Electronic Supplementary Information for

EnzyKR: A Chirality-Aware Deep Learning Model for Predicting the Outcomes of the Hydrolase-Catalyzed Kinetic Resolution

Xinchun Ran[1], Yaoyukun Jiang[1], Qianzhen Shao[1], Zhongyue J. Yang[1-5,*]

[1]*Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235, United States*

[2]*Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37235, United States*

[3]*Vanderbilt Institute of Chemical Biology, Vanderbilt University, Nashville, Tennessee 37235, United States*

[4]*Data Science Institute, Vanderbilt University, Nashville, Tennessee 37235, United States*

[5]*Department of Chemical and Biomolecular Engineering, Vanderbilt University, Nashville, Tennessee 37235, United States*

Contents

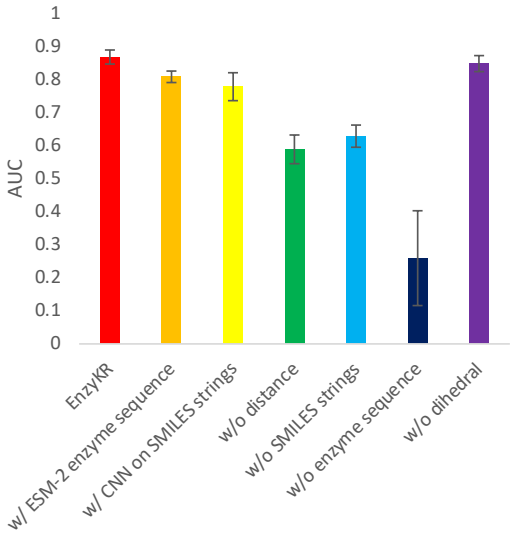

Figure S1. The benchmark of the EnzyKR classifier. The performance is evaluated by area under the curve (AUC) benchmark. The red bar represents the original EnzyKR model, which achieves an AUC of 0.87. The orange bar shows a version of EnzyKR model using evolutionary scaling modeling 2 (ESM-2) embedding[1] for input enzyme sequences instead of one-hot CNN encoding, which results in an AUC of 0.81. The yellow bar illustrates a model employing CNN, instead of GNN, for one-hot encoding of SMILES strings, which yields an AUC of 0.78. The green bar shows a model that excludes substrate-enzyme distance maps from the EnzyKR model, which achieves an AUC of 0.59. The blue bar corresponds to a model excluding SMILES strings from the classifier input, which results in an AUC of 0.63. The dark blue bar indicates a model that excludes enzyme sequences from the model, achieving an AUC of 0.26. The purple bar represents a model that excludes dihedral angles from the model, which resultS in an AUC of 0.85.

Text S1. An alternative sequence encoder for EnzyKR. Using Evolutionary Scaling Modeling-2 (ESM2) embedding for encoding the input enzyme sequences achieves an AUC of 0.81 in the reactive binding pose classification task. The result is comparable to the original architecture with the CNN encoder. This demonstrates the robustness of the model, particularly in scenarios that require alternative sequence encodings for sequences derived from different sources. Our analysis also highlights the importance of encoding structural information, particularly atomic distance maps representing substrate-enzyme interactions, as the second most influential feature. In contrast, dihedral angles have minimal impact on the classification of enzyme-substrate poses, but is likely important for prediction of outcomes of kinetic resolution due to its sharp differentiation of substrate chirality. Furthermore, the inclusion of SMILES strings significantly enhances the model's ability to learn enzyme-substrate pairs, with the exclusion of SMILES strings resulting in a notable AUC drop to 0.63. The utilization of graph convolution layers for encoding substrate SMILES strings further enhances the classifier's performance, underscoring the importance of tailored encoding strategies for chemical data. These findings offer insights into the feature dependencies of the EnzyKR classifier, guiding model construction and data preprocessing strategies in the future.

Text S2. The method used to obtain substrate 3D structure. The substrate SMILES string was converted into 2D connectivity graph with RDKit (https://www.rdkit.org/).[2] For a given substrate graph, the molecules can be categorized as two parts, including: nodes (atom) and edges

(connectivity). The nodes use one hot embedding to generate the atom tensor with a dimension of 10. The edges are directly encoded to the graph neural network based on the order of atom tensor in the substrate graph[3]. The nodes and edges were then fed into the graph convolution layer after a ReLU activation function. For the enantiomeric reactions listed in Figure 4 of the main text, we first constructed the substrates using ChemDraw 22.0.0[4] and obtained the SMILES strings using Chemdraw. We then converted the SMILES strings to sdf files for subsequent docking simulations.

Text S3. The method used to obtain the substrate-enzyme complexes. The hydrolase-substrate complexes were constructed with RosettaLigand docking protocol.[5] First, we adopted the enzyme PDB ID from IntEnzyDB[6] and then applied the clean_pdb.py script in Rosetta to downloaded the structure file. The substrate coordinate files were downloaded from the PubChem by searching the substrate names stored in IntEnzyDB. If a substrate name fails to match with a 3D structure in PubChem, we used the online SMILES Translator (https://cactus.nci.nih.gov/translate/) to generate the 3D sdf file. For each substrate structure sdf file, we generated 250 conformations using BCL::Conf[7] with the default setting. We then used the molfile_to_params.py script to generate the parameter files for the Rosetta docking. Eventually, we run the main script of Rosetta, which performs docking and generates the hydrolase-substrate complex.

Table S1. The benchmark assessed EnzyKR's feature contributions using five performance metrics: Mean Square Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Pearson R, and Spearman R. The first row presents results after removing the substrate-enzyme atomic distance map from both the classifier and regressor. The second row showcases EnzyKR's performance with the incorporation of ESM embedding for encoding enzyme sequences. The third row excludes substrate dihedral information from both the classifier and regressor of EnzyKR. Finally, the last row represents the final EnzyKR model. The analysis reveals that structural information enhances EnzyKR's performance. Notably, distance-related information contributes to lower MSE, while excluding substrate dihedral information results in a higher correlation coefficient.

	MSE	MAE	RMSE	Pearson R	Spearman R
EnzyKR w/o distance map	4.17	2.01	2.04	0.64	0.68
EnzyKR w/ ESM-2 embedding	3.93	1.95	1.98	0.61	0.62
EnzyKR w/o substrate dihedral	3.65	1.49	1.91	0.73	0.73
EnzyKR	3.75	1.54	1.94	0.72	0.72

Table S2. In Table S3, the impact of varying the training dataset size (e.g., 90%:10%, 85%:15%, 80%:20%, 75%:25%, 70%:30% splits) on EnzyKR's performance is examined. The findings reveal a correlation between reduced training dataset size and decreased performance of EnzyKR. As a result of this analysis, we have opted to adopt a training set: test set ratio of 90%:10%, consistently delivering superior performance among the benchmark ratios, with metrics including MSE 3.75, MAE 1.54, RMSE 1.94, Pearson R 0.72, and Spearman R 0.72. Given the current constraints

related to the availability of training data, our choice to maximize the model's exposure to data is aligned with our overarching goal of enhancing its predictive capabilities.

	MSE	MAE	RMSE	Pearson R	Spearman R
EnzyKR (90%:10%)	3.75	1.54	1.94	0.72	0.72
85%:15% training splits	4.96	2.33	2.23	0.48	0.49
80%:20% training splits	5.23	2.38	2.29	0.43	0.40
75%:25% training splits	5.34	2.52	2.31	0.34	0.32
70%:30% training splits	5.67	2.61	2.38	0.22	0.23

Table S3. The comparison of EnzyKR against two other deep learning models: DLKcat[8] and CPI[9]. Five metrics are used, including Mean Square Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Pearson correlation coefficient R, and Spearman correlation coefficient R. Both DLKcat and CPI models were retrained using the training set in this study based on the code reported in the publication.

	MSE	MAE	RMSE	Pearson R	Spearman R
EnzyKR	3.75	1.54	1.94	0.72	0.72
DLKcat on our datasets	3.76	1.74	1.94	0.64	0.63
CPI on our datasets	3.89	1.83	1.97	0.63	0.65

Table S4: The predicted outcomes for kinetic resolution. The first column corresponds to 28 racemic substrates in the ee% value test set. The labels, 1a to 14, are consistent with Figure 4 in the main text. Positive ee% values indicate the situation in which the enzyme prefers reacting with the *S* enantiomer while retaining the *R* enantiomer, whereas negative *ee*% values indicate the opposite. The second column shows the experimental *ee*% values. The third and fourth columns show the predicted *ee*% values using EnzyKR and DLKcat, respectively.

substrates	exp ee%	EnzyKR ee%	DLKcat ee%
1a	>99	78.47	57.46
1b	>99	60.01	45.66
1c	98	91.95	19.87
1d	>99	-87.98	-10.86
1e	>99	78.73	-20.53
1f	98	75.67	9.81
1g	>99	89.84	-0.41
1h	>99	92.53	12.75
1i	>99	81.26	39.79
4j	<-99	85.11	33.71
4k	<-99	-93.08	0.05
4l	<-99	-35.49	0.00
4m	-95	-41.94	-3.71
4n	<-99	-57.60	0.77
4o	-97	-78.40	-61.25
4p	<-99	31.91	-1.13
4q	-97	-2.58	-38.04
4r	<-99	62.31	57.72
7s	36.4	33.90	0.00
7t	<-99	-90.62	0.82
7u	>99	-85.88	0.00
7v	<-99	-0.03	1.63
7w	88.4	47.94	0.00
7y	47	10.17	0.00
7z	91.1	-34.05	0.00
7aa	95.7	-92.49	0.00
10	8.9	92.46	0.00
14	48.7	31.87	8.15

Table S5: The performance of both EnzyKR and DLKcat on the prediction of *ee*% values. The results are evaluated using four metrics: accuracy score, precision score, F1-Score, and recall score. The *ee*% values, with a range of (-100%, 100%), are split into three categories: (-100%, -50%), [-50%, 50%], and (50%, 100%).

	Accuracy	Precision	F1-Score	Recall
EnzyKR prediction 3 bins	0.55	0.53	0.51	0.58
DLKcat prediction 3 bins	0.21	0.55	0.19	0.39

References

(1) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379* (6637), 1123-1130.

(2) Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **2013**, *8*.

(3) Yang, Z.; Zhong, W.; Zhao, L.; Chen, C. Y.-C. Mgraphdta: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical science* **2022**, *13* (3), 816-833.

(4) Brown, T. ChemDraw. *The Science Teacher* **2014**, *81* (2), 67.

(5) Meiler, J.; Baker, D. ROSETTALIGAND: Protein–small molecule docking with full side-chain flexibility. *Proteins: Structure, Function, and Bioinformatics* **2006**, *65* (3), 538-548.

(6) Yan, B.; Ran, X.; Gollu, A.; Cheng, Z.; Zhou, X.; Chen, Y.; Yang, Z. J. IntEnzyDB: an Integrated Structure–Kinetics Enzymology Database. *Journal of Chemical Information and Modeling* **2022**, *62* (22), 5841-5848.

(7) Kothiwale, S.; Mendenhall, J. L.; Meiler, J. BCL:: Conf: small molecule conformational sampling using a knowledge based rotamer library. *Journal of cheminformatics* **2015**, *7* (1), 1-15.

(8) Li, F.; Yuan, L.; Lu, H.; Li, G.; Chen, Y.; Engqvist, M. K.; Kerkhoven, E. J.; Nielsen, J. Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis* **2022**, *5* (8), 662-672.

(9) Goldman, S.; Das, R.; Yang, K. K.; Coley, C. W. Machine learning modeling of family wide enzyme-substrate specificity screens. *PLoS computational biology* **2022**, *18* (2), e1009853.