Robust enzyme kinetics prediction through pairwise relative learning

Xiongwen Li[1, 2], Zhengkai Li[3], Jiawei Zou[1, 2], Wenjie Chen[1, 2], Shujia Liu[1, 2], Ke Wu[1, 2], Jiahao Luo[1, 2], Yu Chen[4], Feiran Li[1, 2, *]

1 Institute of Biopharmaceutical and Health Engineering, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
2 Key Laboratory for Industrial Biocatalysis, Ministry of Education, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China
3 Department of Computer Science, Boston University, Boston, MA 02215, USA
4 State Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

* Corresponding author:
Feiran Li, Email: feiranli@sz.tsinghua.edu.cn

Abstract

Enzyme kinetics prediction is a fundamental challenge, and numerous computational methods have been developed. However, most methods rely on absolute regression across heterogeneous datasets, which are noisy and prone to regression-to-the-mean bias, limiting accurate identification of highly active enzymes. Here we present DeltaKcat, a siamese neural network framework that reformulates kinetics prediction as a pairwise relative learning problem. By learning pairwise differences within consistent experimental contexts, DeltaKcat improves accuracy, robustness, and generalization over state-of-the-art methods. This framework enables DeltaKcat to (i) mitigate inter-study variability, (ii) generalize to unseen enzymes and substrates, and (iii) prioritize highly active enzymes. We demonstrate these capabilities through systematic benchmarking and external adenylate kinase dataset validation. We anticipate that this approach will be broadly useful for enzyme discovery and engineering, and more generally for learning from noisy biological data.

Introduction

Enzymes are the fundamental functional units of life, governing the rates and efficiencies of biochemical reactions[1-5]. Their activity is quantitatively described by enzyme kinetics, which determines catalytic efficiency[6], including the turnover number (k_{cat}), Michaelis constant (K_m), and catalytic efficiency (k_{cat}/K_m). However, despite the more than 230 million protein sequences in UniProt[7], fewer than 0.1% are linked to experimentally measured kinetic parameters in databases such as BRENDA[8] and SABIO-RK[9], highlighting a substantial gap between enzyme sequence availability and quantitative characterization of enzyme function[10]. Consequently, enzyme kinetics characterization increasingly relies on computational methods[11-14]. Recent machine learning approaches predict enzyme kinetic parameters by jointly modeling enzyme sequences and substrates to predict the kinetic values[15-19]. These methods have evolved from traditional machine learning to deep learning frameworks[15], and more recently to large pretrained protein and molecular models that further improve predictive performance[18, 20-23].

However, two fundamental challenges critically limit current approaches and must be explicitly addressed. First, the enzyme-substrate space is extremely sparse and unevenly sampled, with many enzyme classes represented by only a small number of experimentally characterized instances, severely limiting generalization to unseen sequences and substrates[20, 24-27]. Second, kinetic measurements exhibit substantial cross-study variability, where reported values for the same enzyme-substrate system can differ by orders of magnitude due to differences in experimental conditions and protocols, introducing strong systematic noise into training data[28, 29]. Together, these limitations make direct prediction of absolute kinetic parameters challenging for real-world applications such as enzyme discovery and engineering[30]. In particular, models trained on heterogeneous datasets tend to capture dataset-specific biases rather than true biophysical relationships, limiting their ability to extrapolate beyond curated benchmarks[31]. Therefore, overcoming these limitations requires a fundamental reformulation of enzyme kinetics prediction beyond absolute value estimation toward learning invariant functional relationships.

To address these challenges, we propose DeltaKcat, which reformulates enzyme kinetics prediction as a pairwise relative learning problem, where the model learns kinetic differences between paired enzyme-substrate systems rather than predicting absolute values independently. By comparing enzyme pairs within a consistent experimental context (e.g., within the same study), this formulation effectively reduces cross-study systematic bias and the impact of measurement inconsistency. Importantly, it shifts the learning objective from dataset-dependent values to functional relationships between sequence and substrate changes. This pairwise formulation also increases the effective size of training data, as each study yields multiple valid comparisons, substantially increasing usable data without additional experiments. DeltaKcat is trained on a high-quality, large-scale dataset integrated from BRENDA, SABIO-RK, and literature-mined resources. We further introduce a reference-based reconstruction strategy to recover absolute kinetic parameters from predicted relative differences, enabling benchmarking and evaluation across multiple challenging settings, including low-sequence similarity, cold-start scenarios, and cross-family generalization. Overall, this work establishes a general framework for enzyme kinetics prediction based on pairwise relative functional learning and provides a principled approach for robust prediction of noisy dataset.

Results

Overview of DeltaKcat

We first constructed a comprehensive dataset of enzyme kinetic parameters by integrating k_{cat}, K_m, and k_{cat}/K_m measurements from the BRENDA[8] and SABIO-RK[9] databases, and further expanding coverage using large language model-based literature mined resources, including EnzyExtraDB[32] and Enzyme Co-Scientist[33]. After removing entries lacking substrate simplified molecular input line entry system (SMILES) and duplicate records, the curated dataset contains 135,072 k_{cat}, 208,176 K_m, and 110,624 k_{cat}/K_m measurements, spanning 35,076 unique protein sequences and 13,705 distinct substrates (Supplementary Fig. 1). To reduce cross-study variability, we reorganized the data into paired representations of enzyme-substrate complexes within the same study. By fixing either the enzyme or the substrate within each pair, this design enables the model to disentangle sequence-driven and substrate-driven effects under a consistent experimental context. By separating these two pair types, we could explicitly distinguish kinetic changes caused by enzyme sequence variation from those caused by substrate variation. These capture (i) variation across enzyme sequences while shared substrates, and (ii) variation across substrates with shared enzymes, yielding 117,873/92,187/51,748 and 142,909/160,331/98,669 paired samples for k_{cat}, K_m, and k_{cat}/K_m, respectively (Supplementary Fig. 2).

We developed DeltaKcat, a siamese neural network model that predicts relative enzyme kinetic differences between enzyme-substrate complexes (Fig. 1). The model takes as input a pair of enzyme-substrate complexes and learns the difference in their kinetic properties. For each complex, the protein sequence and substrate molecule are encoded using pretrained models (ProtT5[34] for proteins, and MolT5[35] for substrates), complemented by molecular access system (MACCS) fingerprints, and integrated via a gated cross-attention module to capture enzyme-substrate interactions. The resulting representations from the two complexes are then combined through explicit differencing, together with their individual features, and fed into a multilayer perceptron (MLP) to predict relative k_{cat}, K_m, or k_{cat}/K_m. This formulation shifts the learning objective from absolute values to invariant functional relationships, improving robustness to data heterogeneity.

Prediction of relative enzyme kinetics

DeltaKcat shows strong predictive performance across relative kinetics prediction on both sequence-diversity and substrate-specificity tasks, achieving high correlations across all three relative kinetic prediction targets (Δk_{cat}, ΔK_m, and $\Delta k_{cat}/K_m$), with Pearson correlation coefficients (PCC) reaching up to 0.80 (Fig. 2a). Considering the inherent noise and cross-study variability in enzyme kinetic measurements, these correlation levels indicate that the model is able to recover substantial functional signal from highly heterogeneous data. Notably, we observe a clear task-dependent pattern in predictive performance. Δk_{cat} is most accurately predicted in the sequence-diversity task, whereas ΔK_m is best predicted in the substrate-specificity task (Fig. 2a), suggesting that the model is more sensitive to sequence-driven variations in catalytic turnover and substrate-driven variations in binding affinity. To further assess generalization beyond random-split settings, we further evaluated DeltaKcat under a more challenging unseen-sequence or unseen-substrate scenario (cold-start). While

performance is lower under this setting, ΔK_m remains strongly predictable in the unseen-substrate setting, achieving a PCC of 0.74, highlighting robust generalization to novel substrates (Fig. 2b), indicating that the model retains robust predictive capability when extrapolating to previously unobserved enzyme-substrate combinations.

We next examined performance across different sequence contexts. DeltaKcat achieves higher accuracy when comparing mutants than when comparing distinct wild-type sequences (Supplementary Fig. 3), suggesting that kinetic changes induced by local sequence perturbations are more predictable than those arising from larger evolutionary divergence. The model showed consistently similar performance across all enzyme commission (EC) classes, indicating robust generalization across diverse functional categories (Supplementary Fig. 4). Beyond regression accuracy, DeltaKcat also reliably captures the direction of kinetic changes with an accuracy of 0.71, indicating that the model learns meaningful relative sequence-function relationships (Fig. 2c for k_{cat}, Supplementary Fig. 5 for K_m and k_{cat}/K_m). All results are consistent across three independent runs (Fig. 2a-c), demonstrating the stability of the model across different data splits. Ablation studies showed that each component of DeltaKcat contributes to performance, with module-specific effects observed across sequence-diversity and substrate-specificity tasks (Supplementary Fig. 6).

Absolute kinetic parameter prediction on the sequence-diversity task

While relative kinetic parameters are central to enzyme engineering, absolute kinetic parameters are also essential for systems-level applications. We therefore extended DeltaKcat to absolute value prediction using a reference-based inference approach, in which DeltaKcat's relative outputs are anchored with experimentally characterized values to obtain k_{cat}, K_m, and k_{cat}/K_m values (Methods). Under the random split, DeltaKcat achieves R^2 values of 0.79 for k_{cat}, 0.80 for K_m, and 0.84 for k_{cat}/K_m (Supplementary Fig. 7). Under the cold-start split, DeltaKcat also achieves strong performance across all three tasks. For k_{cat}, it reaches an R^2 of 0.71 and a PCC of 0.85, outperforming all baseline methods, including DLKcat, CataPro, and UniKP (R^2 = 0.14 to 0.48) (Fig. 3a, Supplementary Fig. 8). Comparable gains are observed for K_m and k_{cat}/K_m, with consistent improvements in both R^2 and PCC (Fig. 3a for R^2, Supplementary Fig. 8 for PCC). As cold-start evaluation better reflects realistic application scenarios, we therefore focus on the cold-start results in the following analyses.

To assess performance under realistic experimental conditions, we evaluated DeltaKcat within individual publications. In this publication-level setting, kinetic parameters are measured in the same laboratory under comparable experimental protocols. Under this setting, DeltaKcat achieves a median PCC of 0.78 for k_{cat} prediction, outperforming baseline models (PCCs of 0.11-0.36) (Fig. 3b). Across EC classes, the largest gains are observed in EC 1, EC 2, and EC 6, while consistently outperforming baseline models across the remaining classes (Fig. 3c). Consistent trends are observed for K_m and k_{cat}/K_m prediction (Supplementary Fig. 9-10).

Importantly, DeltaKcat shows clear advantages in identifying highly active enzymes, where existing methods are often biased toward the mean. It better captures extreme kinetic regimes and consistently outperforms all baselines across top-ranked enzyme subsets (top 10-25%),

with the largest gains observed for the top 10% most active enzymes (Fig. 3d). Consistent improvements are observed for K_m and k_{cat}/K_m (Supplementary Fig. 11). This improvement arises because models trained via absolute regression are biased toward the central tendency of the training distribution, leading to systematic underestimation of extreme values. In contrast, modeling relative differences removes this dependency on absolute value distributions and better preserves extreme kinetic signals.

We further evaluated generalization across sequence similarity levels. DeltaKcat maintains strong performance even in low-similarity regimes, achieving PCC values of 0.76 and 0.78 in the 0-40% and 40-60% identity groups for k_{cat} (Fig. 3e). Consistent performance is also observed for K_m and k_{cat}/K_m, where DeltaKcat achieves PCC values above 0.85 and 0.73 in the 0-40% identity group (Supplementary Fig. 12). These results demonstrate robust extrapolation capability across diverse sequence space of DeltaKcat.

Finally, we assessed the consistency of the relative-to-absolute conversion procedure. Predicted absolute values show low variance across different anchor selections, with standard deviations centered around 0.25 (Fig. 3f for k_{cat}, Supplementary Fig. 13 for K_m and k_{cat}/K_m). This indicates that the inference procedure is stable and insensitive to the choice of reference enzyme-substrate complex, supporting the reliability of the proposed conversion framework.

Absolute value prediction on the substrate-specificity task

We next evaluated DeltaKcat in the substrate-specificity setting by converting relative kinetic predictions into absolute k_{cat}, K_m, and k_{cat}/K_m values. Despite substantially increased chemical diversity, DeltaKcat consistently outperforms all baseline models. For k_{cat}, DeltaKcat achieves an R^2 of 0.66 and a PCC of 0.83, exceeding CataPro ($R^2 = 0.42$) and DLKcat ($R^2 = 0.17$) (Fig. 4a). Similar improvements are observed for K_m ($R^2 = 0.47$, Fig. 4b) and k_{cat}/K_m ($R^2 = 0.53$ vs. 0.18 for the best baseline; Fig. 4c). Moreover, DeltaKcat also consistently outperforms baseline models across publication-level analysis (Supplementary Fig. 14). To assess the generalization of DeltaKcat to dissimilar substrates, we stratified test samples based on substrate similarity to the training set. DeltaKcat maintains stable performance across all similarity ranges, including the most challenging 0-80% interval, where it achieves PCCs of 0.80 (k_{cat}), 0.67 (K_m), and 0.77 (k_{cat}/K_m) (Fig. 4d-f). Overall, DeltaKcat generalizes well to substrate-driven kinetic variation and uncharacterized substrates.

Performance of DeltaKcat on an external adenylate kinase dataset

To validate the real-world applicability of DeltaKcat, we tested it on an external adenylate kinase (ADK) dataset that was completely excluded from training[36]. ADKs are key enzymes in energy metabolism that maintain intracellular adenylate homeostasis by catalyzing the interconversion of ATP, ADP, and AMP (Fig. 5a)[37]. This dataset contains 175 kinetic parameter records derived from diverse ADK variants.

Under a zero-shot setting, we evaluated the model's utility for enzyme prioritization. Using the top 5% of true k_{cat} values as a proxy for highly active enzymes, DeltaKcat successfully identified one highly active enzyme within its top 5 candidates, whereas all baseline models

failed to retrieve any highly active enzyme within their top 10 candidates (Fig. 5 b, c). Across the top 30 candidates, DeltaKcat recovers 56% of the true highly active enzymes, compared with 22% for the best-performing baseline UniKP (Fig. 5b, c), highlighting its effectiveness for practical enzyme discovery.

Under the zero-shot setting, DeltaKcat achieves a PCC of 0.40 and a Spearman correlation coefficient (SCC) of 0.33 on this external ADK test set, representing the best performance among all models evaluated, including UniKP (PCC = 0.22, SCC = 0.21) (Fig. 5d). Given the complete absence of the ADK enzyme family from the training data, these findings indicate that the paired comparative learning framework of DeltaKcat provides meaningful generalization capabilities across unseen enzyme classes. Finally, we assessed whether limited family-specific data can further improve performance. Using 20-140 training samples for fine-tuning with a fixed held-out test set, performance increases steadily with data size, with a substantial improvement beyond 60 samples. The fine-tuned model reached a PCC of 0.69, corresponding to a 72% gain over the zero-shot setting (Fig. 5d, Supplementary Fig. 15), and outperformed the previously reported ADK-specific model (ADK evolutionary). These results demonstrate that DeltaKcat not only generalizes to completely unseen enzyme families, but can also be rapidly adapted using small amounts of experimental data for improved predictive accuracy and enzyme prioritization.

Discussion

Despite the rapid expansion of enzyme databases, the limited availability of high-quality kinetic measurements remains a fundamental bottleneck for quantitative modeling and biocatalyst design[24]. Current approaches typically regress absolute k_{cat}, K_m and k_{cat}/K_m values from heterogeneous experimental datasets[20, 21], which introduces systematic variability and noise and leads models to capture dataset-specific biases rather than generalizable enzyme-substrate quantitative function relationships. In addition, standard random splits may overestimate performance due to residual sequence or substrate similarity between training and test data.

To address these limitations, DeltaKcat reformulates enzyme kinetics prediction as a relational learning problem. Instead of directly predicting absolute values, the model learns relative kinetic differences between paired enzyme-substrate complexes under sequence or substrate perturbations. This shifts learning from absolute scale fitting to invariant functional comparison, improving robustness to experimental heterogeneity while focusing on perturbation-driven signals.

Absolute kinetic parameters can then be reconstructed from relative predictions using a reference-based inference strategy. Across diverse settings, including low sequence similarity, low substrate similarity, and zero-shot validation on external ADK dataset, DeltaKcat consistently outperforms state-of-the-art methods and maintains strong performance even on unseen enzyme families, supporting its utility for enzyme discovery and functional annotation.

Despite these advantages, several limitations remain. First, although the dataset used in this study is among the largest curated kinetic resources to date, the overall availability of measurements remains limited, particularly for rare reaction types and underrepresented enzyme classes. Second, key environmental factors such as pH, temperature, and assay conditions, which are known to significantly influence catalytic activity, are not explicitly

incorporated into the current framework[38]. Third, structural information of proteins and enzyme-substrate complexes is not utilized, which may further improve model performance[39, 40].

More broadly, relational learning provides a general framework for biomolecular property prediction under sparse, noisy, and heterogeneous data. This approach may facilitate enzyme engineering, metabolic design, and functional annotation by enabling more reliable inference in uncharacterized sequence and substrate spaces.

Method

Dataset preparation

Experimentally measured k_{cat}, K_m, and k_{cat}/K_m values were collected from the BRENDA and SABIO-RK[9] open-source databases in October 2024. Each data sample included the EC number, protein sequence, organism information, substrate name, substrate SMILES, PubMed ID, UniProt ID, temperature, pH, modification information, and enzyme kinetic parameter values. The SMILES string of the substrate was queried from the PubChem Compound Database[41] using the substrate name or international chemical identifiers (InChI). Enzyme sequences are primarily retrieved from the UniProt database6 using UniProt ID. For entries lacking UniProt IDs, sequences are obtained from the UniProt database based on the corresponding EC numbers and species information. Mutant enzyme sequences were generated by applying the amino acid substitutions recorded in the dataset to the corresponding wild-type sequences. We further incorporated data from EnzyExtraDB and Enzyme Co-Scientist databases, two databases constructed using large language models to extract kinetic information from the literature. The same set of data fields was retained across all sources to ensure consistency. The final dataset contains 135072 unique k_{cat} entries, 208176 K_m entries, and 110624 k_{cat}/K_m entries.

Relative enzyme kinetic parameter datasets preparation

Two types of relative enzyme kinetic parameter datasets were constructed based on specific prediction tasks, the sequence-diversity task and the substrate-specificity task. In both datasets, pairs of enzyme-substrate complexes were constructed under the constraint that samples originated from the same publication, ensuring consistent experimental conditions within each pair. For each pair, the relative kinetic parameter was calculated as the $\log_2$-transformed ratio between the kinetic value of the first complex and that of the second complex. For the sequence-diversity dataset, we retained only records in which at least two distinct protein sequences were associated with the same substrate within the same study. A cold-start split was then applied: a subset of protein sequences was held out as test sequences. Training pairs were formed exclusively from training sequences sharing the same substrate and study, while test pairs were formed by pairing complexes involving test sequences with any other complex under the same constraints. This design ensures that no test sequence appears during training. For the substrate-specificity task, we retained only records in which at least two distinct substrates were associated with the same protein sequence within the same study. A cold-start split was applied by holding out a subset of substrates as test substrates. Test pairs were formed by pairing complexes containing test substrates with other complexes from the same study, while training pairs were formed solely from complexes involving training substrates. This prevents any test substrate from appearing in the training set.

296

297 **Evaluation metrics**

298 In this work, the coefficient of determination (R^2), Pearson's correlation coefficient (PCC),
299 Spearman correlation coefficient (SCC), and root mean squared error (RMSE) was used to
300 assess prediction accuracy. Recall@K was used to assess top-K retrieval performance.

301 $$R^2 = 1 - \frac{\sum_{i=1}^{n}(m_i - z_i)^2}{\sum_{i=1}^{n}(m_i - \overline{m}_i)^2}$$

302

303 $$PCC = 1 - \frac{\sum_{i=1}^{n}(z_i - \overline{z}_i)(m_i - \overline{m}_i)}{\sqrt{\sum_{i=1}^{n}(z_i - \overline{z}_i)^2} \cdot \sqrt{\sum_{i=1}^{n}(m_i - \overline{m}_i)^2}}$$

304 $$SCC = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

305 $$RMSE = 1 - \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - z_i)^2}$$

306 $$Recall@K = \frac{\sum_{i=1}^{K} r_i}{R}$$

307 where m_i is the value of the experimentally measured enzyme kinetic parameter, z_i denotes the
308 value of the model-predicted enzyme kinetic parameter. d_i denotes the rank difference between
309 the predicted and experimental values for a given sample. $\overline{m}_i$ denotes the mean value of the
310 experimentally measured enzyme kinetic parameter, $\overline{z}_i$ denotes the mean value of the model-
311 predicted enzyme kinetic parameter, and n denotes the number of samples. K represents the
312 predetermined number of top ranked candidates being evaluated. r_i is an indicator variable that
313 equals 1 if the candidate at rank i is a true positive and 0 if it is a false positive. R represents the
314 total number of actual positive samples present in the entire dataset.

315

316 **Absolute kinetic parameter inference from relative predictions**

317 To extend DeltaKcat from relative prediction to absolute kinetic parameter estimation, we
318 applied a reference-based inference approach that converts predicted relative ratios into
319 absolute values. For each test sample, we first identified whether a valid reference complex was
320 available in the training set. Specifically, we retained only those comparative pairs in which
321 one enzyme-substrate complex was present in the training set and had an experimentally
322 measured kinetic parameter. This known complex was treated as the reference anchor, whereas
323 the other complex was regarded as the target whose absolute parameter was to be inferred.
324 DeltaKcat predicts the relative kinetic relationship between the target complex and the
325 reference complex. We then combined the predicted relative ratio with the experimentally
326 determined kinetic value of the reference anchor to infer the absolute kinetic parameter of the
327 target complex. When multiple valid training-set anchors were available for the same target
328 complex, we performed inference independently for each anchor and then averaged the
329 resulting absolute values to obtain the final prediction. Using this approach, DeltaKcat projects
330 comparative predictions into the absolute parameter space through experimentally reference
331 points.

332

External ADK dataset and evaluation

The external ADK dataset contains 175 kinetic parameter records collected from diverse ADK variants. For zero-shot evaluation, we applied the relative prediction framework without using any ADK data for training. Five ADK enzymes were randomly selected as reference enzymes with known experimental kinetic parameters, and the remaining enzymes were treated as target enzymes. Each target enzyme was paired with all five references to obtain relative predictions. These predictions were then converted to absolute kinetic parameters using the corresponding reference values, and the five inferred values were averaged to obtain the final prediction for each target enzyme. For fine-tuning experiments, 30 ADK samples were randomly selected as a fixed test set. From the remaining samples, training subsets of different sizes were constructed, ranging from 20 to 140 samples. DeltaKcat was fine-tuned on each training subset and evaluated on the fixed test set.

Encoding for enzyme sequence, structure and substrates

We used pre-trained language models to characterize proteins and substrates, aiming to obtain more expressive and transferable encodings. Enzyme sequences were encoded as residue-level embeddings from ProtT5-XL-UniRef, yielding an L×1024 tensor (L is the sequence length). To control computation and memory, sequences longer than 1024 residues were truncated at residue 1024. Substrates were represented with two features: a 768-dimensional continuous vector from MolT5 applied to normalized SMILES, and a 167-bit MACCS key fingerprint. The MACCS keys were generated using the RDKit package.

DeltaKcat model and architecture

DeltaKcat was developed to predict relative enzyme kinetic parameters between two enzyme-substrate complexes, including relative k_{cat}, relative K_m, and relative k_{cat}/K_m. Instead of predicting the absolute kinetic value of each complex independently, the model takes a pair of enzyme-substrate complexes as input and directly learns the kinetic difference between them. For each complex $i \in [1, 2]$, the protein sequence and substrate SMILES were encoded separately using pretrained language models. The protein sequence was represented by ProtT5 as a residue-level embedding matrix.

$$P_i \in \mathrm{R}^{L_i \times d_p}$$

where L_i is the sequence length and d_p=1024. The substrate SMILES was represented by MolT5 as a molecular embedding.

$$S_i \in \mathrm{R}^{d_s}$$

where d_s=768. To capture enzyme-substrate interactions, DeltaKcat used a gated cross-attention module. The substrate representation was used to query the protein sequence representation, enabling the model to identify enzyme features most relevant to the corresponding substrate. A gating unit was added after the softmax operation in the attention layer to introduce an additional nonlinear modulation step. This design improved the flexibility of attention allocation and strengthened the representation of enzyme-substrate interactions.

The gated cross-attention module was applied independently to the two input complexes, generating two interaction representations, H_1 and H_2. A difference representation was then constructed as

$$H_\Delta = H_1 - H_2$$

The final pair representation was obtained by concatenating the two complex-specific interaction vectors and their difference

$$H_{pair} = H_1 || H_2 || H_\Delta$$

where || denotes vector concatenation. This representation was fed into a multilayer perceptron

$$y = MLP(H_{pair})$$

In this way, DeltaKcat learns relative kinetic changes directly from paired enzyme-substrate complexes.

Data availability

All relevant data supporting the key findings of this study are available within the article and its Supplementary Information files. All the data analyzed in this study is publicly available from either public database, including BRENDA (https://www.brenda-enzymes.org/), SABIO-RK (https://sabiork.h-its.org/), UniProt (https://www.uniprot.org/), PubChem (https://www.uniprot.org/) databases or supplementary datasets of referenced articles (https://github.com/ChemBioHTP/EnzyExtract, https://github.com/JackKuo666/LLM-BioDataExtractor). All data used in the paper can be found in the GitHub repository: https://github.com/LiLabTsinghua/DeltaKcat.

Code availability

In order to facilitate additional utilization, we have made available all of the codes and thorough instructions in our GitHub repository located at https://github.com/LiLabTsinghua/DeltaKcat.

Acknowledgements

We acknowledge financial support from National Key R&D Program of China (2024YFA0920300), Shenzhen Medical Research Fund (A2403013) and the National Natural Science Foundation of China (22478223 and 62532017).

Author Contributions

F.L., Y.C. and X.L. and Z.L. designed the research. X.L. and Z.L. performed the research. X. L., Z. L., J. Z., S. L., W. C., K. W., J. L., Y. C., and F. L. analyzed the data. F.L. and X.L. wrote the paper. All authors approved the final paper.

Competing Interests Statement

The authors declare no competing interests.

Reference:

1. Currin, A., Swainston, N., Day, P.J. & Kell, D.B. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem Soc Rev* **44**, 1172-1239 (2015).
2. Riva, S. Laccases: blue enzymes for green chemistry. *Trends Biotechnol* **24**, 219-226 (2006).
3. Chen, Y. & Nielsen, J. Energy metabolism controls phenotypes by protein efficiency and allocation. *Proc Natl Acad Sci U S A* **116**, 17592-17597 (2019).
4. Meghwanshi, G.K. et al. Enzymes for pharmaceutical and therapeutic applications. *Biotechnol Appl Biochem* **67**, 586-601 (2020).
5. Raveendran, S. et al. Applications of Microbial Enzymes in Food Industry. *Food Technol Biotechnol* **56**, 16-30 (2018).
6. Goldman, S., Das, R., Yang, K.K. & Coley, C.W. Machine learning modeling of family wide enzyme-substrate specificity screens. *PLoS Comput Biol* **18**, e1009853 (2022).
7. Bateman, A. et al. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research* **52**, D609-D617 (2024).
8. Chang, A. et al. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Research* **49**, D498-D508 (2021).
9. 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).
10. Nilsson, A., Nielsen, J. & Palsson, B.O. Metabolic Models of Protein Allocation Call for the Kinetome. *Cell Syst* **5**, 538-541 (2017).
11. Zhang, D. et al. Discovery of Toxin-Degrading Enzymes with Positive Unlabeled Deep Learning. *ACS Catalysis* **14**, 3336-3348 (2024).
12. Cui, H. et al. Enzyme specificity prediction using cross-attention graph neural networks. *Nature* **647**, 639-647 (2025).
13. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems* **34**, 29287-29303 (2021).
14. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
15. Li, F. et al. Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis* **5**, 662-672 (2022).
16. Wang, J. et al. MPEK: a multitask deep learning framework based on pretrained language models for enzymatic reaction kinetic parameters prediction. *Brief Bioinform* **25** (2024).
17. Hua, C. et al. Reactzyme: A benchmark for enzyme-reaction prediction. *Advances in Neural Information Processing Systems* **37**, 26415-26442 (2024).
18. Kroll, A., Rousset, Y., Hu, X.P., Liebrand, N.A. & Lercher, M.J. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *Nat Commun* **14**, 4139 (2023).
19. Shen, X. et al. EITLEM-Kinetics: A deep-learning framework for kinetic parameter prediction of mutant enzymes. *Chem Catalysis* (2024).
20. Wang, Z. et al. Robust enzyme discovery and engineering with deep learning using CataPro. *Nat Commun* **16**, 2736 (2025).
21. Yu, H., Deng, H., He, J., Keasling, J.D. & Luo, X. UniKP: a unified framework for the prediction of enzyme kinetic parameters. *Nat Commun* **14**, 8211 (2023).

22. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* **118** (2021).
23. Honda, S., Shi, S. & Ueda, H.R. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738* (2019).
24. Sapoval, N. et al. Current progress and open challenges for applying deep learning across the biosciences. *Nat Commun* **13**, 1728 (2022).
25. Lyu, B. et al. GotEnzymes2: expanding coverage of enzyme kinetics and thermal properties. *Nucleic Acids Res* **54**, D583-D592 (2026).
26. Qiu, S., Saeed, H., Leonard, W., Li, F. & Yang, A. Machine learning for enzyme catalytic activity: current progress and future horizons. *Brief Bioinform* **27** (2026).
27. Sagi, O. & Rokach, L. Ensemble learning: A survey. *Wiley interdisciplinary reviews: data mining and knowledge discovery* **8**, e1249 (2018).
28. Jimenez-Luna, J. et al. DeltaDelta neural networks for lead optimization of small molecule potency. *Chem Sci* **10**, 10911-10918 (2019).
29. Yu, J. et al. Computing the relative binding affinity of ligands based on a pairwise binding comparison network. *Nat Comput Sci* **3**, 860-872 (2023).
30. Sajeevan, K.A. et al. Robust prediction of enzyme variant kinetics with RealKcat. *bioRxiv* (2025).
31. Malli, A., Vasyutyn, D. & Kim, J.R. Advances in Machine Learning Models for Predicting Enzyme Kinetic Parameters. *J Chem Inf Model* **66**, 42-60 (2026).
32. Wei, G., Ran, X., Ai-Abssi, R. & Yang, Z. Finding the dark matter: Large language model-based enzyme kinetic data extractor and its validation. *Protein Sci* **34**, e70251 (2025).
33. Jiang, J. et al. Enzyme Co-scientist: harnessing large language models for enzyme kinetic data extraction from literature. *BioRxiv*, 2025.2003. 2003.641178 (2025).
34. Elnaggar, A. et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell* **44**, 7112-7127 (2022).
35. Edwards, C. et al. in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing 375-413 (2022).
36. Muir, D.F. et al. Evolutionary-scale enzymology enables exploration of a rugged catalytic landscape. *Science* **388**, eadu1058 (2025).
37. Kerns, S.J. et al. The energy landscape of adenylate kinase during catalysis. *Nat Struct Mol Biol* **22**, 124-131 (2015).
38. Bar-Even, A. et al. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **50**, 4402-4410 (2011).
39. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
40. Song, Y. et al. Accurately predicting enzyme functions through geometric graph learning on ESMFold-predicted structures. *Nat Commun* **15**, 8180 (2024).
41. Kim, S. et al. PubChem 2025 update. *Nucleic Acids Res* **53**, D1516-D1525 (2025).

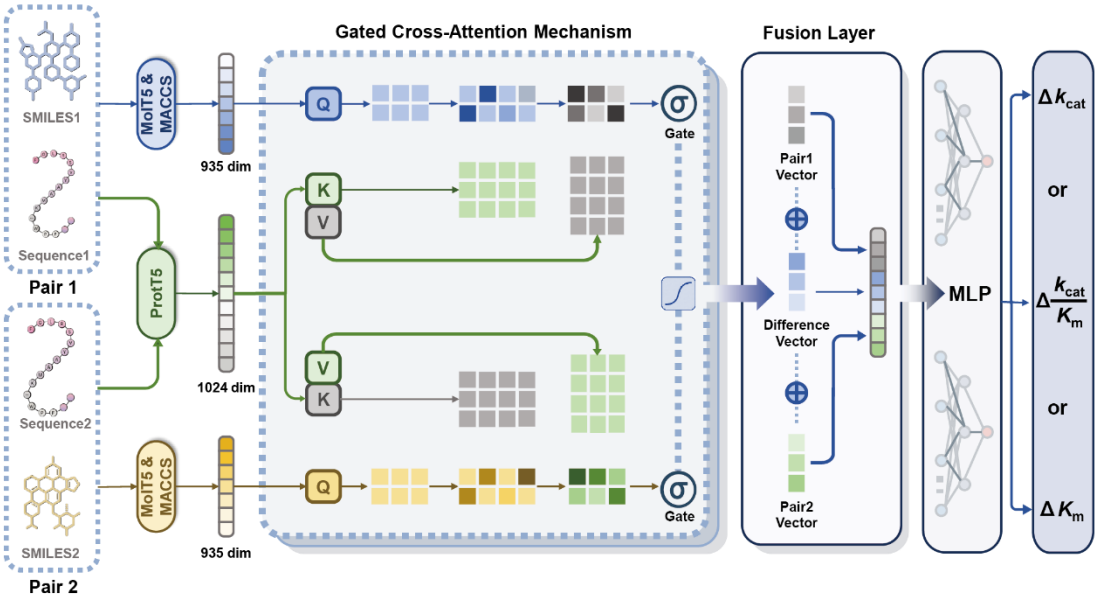

Fig. 1 DeltaKcat framework for predicting relative enzyme kinetic parameters. DeltaKcat takes a pair of enzyme-substrate complexes as input and predicts their relative k_{cat}, K_m, or k_{cat} / K_m values as the output, respectively. Protein sequences are encoded by ProtT5, and substrates are represented using MolT5 and MACCS features. Sequence and substrate embeddings are integrated through a gated cross-attention module, in which substrate-derived queries attend to protein-derived keys and values to capture enzyme-substrate interactions. A gating unit is applied after the softmax step to refine attention allocation. The resulting interaction vectors for the two complexes, together with their difference vector, are concatenated and passed to a multilayer perceptron for prediction.

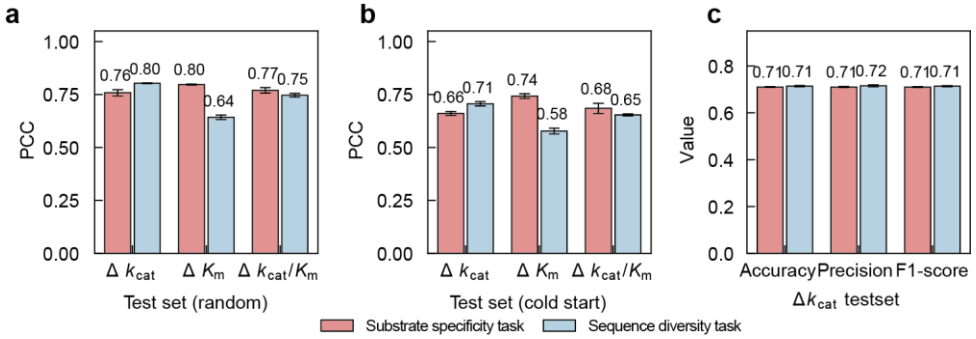

Fig. 2 Prediction of relative enzyme kinetic parameters by DeltaKcat. (a, b) Predictive performance of DeltaKcat on the sequence-diversity dataset and substrate-specificity dataset under random-split (a) and cold-start (b) settings for relative k_{cat}, relative K_m, and relative k_{cat}/K_m. In the cold-start split, each test pair contained at least one protein sequence not observed during training. (c) Classification performance of DeltaKcat for predicting the direction of kinetic changes in the sequence-diversity and substrate-specificity tasks. The data in a-c are presented as mean values, and the error bars represent the standard deviation (s.d.) across three independent experiments using different random seeds.

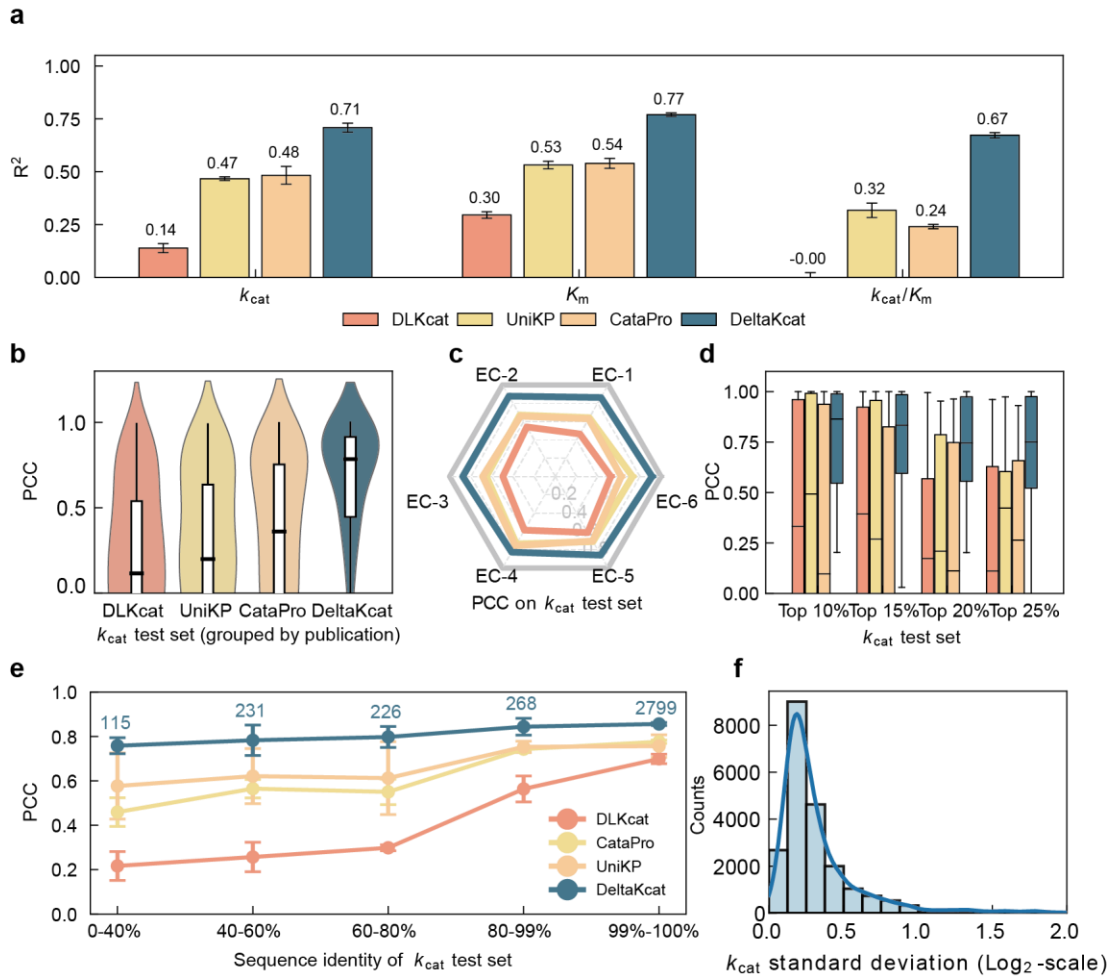

Fig. 3 Absolute kinetic parameter prediction on the sequence-diversity dataset. (a) Comparison of R^2 values for absolute parameter prediction. Performance of DeltaKcat versus DLKcat, UniKP, and CataPro baseline models on inferring absolute k_{cat}, K_m, and k_{cat}/K_m. (b) Publication-level Pearson correlation coefficients for absolute prediction across studies with more than five data points. (c) PCC results of different model based on EC classifications 1-6. (d) Evaluation of k_{cat} prediction (PCC) on subsets of the test set representing the top 10%, 15%, 20%, and 25% of absolute k_{cat} values within each EC class. (e) PCC evaluation across varying thresholds of maximum sequence similarity to the training set. (f) Histogram showing the distribution of the standard deviation of multiple absolute k_{cat} estimates for the same test complex, derived from different training-set reference anchors. The solid blue line is a kernel density estimate. The data in a, e and f are presented as mean values, and the error bars represent the standard deviation (s.d.) across three independent experiments using different random seeds. In each box plot, the lower and upper boundaries of the box represent the first quartile (Q1) and the third quartile (Q3), respectively. The whiskers extend from the quartiles to the minimum and maximum values within 1.5 times the interquartile range. The black line inside the box represents the median.

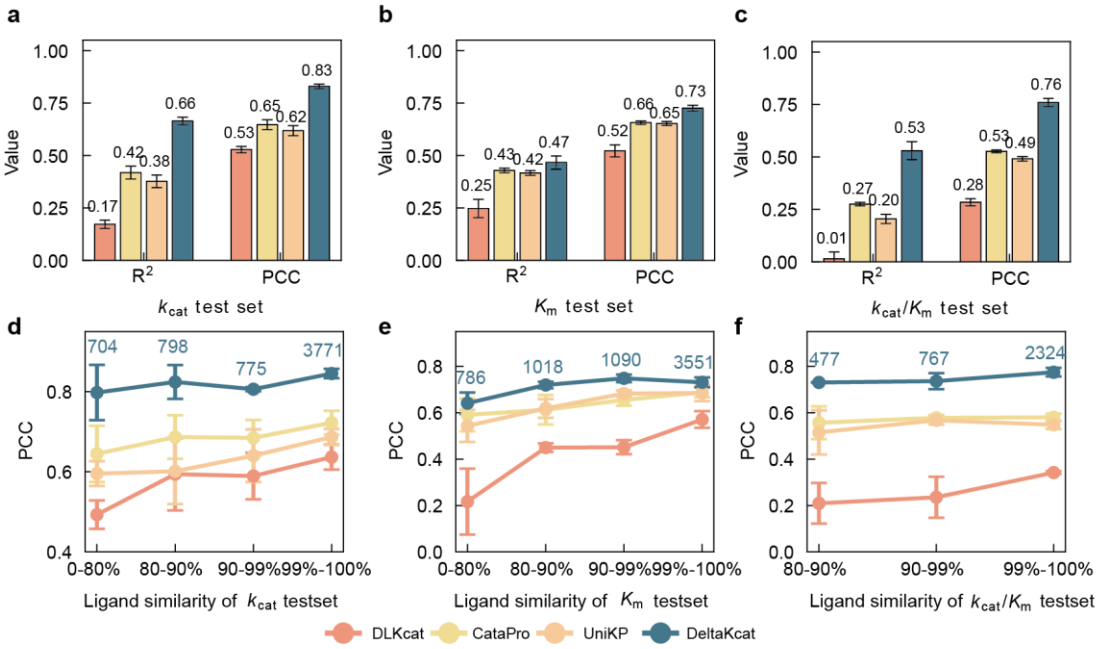

Fig. 4 Absolute kinetic parameter prediction on the substrate-specificity dataset. (a-c) Comparison of R^2 and PCC values for absolute parameter prediction. Performance of DeltaKcat versus DLKcat, UniKP, and CataPro baseline models on inferring absolute k_{cat}, K_m, and k_{cat}/K_m. (d-f) PCC evaluation across varying thresholds of maximum ligand similarity to the training set on k_{cat}, K_m, and k_{cat}/K_m. The data in a-f are presented as mean values, and the error bars represent the standard deviation (s.d.) across three independent experiments using different random seeds.

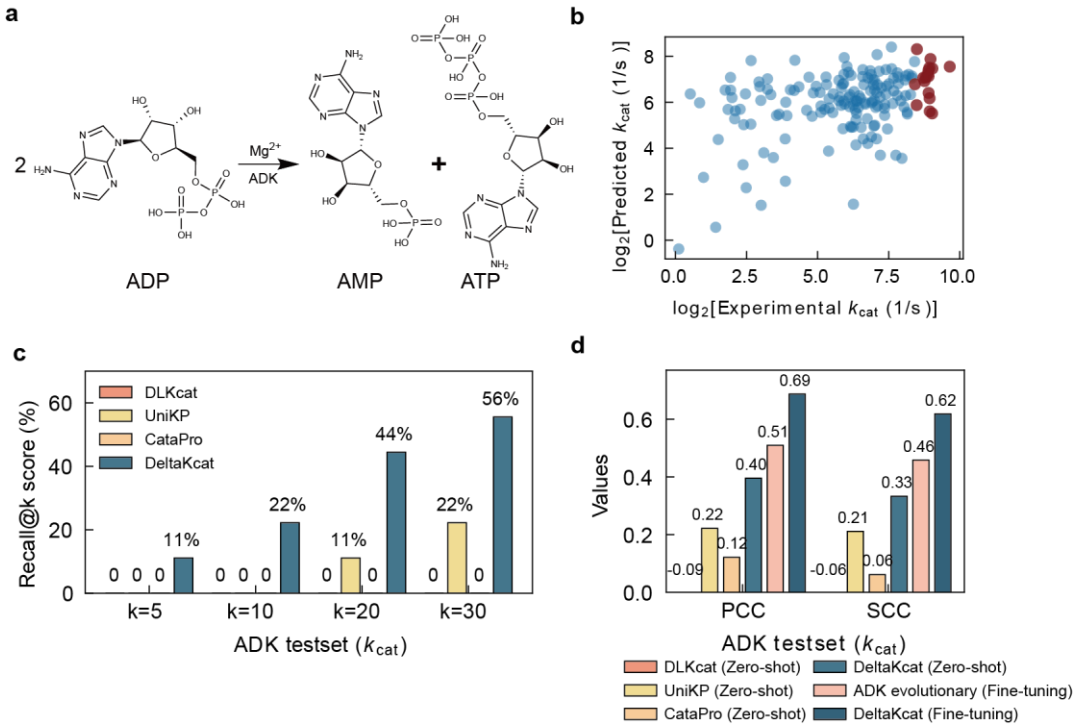

Fig. 5 Evaluation of DeltaKcat on an external adenylate kinase (ADK) dataset. (a) Adenylate kinase catalyzes the reaction. (b) Scatter plot of predicted versus observed absolute k_{cat} values for the external ADK dataset. Red points denote enzymes with observed k_{cat} values in the top 5% of the dataset. (c) Recall of highly active enzymes among top-ranked candidates. Highly active enzymes were defined as the top 5% by true k_{cat}. (d) Comparison of prediction performance between zero-shot and fine-tuned models on the ADK dataset, evaluated by Pearson and Spearman correlation coefficients.