Supplementary material

KcatNet: A Geometric Deep Learning Framework for Genome-Wide Prediction of Enzyme Catalytic Efficiency

Tong Pan[1], Xin Cui[2], Huan Yee Koh[3], Yue Bi[1], Xiaoyu Wang[1], Yumeng Zhang[1], Shantong Hu[2], Geoffrey I. Webb[3], Lukasz Kurgan[4,*], Guimin Zhang[2,*], Jiangning Song[1,5,*]

[1]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia

[2]College of Life Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

[3]Department of Data Science and Artificial Intelligence, Monash University, VIC 3800, Australia

[4]Department of Computer Science, Virginia Commonwealth University, Richmond 23284, USA

[5]Department of Veterinary Biosciences, Melbourne Veterinary School, The University of Melbourne, Parkville 3010, Australia

*Corresponding Authors: Jiangning Song, Email: jiangning.song@monash.edu; Guimin Zhang, Email: 2021500017@buct.edu.cn; Lukasz Kurgan, Email: lkurgan@vcu.edu.

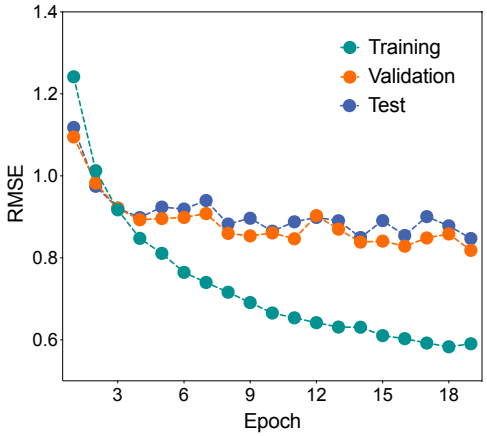

Figure S1 The root mean square error (RMSE) of k_{cat} prediction during model training process for training, validation and test datasets

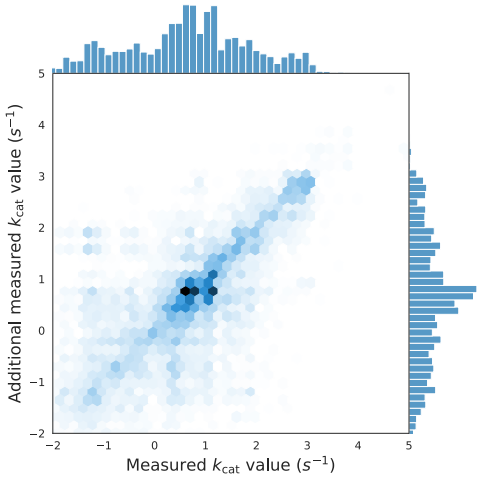

Figure S2 Different measurements for the same enzyme-reaction pair can be very noisy

For all enzyme-reaction pairs with multiple experimentally measured k_{cat}values in our dataset, we plotted k_{cat}values against different measurements for the same enzyme-reaction pair.

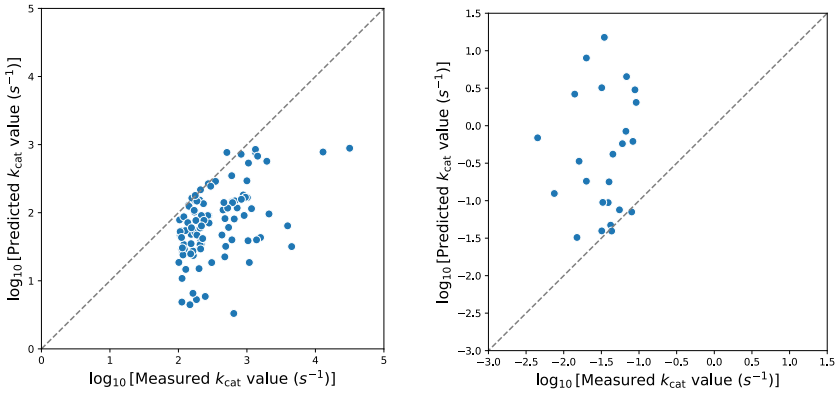

Figure S3 TurNuP performance on extremely low (< 0.1 s^{-1}, left) and high K_{cat} values (> 100 s^{-1}, right)

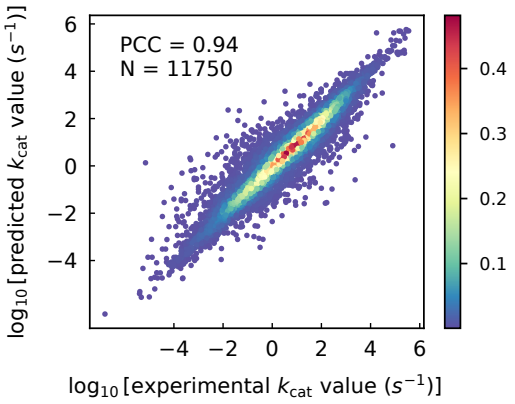

Figure S4 KcatNet performance on the entire dataset

A strong correlation exists between KcatNet's predictions and the experimentally measured K_{cat} values, with a Pearson correlation coefficient (PCC) of 0.94 on the entire dataset.

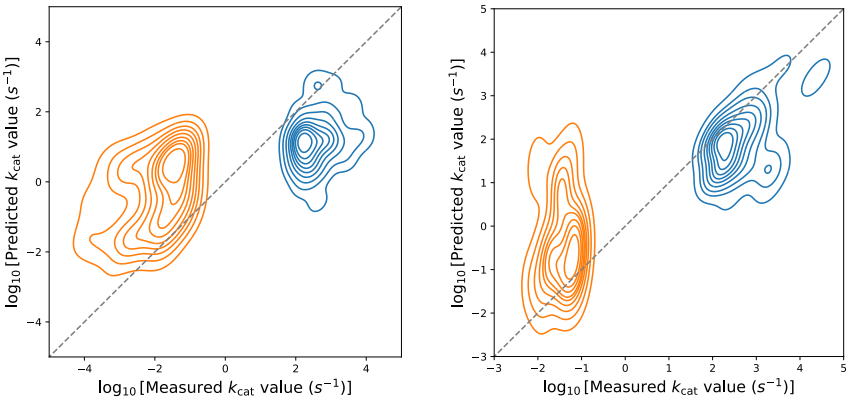

Figure S5 DLKcat (left) and UniKP performances on extremely low (<0.1, left) and high k_{cat} values (>100, right)

The models often systematically overestimate low and underestimate high k_{cat} values.

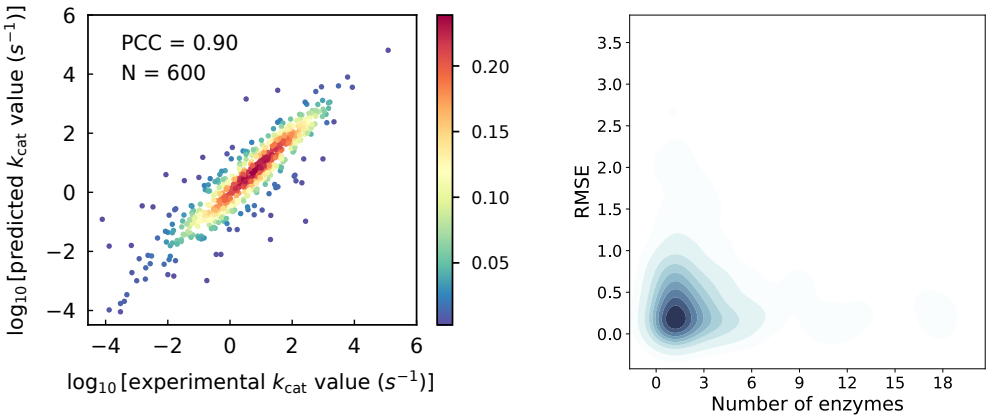

Figure S6 Performance of KcatNet on enzymes very similar to those in the training set

The figure evaluates KcatNet's performance in predicting k_{cat} for enzymes identical to those in the training set, based on Pearson correlation coefficient (PCC, left) and root mean square error (RMSE, right). KcatNet achieves a PCC of 0.90 and an RMSE of 0.35. The impact of the number of identical enzyme sequences in the training dataset on prediction performance is also analyzed.

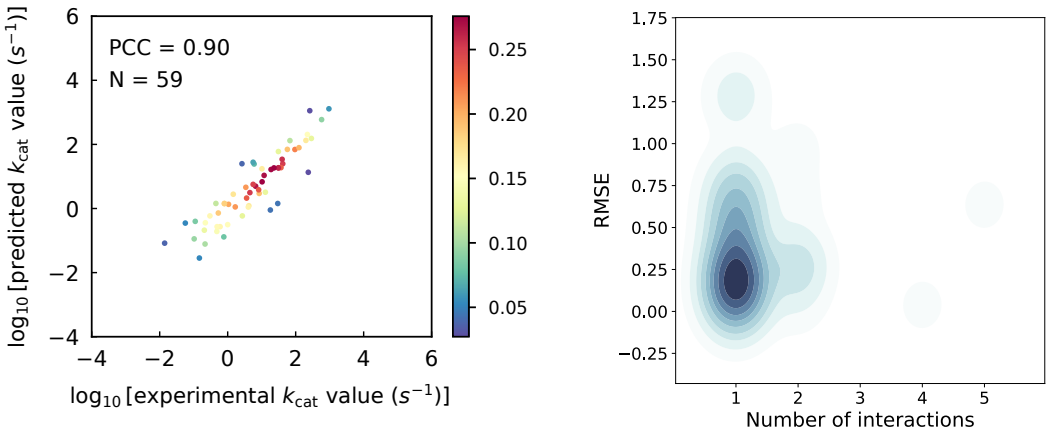

Figure S7 Performance of KcatNet on chemical reactions identical to those in the training set

The figure evaluates KcatNet's performance in predicting k_{cat} for catalytic reactions identical to those in the training set, based on Pearson correlation coefficient (PCC, left) and root mean square error (RMSE, right). KcatNet achieves a PCC of 0.90 and an RMSE of 0.25. The impact of the number of identical catalytic reactions in the training dataset on prediction performance is also analyzed.

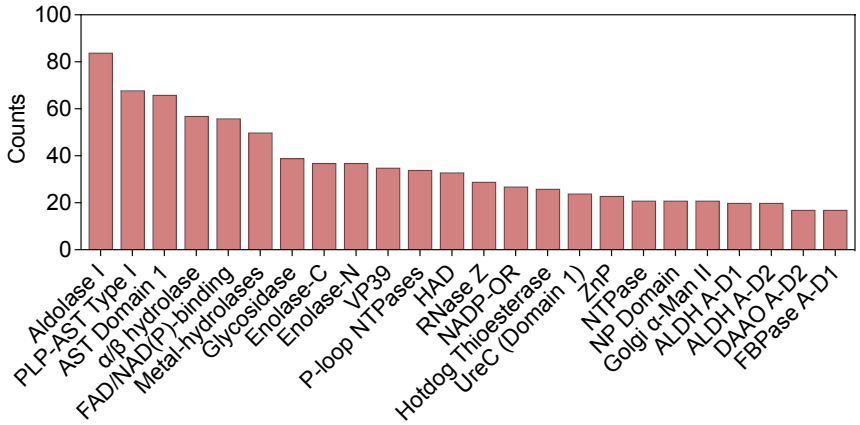

Figure S8 Abundances of various protein families (as appearing in CATH annotations) in constructed benchmark set

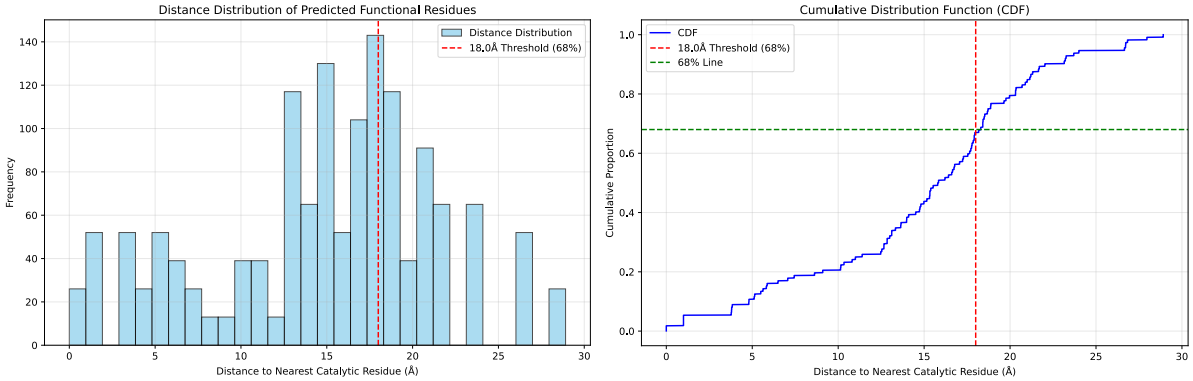

Figure S9 The spatial distribution of identified functional residues by KcatNet

All residues given high activation scores, while located at distinct sequence positions, were spatially close to the catalytic residues. On average, 68% of the are within 18Å of the nearest catalytic residue (left). Cumulative distribution function (CDF) of the experimental K_{cat} values (right). The green dashed line represents the 68% line for reference, and the red dashed line indicates the threshold corresponding to 68% of the data.

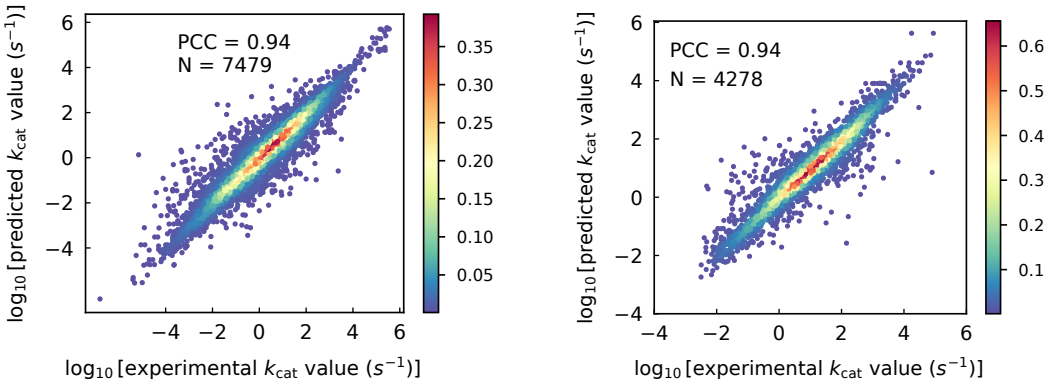

Figure S10 KcatNet performance for both mutated enzymes and wild-type enzymes on the entire dataset

We observed a strong correlation between KcatNet's predicted and experimentally measured k_{cat} values for both mutated enzymes (PCC = 0.94, left) and wild-type enzymes (PCC = 0.94, right).

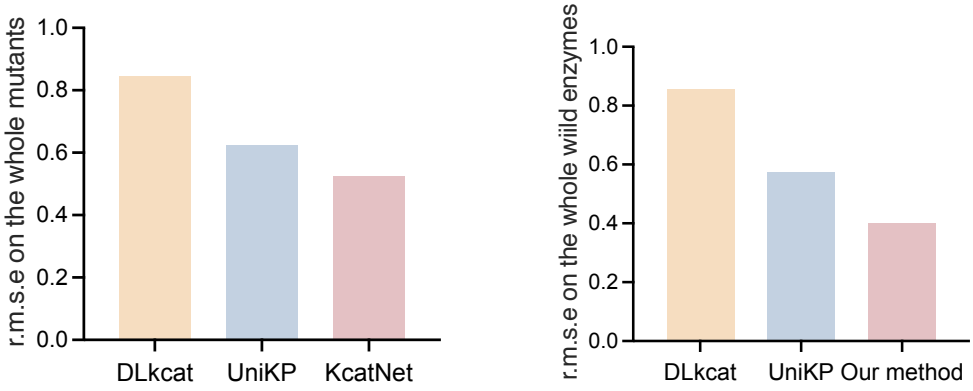

Figure S11 Performance comparison of KcatNet, DLKcat, and UniKP with metric Root Mean Squared Error (RMSE) for mutated enzymes (left) and wild-type enzymes across the entire dataset (right)

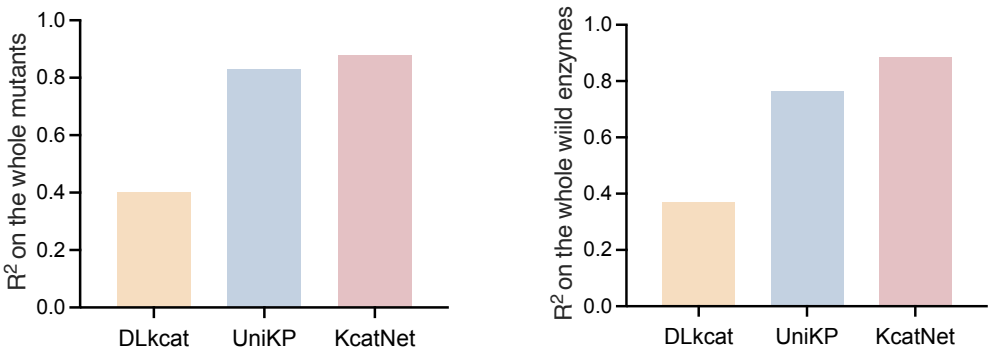

Figure S12 Performance comparison of KcatNet, DLKcat, and UniKP with metric R^2 for mutated enzymes (left) and wild-type enzymes across the entire dataset (right)

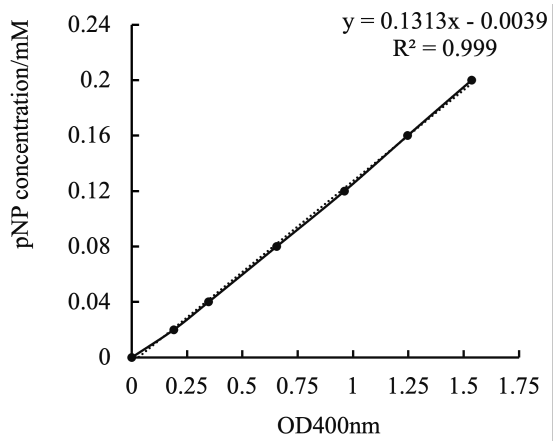

Figure S13 Standard curve correlatingρ-nitrophenol (pNP) concentration with absorbance at 400 nm

The absorbance (OD400) of pNP solutions at various concentrations was measured to generate a standard curve for quantifying pNP concentrations in experimental samples.

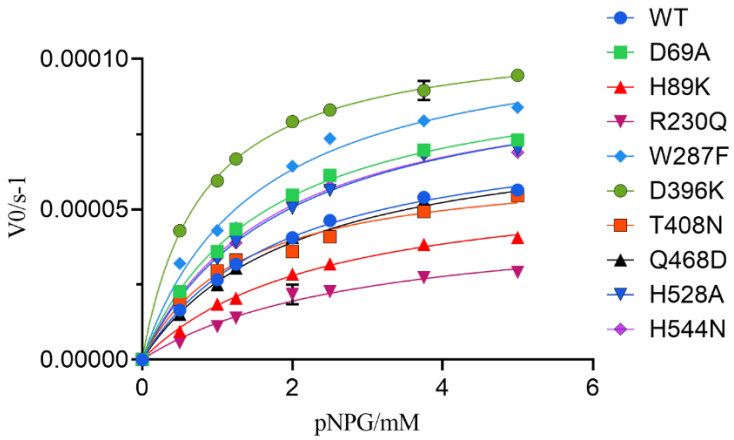

Figure S14 The dependence of initial velocities of purified BspAG13_31A(WT) and mutants on the pNPG concentration

Table S1 The kinetic characteristics of wild-type BspAG13_31A(WT) and mutants

Name	K_{cat}/s^{-1}	K_m/mM	K_{cat}/K_m(s^{-1}/mM)
WT	4.25	1.90	2.24
D69A	5.44	1.65	3.30
H89K	4.60	2.42	1.90
R230Q	3.93	2.92	1.35
W287F	6.25	1.52	4.11
D396K	4.63	0.80	5.79
T408N	3.39	1.32	2.56
Q468D	4.94	2.06	2.39
H528A	5.47	1.92	2.85
H544N	5.73	1.80	3.18

Table S2 Model performance comparison with TurNup

Datasets	Methods	PCC	RMSE	MAE	R^2
Test dataset	KcatNet	0.84	0.78	0.55	0.7
	TurNup	0.8	0.71	0.5	0.62
The whole dataset	KcatNet	0.94	0.49	0.3	0.89
	TurNup	0.94	0.44	0.2	0.86

Table S3 P-values between KcatNet and baselines on test dataset

Benchmark set two-sided t-test P-values between KcatNet and existing methods			P-value
R^2	KcatNet	DLKcat	1.42E-10
R^2	KcatNet	UniKP	4.07E-04
RMSE	KcatNet	DLKcat	4.35E-10
RMSE	KcatNet	UniKP	3.15E-04
MAE	KcatNet	DLKcat	2.39E-11
MAE	KcatNet	UniKP	1.32E-07

Table S4 P-values between KcatNet and baselines for unseen enzymes and reaction

Benchmark set two-sided t-test P-values between KcatNet and existing methods				P-value
R^2	Unseen enzyme	KcatNet	UniKP	2.25E-02
		KcatNet	DLKcat	2.07E-04
	Unseen reaction	KcatNet	UniKP	8.10E-04
		KcatNet	DLKcat	1.19E-05
RMSE	Unseen enzyme	KcatNet	UniKP	2.13E-02
		KcatNet	DLKcat	7.98E-05
	Unseen reaction	KcatNet	UniKP	1.39E-03
		KcatNet	DLKcat	6.38E-06

Table S5 The primers for mutants

The primers name	Sequence (5'-3')
D69A-F	CGACTACAAAGCCATCATGAAAGACTTCGGTAC
D69A-R	GTCTTTCATGATGGCTTTGTAGTCGGAGATGTCG
R230Q-F	GTCACATGAACCAGCCAGGTATCCAAGAACACCTG
R230Q-R	CTTGGATACCTGGCTGGTTCATGTGACCTTCGAAAGACG
H89K-F	CGAAGTTCACAAGCGTGGTATGAAGCTGATCATG
H89K-R	CTTCATACCACGCTTGTGAACTTCGTCCAGCAGTTCGTC
W287F-F	CATGGGTCTTTTCGACAAAGGTGAAGAGAAACC
W287F-R	CACCTTTGTCGAAAAGACCCATGTGTTCGAACTG
N274A-F	GAAGATGGTGGTGCCTTCAACATGATCTTCCAGTTC
N274A-R	GTTGAAGGCACCACCATCTTCCGCAACCCATTC
D396K-F	GTGAGAAAGGTAAGTCTCACGACGAAATCATGAAGG
D396K-R	GTCGTGAGACTTACCTTTCTCACGCTGGATACGGTAG
T408N-F	GTAATCTGGGAAAACGGTCGTGACAACAGCCGTAC
T408N-R	GTCACGACCGTTTTCCCAGATTACCTTCATGATTTCG
Q468D-F	GATCCGTAAAGACCACGACGTACTGGTTTACGG
Q468D-R	CAGTACGTCGTGGTCTTTACGGATCTTGATCAGCTG
H528A-F	GAAGTGCTGATCGCCAACTATCCACTGGACAAGAACG
H528A-R	GTGGATAGTTGGCGATCAGCACTTCGATGTTGGTG
Q540N-F	GAAACTCTGGAAAACTGTACTCTGCATCCGTACG
Q540N-R	GCAGAGTACAGTTTTCCAGAGTTTCGTTCTTGTCCAG
C541I-F	CTCTGGAACAGATCACTCTGCATCCGTACGAAAC
C541I-R	GGATGCAGAGTGATCTGTTCCAGAGTTTCGTTCTTGTCC
H544N-F	GTGTACTCTGAACCCGTACGAAACTCGTGTTTAC
H544N-R	GAGTTTCGTACGGGTTCAGAGTACACTGTTCCAGAGTTTC

Sequence: The synthetic gene GH13_31 α-Glucosidase (BspAG13_31A) from Bacillus sp. AHU2216 after codon optimization

GAATTCATGGAGAAGAAATGGTGGAAAGAAGCGGTAGCGTACCAGATCTATCCGC
GTTCTTTCATGGACTCTAACGGTGACGGTATCGGTGACATCCAGGGTGTTATCAGC
AAACTGGATTACCTGTCTGACCTGGGTATCGACGTTATCTGGATCTGTCCGATCTA
CCAGTCTCCGAACGACGACAACGGTTACGACATCTCCGACTACAAAGACATCATG
AAAGACTTCGGTACTATGGAGGACTTCGACGAACTGCTGGACGAAGTTCACCACC
GTGGTATGAAGCTGATCATGGACCTGGTTATCAACCACACCTCTGACGAACATCCG
TGGTTTCTGGAATCTCGTTCTGCGAAAGAAAATCCGTACCGTGACTACTATATCTG
GCACGAAGGTAAAGACGGTAAAGAACCGAACAACTGGGAATCCATCTTCTCTGGT
TCTGCATGGGAATTTGACGAGAAAACCAAAGAATACTACATGCACGTGTTCTCCA
AGAAACAGCCGGACCTGAACTGGGAGAACGAGAAAGTTCGTCACGAACTGTACG
AAATGGTGAACTGGTGGCTGGACAAAGGTATCGATGGTTTCCGTGTTGACGCGAT
CAGCCACATCAAGAAAGTTGCAGGCTTTCCGGACCTGCCGAACCCGGAGAAACTG
GACTACGTTCCGTCTTTCGAAGGTCACATGAACCGTCCAGGTATCCAAGAACACCT
GAAAGAACTGAAAGAGAAAACCTTCGCGAAATACGACATCATGACCGTTGGTGAA
GCGAACGGTGTTACCTCTGATTCTGCGGATGAATGGGTTGCGGAAGATGGTGGTA
ACTTCAACATGATCTTCCAGTTCGAACACATGGGTCTTTGGGACAAAGGTGAAGA
GAAACCACTGGACCTGATCGAACTCAAGACCATCTTGACCAACTGGCAGAATGGT
CTGGAAAAGATCAACGGTTGGAACGCTCTGTACCTGGAGAACCACGACCAGATCC
GTTCCGTTAACAAGTTCGGTTCCACCGCTTACCGTGTTGAATCTGCGAAATGCCTT
GCTGCACTGTACTTTCTGATGAAAGGTACTCCGTTCATCTACCAGGGTCAAGAACT
GGGTATGACCAACGTTAAATTCGACTCTATCGACGACTACGACGATGTGGGTATG
ATCAACTACTACCGTATCCAGCGTGAGAAAGGTGACTCTCACGACGAAATCATGA
AGGTAATCTGGGAAACTGGTCGTGACAACAGCCGTACTCCGATGCAGTGGAACAC
CGAGAAGAACGCTGGTTTCTCTACCGGTAATCCGTGGATGAAGGTTAATCCGAACT
ACGTTGACATCAACGTTGAAGAGCAGAAATCCGACAAGAACTCCGTACTGAACTT
CTACAAACAGCTGATCAAGATCCGTAAACAGCACGACGTACTGGTTTACGGTACG
TACAAACTGCTGGCTGAAGAAGATTCTGCGATCTACGCGTACACTCGTACTCTGGA
AGGTAAGACTGCTGTTGTTATCTGCAACATGTCTCCAAACAACCAGACCTTCGAAT

TTCCGTCTGAATCCAGCTTCACCAACATCGAAGTGCTGATCCACAACTATCCACTG
GACAAGAACGAAACTCTGGAACAGTGTACTCTGCATCCGTACGAAACTCGTGTTT
ACCTGCTGTCTCTCGAG

Evaluation metrics

Four widely used measures, including coefficient of determination (R^2), the Pearson correlation coefficient (PCC), the root mean square error (RMSE), and the mean absolute error (MAE) were used to evaluate the model performance.

The R^2 ([0,1]) measures how well the observed experimental K_{cat} values are replicated by the predicted K_{cat} values of models, based on the proportion of total K_{cat} values variation in the data explained by models:

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2},$$

where y_i is observed experimental K_{cat} value and $\hat{y}_i$ indicates predicted experimental K_{cat} value , and $\bar{y}$ is the mean of the experimental K_{cat} values.

The PCC ([-1,1]) quantifies the linear relationship between observed and predicted K_{cat} values.

$$\mathrm{PCC} = \frac{\Sigma(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\Sigma(y_i - \bar{y})^2}\sqrt{\Sigma(\hat{y}_i - \bar{\hat{y}})^2}},$$

where y_i and $\hat{y}_i$ are individual data point of the observed experimental K_{cat} value and predicted K_{cat} value separately, $\bar{y}$ and $\bar{\hat{y}}$ are mean values of experimental and predicted K_{cat} values separately.

The RMSE measures the average magnitude of prediction errors by taking the square root of the average squared differences between observed and predicted K_{cat} values. It penalizes large errors more heavily due to squaring.

$$\mathrm{RMSE} = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n}},$$

where y_i is observed experimental K_{cat} value and $\hat{y}_i$ indicates predicted experimental K_{cat} value, n is the number of observed experimental K_{cat} values.

The MAE measures the average magnitude of errors between observed and predicted K_{cat} values, providing a straightforward interpretation of the average error magnitude.

$$\mathrm{MAE} = \frac{\sum|y_i - \hat{y}_i|}{n}$$

where y_i is observed experimental K_{cat} value and $\hat{y}_i$ indicates predicted experimental K_{cat} value, n is the number of observed experimental K_{cat} values.