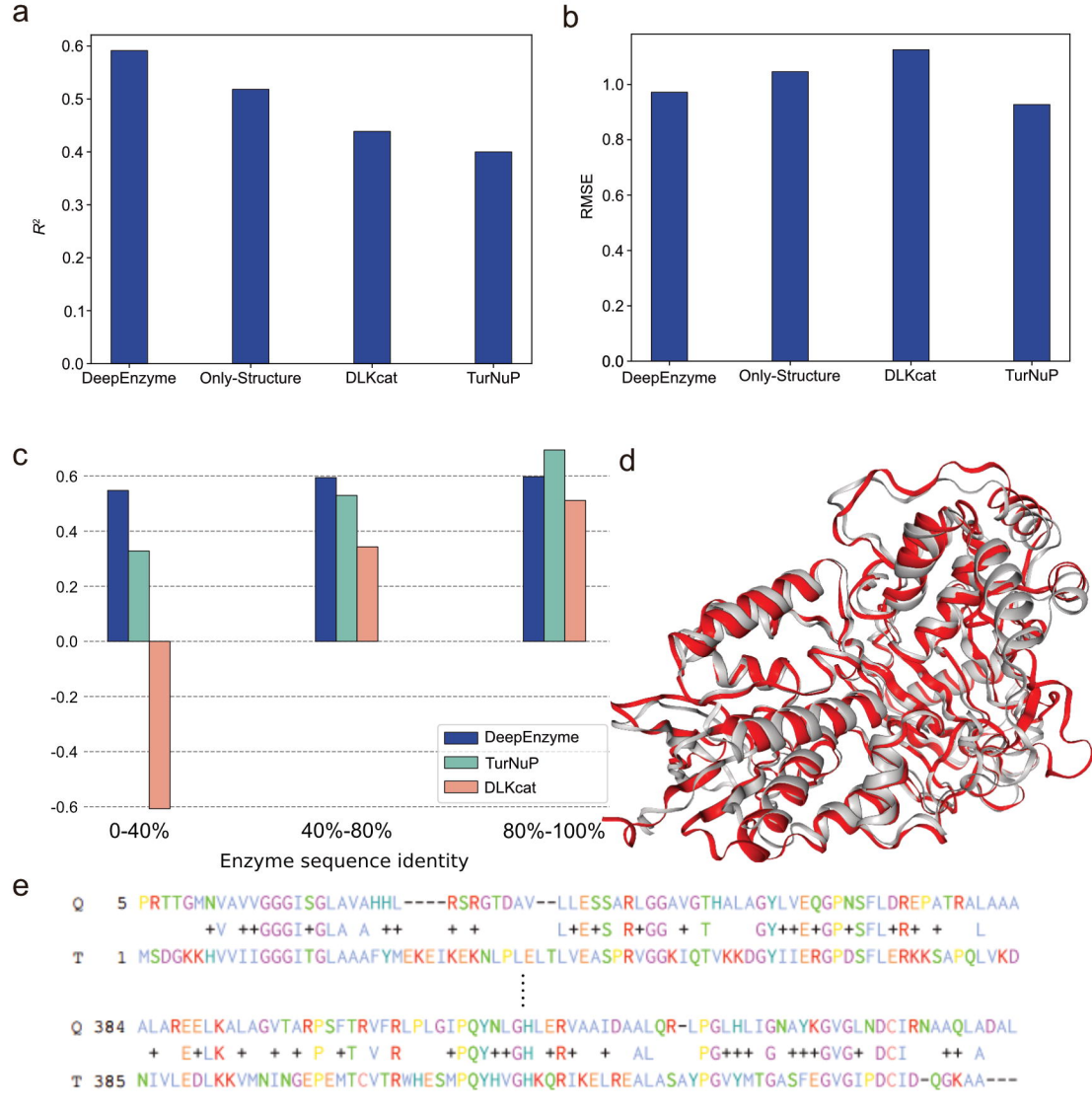
1. Wendering, P., Arend, M., Razaghi-Moghadam, Z. & Nikoloski, Z. Data integration across conditions improves turnover number estimates and metabolic predictions. *Nature Communications* **14**, 1485 (2023).
2. Davidi, D., *et al.* Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro k cat measurements. *Proceedings of the National Academy of Sciences* **113**, 3401-3406 (2016).
3. Nilsson, A., Nielsen, J. & Palsson, B.O. Metabolic models of protein allocation call for the kinetome. *Cell Systems* **5**, 538-541 (2017).
4. Sánchez, B.J., *et al.* Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular systems biology* **13**, 935 (2017).
5. Yang, L., Yurkovich, J.T., King, Z.A. & Palsson, B.O. Modeling the multi-scale mechanisms of macromolecular resource allocation. *Current opinion in microbiology* **45**, 8-15 (2018).
6. Schomburg, I., *et al.* The BRENDA enzyme information system–From a database to an expert system. *Journal of biotechnology* **261**, 194-206 (2017).
7. 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**, D656-D660 (2018).
8. UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Research* **51**, D523-D531 (2023).
9. Jumper, J., *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
10. Gligorijević, V., *et al.* Structure-based protein function prediction using graph

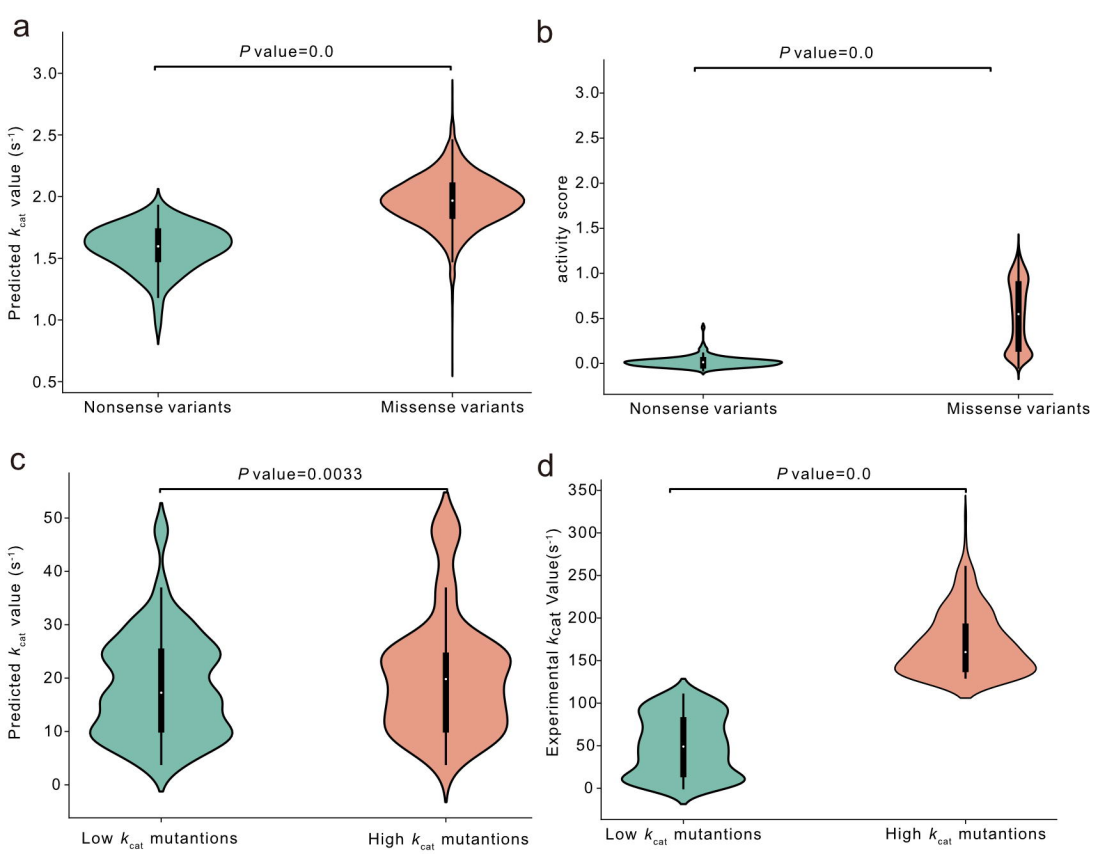
convolutional networks. *Nature communications* **12**, 3168 (2021).
11. Zrimec, J.*, et al.* Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nature Communications* **11**, 6141 (2020).
12. Heckmann, D.*, et al.* Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nature communications* **9**, 5252 (2018).
13. Li, F.*, et al.* Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis* **5**, 662-672 (2022).
14. Kroll, A., Hu, X.-P., Liebrand, N.A. & Lercher, M.J. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *bioRxiv*, 2022.2011.2010.516024 (2022).
15. Goodsell, D.S. The machinery of life. (2009).
16. Leveson-Gower, R.B., Mayer, C. & Roelfes, G. The importance of catalytic promiscuity for enzyme design and evolution. *Nature Reviews Chemistry* **3**, 687-705 (2019).
17. Smith, M.R.*, et al.* Elucidating structure–performance relationships in whole-cell cooperative enzyme catalysis. *Nature catalysis* **2**, 809-819 (2019).
18. Volkenshtein, M.V., R.R. Dogonadze, A.K. Madumarov, Z.D. Urushadze, & Yu.I. Kharkats. Theory of Enzyme Catalysis. *Molekuliarnaya Biologia* **431-439**(1972).
19. Li, B., Yang, Y.T., Capra, J.A. & Gerstein, M.B. Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. *PLOS Computational Biology* **16**, e1008291 (2020).
20. Torng, W. & Altman, R.B. High precision protein functional site detection using 3D convolutional neural networks. *Bioinformatics* **35**, 1503-1512 (2018).
21. Baek, M.*, et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021).
22. Mirdita, M.*, et al.* ColabFold: making protein folding accessible to all. *Nature methods* **19**, 679-682 (2022).
23. Kipf, T.N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
24. Duvenaud, D.K.*, et al.* Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems* **28**(2015).
25. Coley, C.W., Barzilay, R., Green, W.H., Jaakkola, T.S. & Jensen, K.F. Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of chemical information and modeling* **57**, 1757-1772 (2017).
26. Vaswani, A.*, et al.* Attention is all you need. *Advances in neural information processing systems* **30**(2017).
27. Lin, Z.*, et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
28. Madani, A.*, et al.* Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 1-8 (2023).
29. Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **8**, 31 (2013).
30. Kroll, A., Ranjan, S., Engqvist, M.K.M. & Lercher, M.J. A general model to predict small

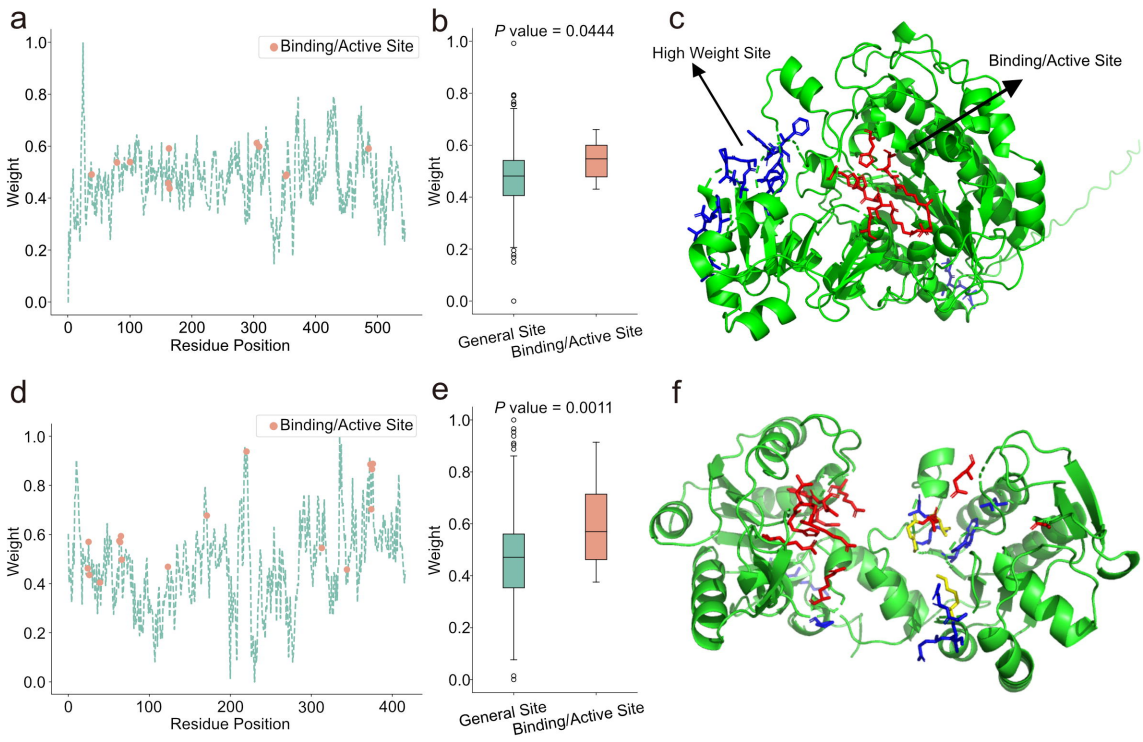
molecule substrates of enzymes based on machine and deep learning. *Nature Communications* **14**, 2787 (2023).
31. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**, 1026-1028 (2017).
32. van Kempen, M.*, et al.* Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* (2023).
33. Zhang, C., Shine, M., Pyle, A.M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature Methods* **19**, 1109-1115 (2022).
34. Rettie, A.E. & Jones, J.P. Clinical and toxicological relevance of CYP2C9: drug-drug interactions and pharmacogenetics. *Annu. Rev. Pharmacol. Toxicol.* **45**, 477-494 (2005).
35. Zanger, U.M. & Schwab, M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & therapeutics* **138**, 103-141 (2013).
36. Steward, D.J.*, et al.* Genetic association between sensitivity to warfarin and expression of CYP2C9* 3. *Pharmacogenetics* **7**, 361-367 (1997).
37. Amorosi, C.J.*, et al.* Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. *The American Journal of Human Genetics* **108**, 1735-1751 (2021).
38. Markin, C.J.*, et al.* Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics.
39. Jencks, W.P. *Catalysis in chemistry and enzymology*, (Catalysis in chemistry and enzymology /, 1987).
40. Kroll, A., Engqvist, M.K., Heckmann, D. & Lercher, M.J. Deep learning allows genome-scale prediction of Michaelis constants from structural features. *PLoS biology* **19**, e3001402 (2021).
41. Yu, H., Deng, H., He, J., Keasling, J. & Luo, X. Highly accurate enzyme turnover number prediction and enzyme engineering with PreKcat. (2023).
42. Wang, Z.*, et al.* Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction. *Scientific reports* **12**, 6832 (2022).
43. Bahdanau, D., Cho, K. & Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *Computer Science* (2014).
44. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning.
45. Virtanen, P.*, et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* **17**, 261-272 (2020).

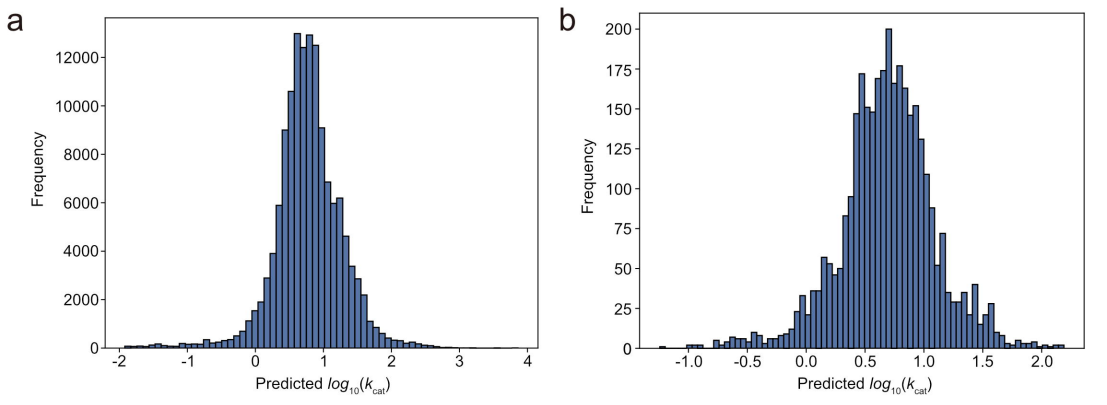

Input
Features Extraction
Features Combination
Neural Attention Mechanism
Output
DeepEnzyme Database PLDDT Score
Mean PLDDT Score = 92.67
pLDDT Score >= 75 ≈ 98.13%
Frequency
pLDDT Score
Substrate
Protein sequence
M K S E G ...
ColabFold
Protein Structure
GCN
Transformer
Fully Connection
Turnover Numbers (k_{cat})
Protein 3D Features Extraction
A1
A2
A3
A4
A5
Contact Map
GCN
Features


a
R^2 = 0.60
P value = 0
$\log_{10}$[predicted K_{cat} value (s^{-1})]
$\log_{10}$[experimental K_{cat} value (s^{-1})]
Density
b
R^2 = 0.45
P value = 0
Predicted K_{cat} value
Wild-type
$\log_{10}$[experimental K_{cat} value (s^{-1})]
Density
c
R^2 = 0.70
P value = 0
Predicted K_{cat} value
Mutant
$\log_{10}$[experimental K_{cat} value (s^{-1})]
Density
d
R^2
EC 1.
EC 2.
EC 3.
EC 4.
EC 5.
EC 6.
e
R^2
DeepEnzyme
Only-Structure
Only-Sequence
f
R^2
Testing R^2
Training R^2


a
R^2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
DeepEnzyme
Only-Structure
DLKcat
TurNuP
b
RMSE
0.0
0.2
0.4
0.6
0.8
1.0
DeepEnzyme
Only-Structure
DLKcat
TurNuP
c
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
DeepEnzyme
TurNuP
DLKcat
0-40%
40%-80%
80%-100%
Enzyme sequence identity
d
e
Q 5 PRTTGMNVAVVGGGISGLAVAHHL----RSRGTDAV--LLESSARLGGAVGTHALAGYLVEQGPNSFLDREPATRALAAA
+V ++GGGI+GLA A ++ + + L+E+S R+GG + T GY++E+GP+SFL+R+ + L
T 1 MSDGKKHVVIIGGGITGLAAAFYMEKEIKEKNLPLELTLVEASPRVGGKIQTVKKDGYIIERGPDSFLERKKSAPQLVKD
⋮
Q 384 ALAREELKALAGVTARPSFTRVFRLPLGIPQYNLGHLERVAAIDAALQR-LPGLHLIGNAYKGVGLNDCIRNAAQLADAL
+ E+LK + + + P +T V R +PQY++GH +R+ + AL PG+++ G +++GVG+ DCI ++ A
T 385 NIVLEDLKKVMNINGEPEMTCVTRWHESMPQYHVGHKQRIKELREALASAYPGVYMTGASFEGVGIPDCID-QGKAA---


a
P value=0.0
Predicted k_{cat} value (s^{-1})
3.0
2.5
2.0
1.5
1.0
0.5
Nonsense variants
Missense variants
b
P value=0.0
activity score
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Nonsense variants
Missense variants
c
P value=0.0033
Predicted k_{cat} value (s^{-1})
50
40
30
20
10
0
Low k_{cat} mutantions
High k_{cat} mutantions
d
P value=0.0
Experimental k_{cat} Value(s^{-1})
350
300
250
200
150
100
50
0
Low k_{cat} mutantions
High k_{cat} mutantions


a
Binding/Active Site
Weight
Residue Position
b
P value = 0.0444
Weight
General Site
Binding/Active Site
c
High Weight Site
Binding/Active Site
d
Binding/Active Site
Weight
Residue Position
e
P value = 0.0011
Weight
General Site
Binding/Active Site
f


a
Frequency
0
2000
4000
6000
8000
10000
12000
-2
-1
0
1
2
3
4
Predicted $log_{10}(k_{cat})$
b
Frequency
0
25
50
75
100
125
150
175
200
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
Predicted $log_{10}(k_{cat})$