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, *]

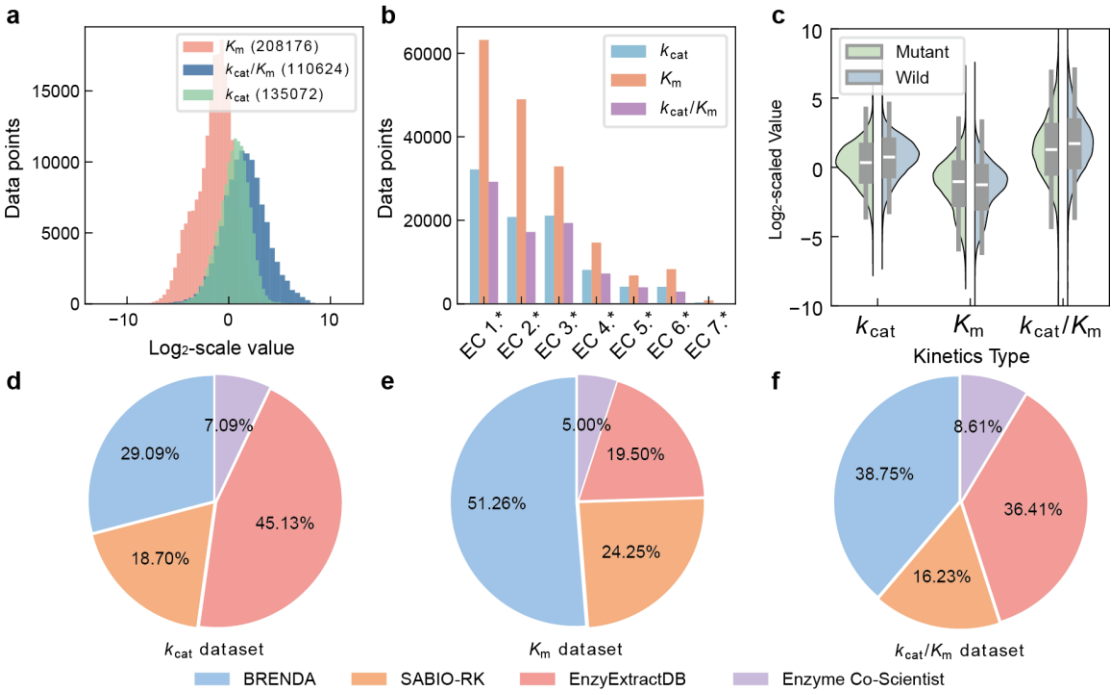
1 Institute of Biopharmaceutical and Health Engineering, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China

2 Key Laboratory for Industrial Biocatalysis, Ministry of Education, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China

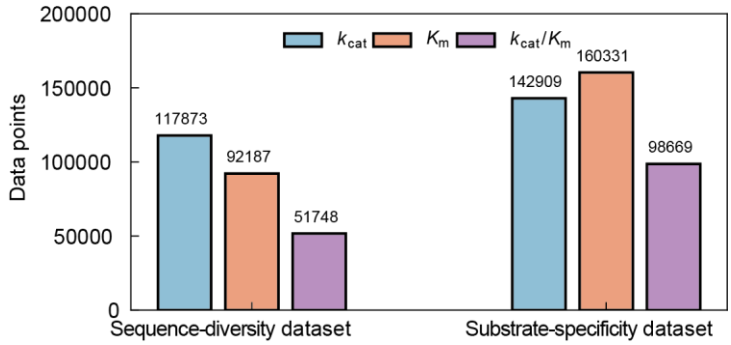
3 Department of Computer Science, Boston University, Boston, MA 02215, USA

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

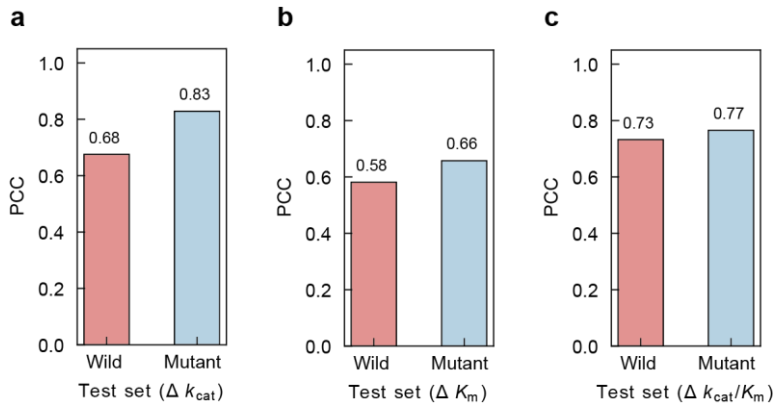
* Corresponding author:
Feiran Li, Email: feiranli@sz.tsinghua.edu.cn



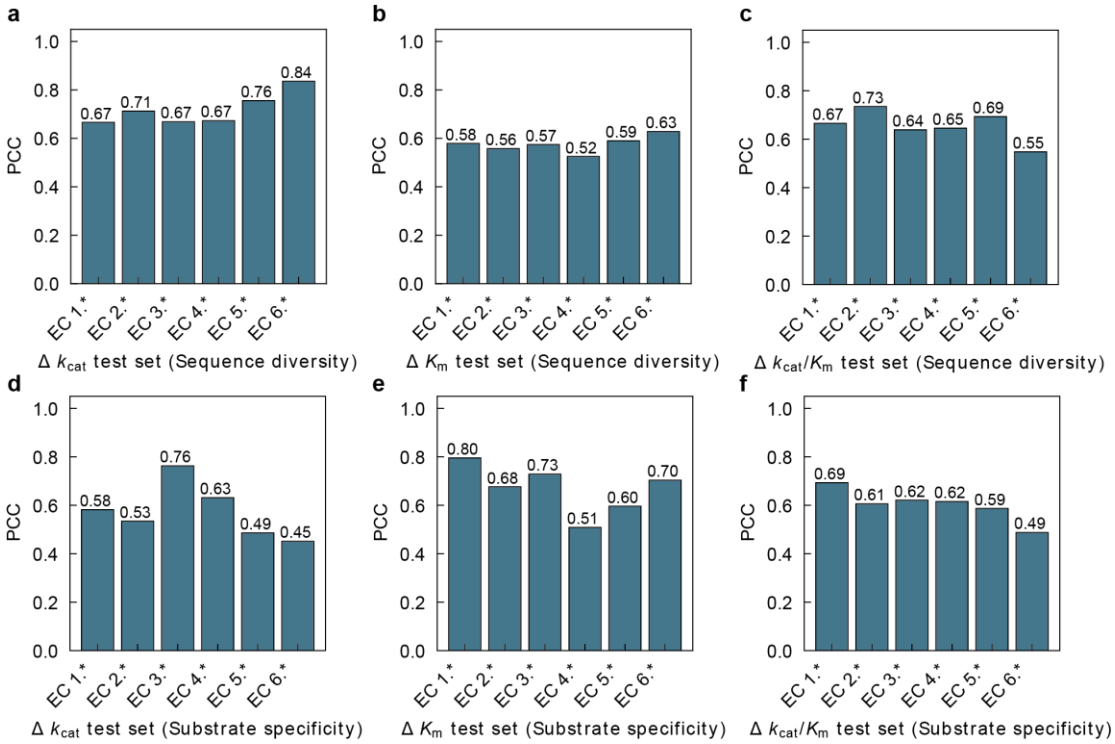
Supplementary Fig. 1. Overview of the processed k_{cat}, K_m and k_{cat}/K_m data. a Distribution of k_{cat}, K_m and k_{cat}/K_m datasets. **b** Number of data points in the k_{cat}, K_m and k_{cat}/K_m datasets classified by the first digit of the EC number. **c** Comparison of $\log_2$-scaled distributions of k_{cat} (s^{-1}), K_m (mM) and k_{cat}/K_m ($mM^{-1}s^{-1}$) between mutant and wild-type enzymes. **d–f**, Distribution of data sources in the absolute k_{cat} (**d**), K_m (**e**), and k_{cat}/K_m (**f**) datasets. Pie charts show the proportions of entries originating from different data sources, with percentages indicated for each source.



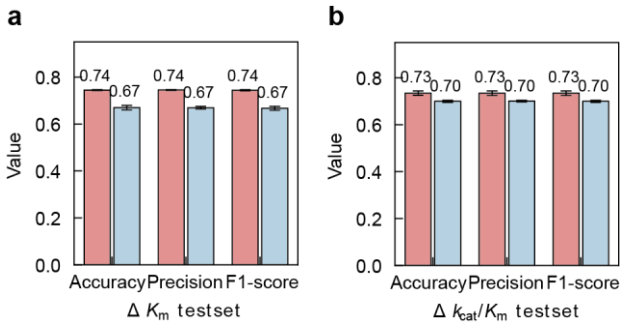
Supplementary Fig. 2. Data sizes of the sequence-diversity and substrate-specificity relative kinetic datasets. The numbers of paired entries for k_{cat}, K_m, and k_{cat}/K_m are shown for the sequence-diversity task and the substrate-specificity task. The sequence-diversity dataset was constructed by pairing records with the same substrate and different protein sequences from the same study, whereas the substrate-specificity dataset was constructed by pairing records with the same protein sequence and different substrates.



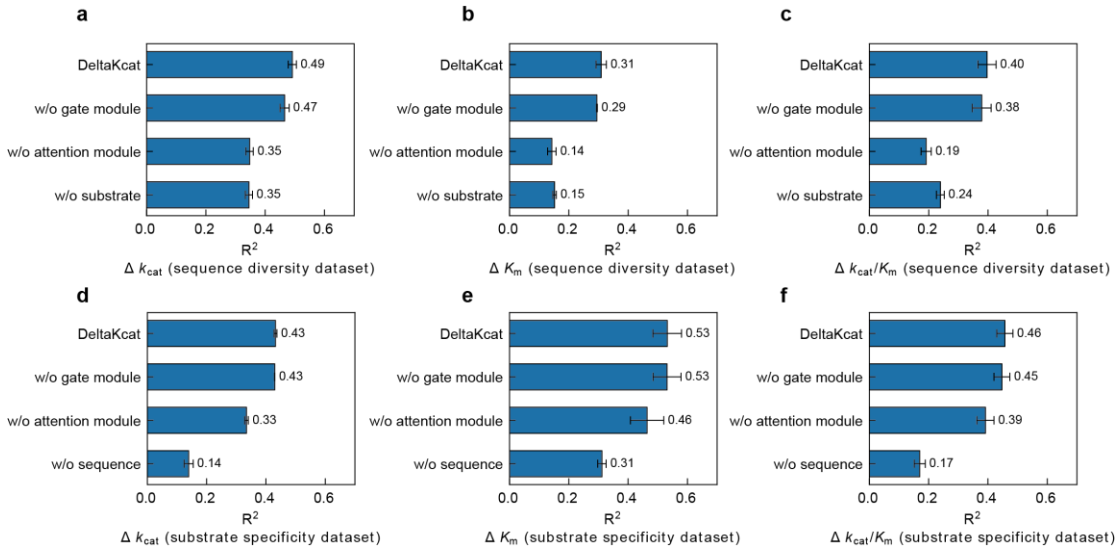
Supplementary Fig. 3. Comparison of PCC values for relative kinetics prediction between mutant pairs and distinct wild-type pairs. Performance of DeltaKcat on Δk_{cat} (**a**), ΔK_m (**b**), and $\Delta k_{cat}/K_m$ (**c**) prediction across different sequence contexts.



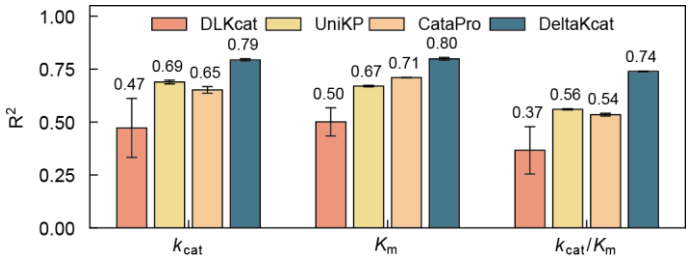
Supplementary Fig. 4. EC class-specific PCC values for relative kinetics prediction. Performance of DeltaKcat across different EC classes on (**a**) Δk_{cat}, (**b**) ΔK_m, and (**c**) $\Delta k_{cat}/K_m$ prediction in the sequence-diversity task. Performance of DeltaKcat across different EC classes on (**d**) Δk_{cat}, (**e**) ΔK_m, and (**f**) $\Delta k_{cat}/K_m$ prediction in the substrate-specificity task.



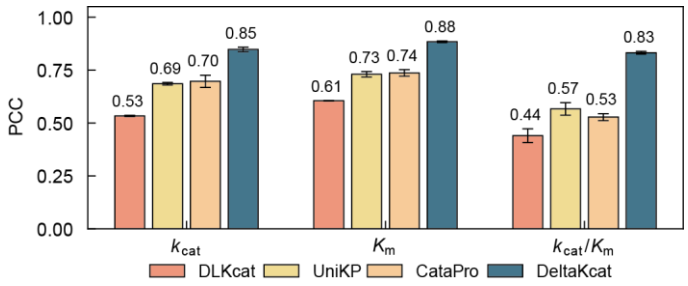
Supplementary Fig. 5. Classification performance of DeltaKcat for predicting the direction of kinetic changes in the sequence diversity and substrate specificity tasks. Classification accuracy for Δ K_m (**a**) and Δ k_{cat}/K_m (**b**) prediction in mutant pairs and distinct wild-type pairs. Error bars represent the standard deviation across three independent experiments using different random seeds.



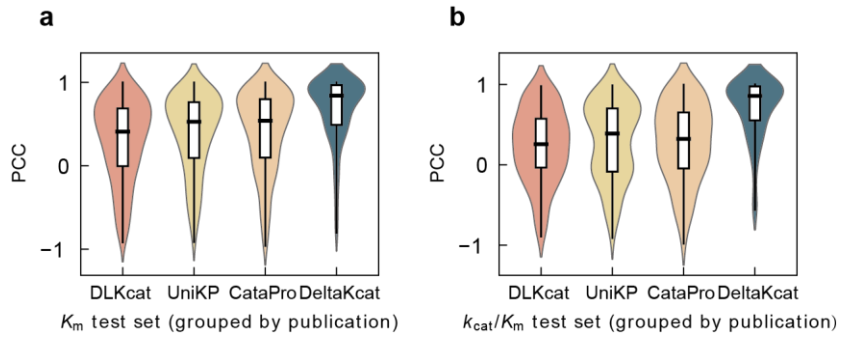
Supplementary Fig. 6. Ablation analysis of DeltaKcat for relative kinetic parameter prediction. a-c, Ablation results on the sequence-diversity task for (**a**) Δk_{cat}, (**b**) ΔK_m, and (**c**) $\Delta k_{cat}/K_m$ prediction. **d-f**, Ablation results on the substrate-specificity task for (**d**) Δk_{cat}, (**e**) ΔK_m, and (**f**) $\Delta k_{cat}/K_m$ prediction. The full DeltaKcat model is compared with variants lacking the gated module, attention module, sequence module or substrate module. Error bars represent the standard deviation across three independent experiments using different random seeds.



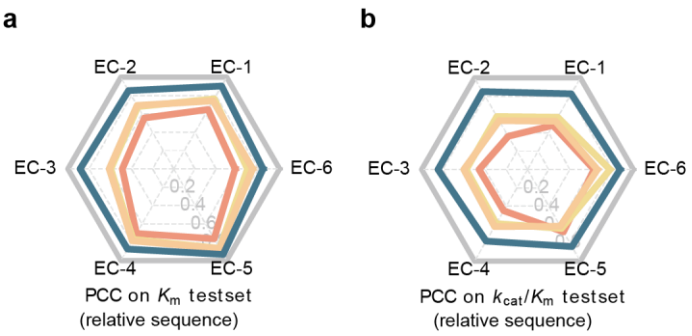
Supplementary Fig. 7. Comparison of model performance for absolute kinetics prediction under the sequence diversity task with random data splitting. R^2 of DLKcat, UniKP, CataPro, and DeltaKcat for k_{cat}, K_m, and k_{cat}/K_m prediction. Error bars represent the standard deviation across three independent experiments using different random seeds.



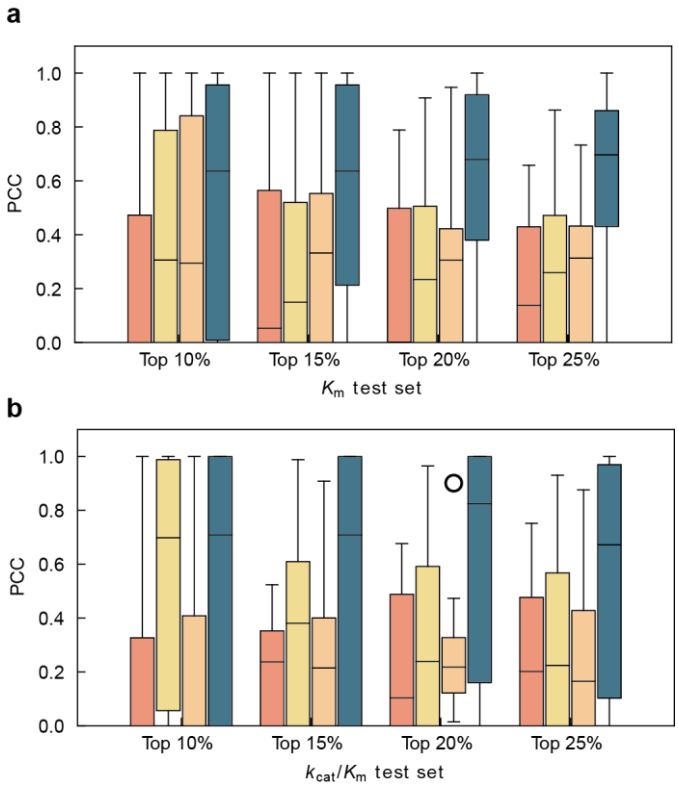
Supplementary Fig. 8. Comparison of PCC values for absolute parameter prediction. Performance of DeltaKcat versus retrained DLKcat, UniKP, and CataPro baseline models on inferring absolute k_{cat}, K_m, and k_{cat}/K_m. Error bars represent the standard deviation across three independent experiments using different random seeds.



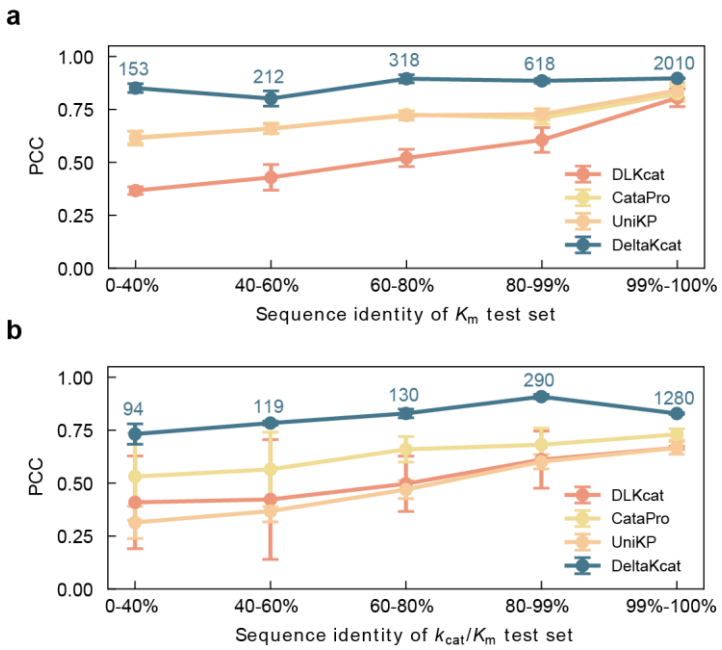
Supplementary Fig. 9. Publication-level Pearson correlation coefficients on K_m (a) and k_{cat}/K_m (b) test set for absolute prediction across studies with more than five data points. The lower and upper boundaries of the box represent the first quartile and the third quartile, respectively. The whiskers 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.



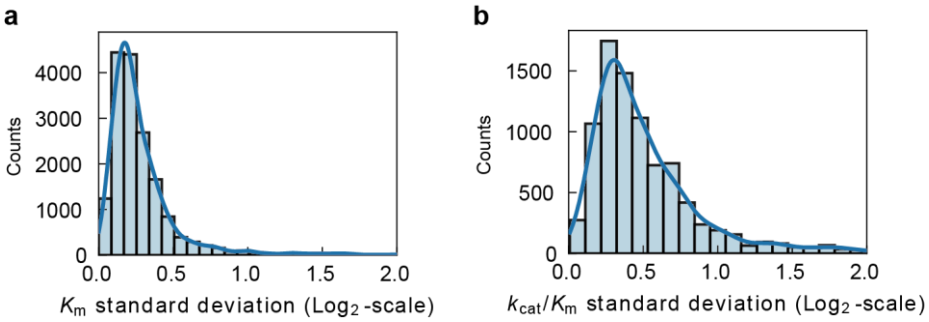
Supplementary Fig. 10. PCC results of different model based on EC classifications 1-6 on K_m (a) and k_{cat}/K_m (b) test set.



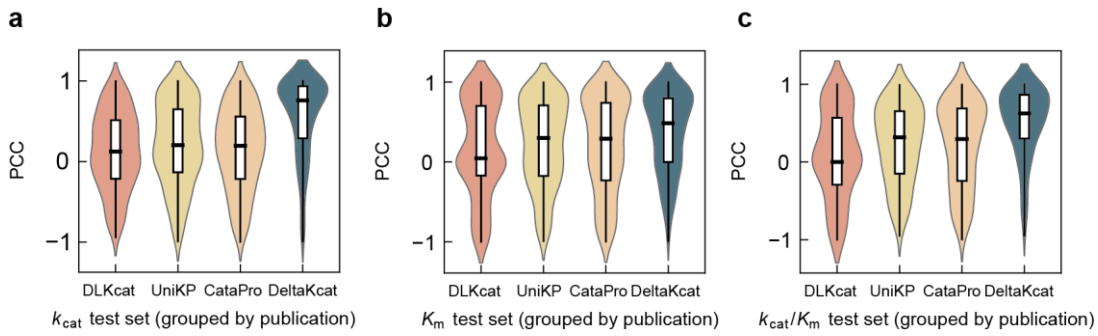
Supplementary Fig. 11. Evaluation of K_m (a) and k_{cat}/K_m (b) prediction (PCC) on subsets of the test set representing the top 10%, 15%, 20%, and 25% of absolute k_{cat} values within each EC class. The lower and upper boundaries of the box represent the first quartile and the third quartile, respectively. 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.



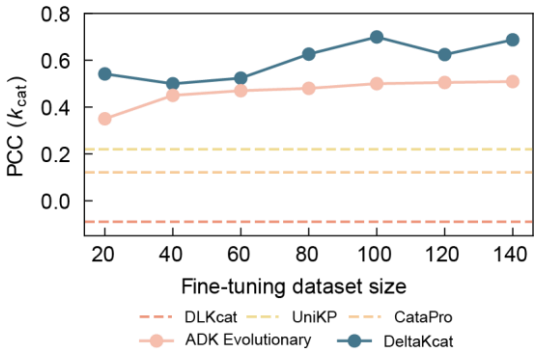
Supplementary Fig. 12. PCC evaluation on K_m (a) and k_{cat}/K_m (b) test set across varying thresholds of maximum sequence similarity to the training set. Error bars represent the standard deviation across three independent experiments using different random seeds.



Supplementary Fig. 13. Histogram showing the distribution of the standard deviation of multiple absolute K_m (a) and k_{cat}/K_m (b) estimates for the same test complex, derived from different training-set reference anchors. The solid blue line is a kernel density estimate.



Supplementary Fig. 14. Publication-level Pearson correlation coefficients on K_m (a) and k_{cat}/K_m (b) test set (substrate diversity task) for absolute prediction across studies with more than five data points. The lower and upper boundaries of the box represent the first quartile and the third quartile, 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.



Supplementary Fig. 15. Improving k_{cat} prediction for ADK sequences with fine-tuning. Improvement in prediction performance after fine-tuning DeltaKcat on increasing numbers of ADK samples, using a fixed test set of 30 samples.