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

Xinchun Ran[1], Yukunyao 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*

Abstract

Hydrolase-catalyzed kinetic resolution is a well-established biocatalytic process. However, the computational tools that predict the favorable enzyme scaffolds for separating racemic mixture are underdeveloped. To address this challenge, we trained a deep learning framework, EnzyKR, to automate the selection of hydrolases for stereoselective biocatalysis. EnzyKR adopts a classifier-regressor architecture that first identifies the reactive binding conformer of an enantiomer-hydrolase complex, and then predicts its activation free energy. A structure-based encoding strategy was used to depict the chiral interactions between hydrolases and enantiomers. EnzyKR was trained using 204 enantiomer-hydrolase complexes curated from IntEnzyDB, and was tested using a pre-split dataset of 20 complexes on the task of active free energy prediction.

EnzyKR results in a Pearson R of 0.66, a Spearman R of 0.70, and an MAE of 1.48 kcal/mol. EnzyKR was further tested on the task of predicting enantiomeric excess ratios for 18 hydrolytic reactions catalyzed by fluoroacetate dehalogenase RPA1163 and halohydrin HheC, where the performance of EnzyKR was compared against a recently-developed kinetic predictor, DLKcat. EnzyKR outperformed the DLKcat in 13 out of 18 catalytic reactions. EnzyKR provides a novel computational strategy for an accurate prediction of enantiomeric outcome of hydrolase-catalyzed kinetic resolution reactions.

Keywords: Kinetics resolution, Enantiomer-enzyme complex, Hydrolase, Deep learning

1. Introduction

Stereoselective biocatalysis provides strategies to differentiate enantiomers in the synthesis of pharmaceuticals, agrochemicals, and other fine chemicals.[1] Hydrolases have been widely employed for kinetic resolution in industrial chemical synthesis. For instance, lipases and esterases, such as lipase B Candida antarctica (CAL-B),[2] lipoprotein lipase,[3] gluconolactonase, acetylcholine esterase,[4] thermolysin, catalyze the formation of chiral esters with high enantio- or regioselectivity.[5] Dehalogenases, such as fluoroacetate dehalogenase RPA1163, accelerate the stereoselective synthesis of fluorocarboxylic acid.[6] Epoxide hydrolases have been used to generate enantiopure diols and unreacted epoxides for pharmaceutical uses.[7] Chiral biocatalysts receive broad attentions due to their ability to catalyze reactions with high specificity, efficiency, mild operating conditions, and environmental sustainability.

However, for a non-native substrate, identifying biocatalysts with high stereoselectivity for kinetic resolution can be challenging due to the unknown structure-function relationships.[8] To address this, empirical and computational models have been developed to facilitate the prediction of stereoselective outcomes of hydrolase-catalyzed kinetic resolution. In 1998, Kazlauskas et al.[9] established a model that links the size or hydrophobicity of stereocenter substituents with enantioselectivity for ~130 esters derived from secondary alcohols. In 2002, Tomić et al.[10] used quantitative structure-activity relationship (QSAR) analysis to predict the enantioselectivity of *Burkholderia cepacia* lipase (BCL)-catalyzed acylation reactions involving thirteen racemic 3-(aryloxy)-1,2-propanediols. In recent years, machine learning has emerged as a powerful tool to predict stereoselective biocatalytic processes.[11] For one, Cadet et al.[12] developed a machine learning model to predict the impact of mutations on the enantioselectivity for epoxide hydrolase. The model was trained using 512 possible single point mutations variants and achieves an R^2 of 0.81 on the a test set containing 28 mutants. Despite the significant advances in models that specialize in enantiomeric prediction for certain types of hydrolases, the "generalist" models that can predict enantioselectivity across a broad spectrum of hydrolase scaffolds, mechanisms, and substrate types remain undeveloped.[11]

One promising strategy is to directly predict the kinetic parameters for an enzymatic reaction, because the apparent selectivity in kinetic resolution directly connects to the difference of hydrolytic rates between enantiomers. In recent years, the predictive models for enzyme turnover number (i.e., k_{cat}) have been developed for metabolic engineering.[11] For example, Heckmann et al.[13] used elastic net regression, random forest, and deep neural network models to predict k_{cat} values in *Escherichia coli*, achieving a cross-validated Pearson R^2 value of 0.31 for k_{cat} and 0.76 for $k_{app,max}$. Li et al[14]. developed a deep learning model, DLKcat, to predict genome-scale

k_{cat} values for over 300 yeast species, achieving a Pearson R value of 0.94. However, one major pitfall in the existing models is lack of chirality representation of the substrates. As such, these models likely fail in the task of enantiomeric prediction.

To address this limitation, here we developed a deep learning model, EnzyKR, to predict the enantiomeric outcome of hydrolase-catalyzed kinetic resolution reactions. EnzyKR adopts a graph neural network architecture and a multi-task learning approach to predict k_{cat} values for hydrolase-enantiomer pairs. Distinct from existing k_{cat} predictors, EnzyKR encodes the chirality information of substrates through geometric features extracted from hydrolase-enantiomer pairs. As the difference of k_{cat} values between enantiomers informs stereoselectivity, EnzyKR can be potentially used to screen and select hydrolase scaffolds for stereoselective biocatalysis applications.

2. Computational Methods

Model design and architecture. EnzyKR is comprised of a classifier and a regressor. The classifier identifies the reactive hydrolase-enantiomer complexes from unreactive ones. The input data for the classifier involve the complex structure, enzyme sequence, and simplified molecular-input line-entry system (SMILES) string. The complex structure is represented as a distance map, with the distances between the substrate's geometric center and specific residues encoded using a 2D convolutional neural network (CNN) with three layers. The distance map encoder has a filter size of 11, a padding size of 1, and a ReLU activation function, and produces 1673 x 512 tensors as output. EnzyKR also employs an enzyme sequence encoder, which takes in the enzyme sequence profile generated by aligning against the UniRef50[15] database using HMMER[16]. The resulting multiple sequence alignment (MSA) of the enzyme is then processed through 2D CNN layers that involve an identical architecture to the distance map encoder. The enzyme sequence

encoder produces 2385 x 512 tensors as output. To encode the substrate SMILES strings, EnzyKR uses a graph neural network (GNN) encoder with three graph convolution layers (Supporting information, Text S1).[17] The RDKit package was used to represent the topology of substrates by separating their atoms and bonds into nodes and edges for use in the GNN encoder.[18] The input dimensions for the graph convolution layer and the multilayer perceptron layer are both 16. The output of the classifier uses the cross-entropy loss function to evaluate the predictive accuracy for the binary classification of reactive versus unreactive hydrolase-enantiomer complexes.

In the regressor, the input involves the embeddings of the classifier concatenated with the interaction map derived from the structures of the hydrolase-enantiomer complexes. Different from the distance map used in the classifier that only measures the distances between the geometric center of the substrate and the hydrolase residues, the interaction map stacks the distances between each atom of the substrate and the C_α and C_β atoms of the hydrolase residues in one matrix. To encode the embeddings, the regressor uses one module of cross-attention with 10 attention heads and a dropout rate of 0.5. The attention module is followed by residual blocks to extract features with a dimension of 8316×512 from the cross-attention embeddings. The residual blocks consist of three 2D dilated convolution layers with a filter size of 11 and a padding size of 1, one 2D batch norm layer, and one ReLU layer. Subsequently, two layers a of fully connected neural network (i.e., multiple-layer perceptron) is employed to conduct regression between the extracted feature and the activation free energy (i.e., $\Delta G^{\ddagger}$).

Data curation. The training data consists of the enzyme sequences, substrate SMILES strings, and hydrolase-substrate complexes. The training data contains 204 hydrolase-substrate complexes and the test data contains 20 complexes (Supporting Information, dataset.zip). The dataset involves 63 distinct types of hydrolases and 182 distinct types of substrates (i.e., 111 chiral

versus 71 achiral substrates). The data for hydrolase sequence, structure, substrate SMILES, and enzyme turnover rate (i.e., k_{cat}) were curated from IntEnzyDB, an integrated enzyme structure-kinetics database developed by our lab.[19, 20] The dataset contains 12 subclasses of hydrolase based on the enzyme commission (EC) number. The major subclasses are 3.1 and 3.2 – they have 63 and 56 enzymes, respectively. There are 27 enzymes shared by both subclasses.

The structural models for hydrolase-enantiomer complexes were constructed using RosettaLigand[21] (Supporting Information, Text S2). Each substrate sdf file was obtained from PubChem API by searching their SMILES string. Conformational sampling was conducted for each substrate to generate 250 conformers using BCL::Conf web interface from the Meiler Lab.[22] These conformers were used as input to dock into the active site of their corresponding hydrolase using RosettaLigand. The docked hydrolase-enantiomer complexes were divided into two categories based on the spatial proximity between enzymes' catalytic residues (i.e., catalytic triad) and the geometric center of the reacting functional group on the substrate. If the distances are all within 4.0 Å, the substrate-enzyme complexes were classified as reactive substrate-enzyme complexes. Otherwise, the complexes were classified to be unreactive. Each reactive complex was also visually inspected to ensure optimal positioning of the substrate into the active site. In total, we curated 224 reactive hydrolase-enantiomer complexes versus 448 unreactive ones. To examine the capability of EnzyKR to differentiate enantiomers, we curated an independent test set of 18 hydrolytic reactions catalyzed by fluoroacetate dehalogenase RPA1163 (PDB ID: 5K3F)[6] and halohydrin HheC (PDB ID: 1PWX)[23]. The data for the enantiomer excess (*ee*) ratio were manually curated from the publication. For each of the 36 hydrolase-enantiomer complexes, we adopted the above-mentioned docking approach to build the structural model.

3. Results and Discussion

3.1 The Model Architecture of EnzyKR

EnzyKR is a deep learning model designed for predicting the activation free energy of a hydrolase-substrate complex in a chirality-resolved fashion. EnzyKR consists of two parts: a classifier and a regressor. The classifier distinguishes reactive hydrolase-enantiomer complexes from unreactive binding poses, while the regressor predicts the hydrolytic activation free energy (i.e., $\Delta G^{\ddagger}$) for the reactive complex. The classifier employs different neural network architectures to separately encode enzyme sequences, substrate SMILES strings, and the distance map between the substrate enantiomer and enzymes (detailed in the Computational Methods section). Notably, the distance map contains the center-of-mass distances between the substrate and enzyme residues. The representation informs the spatial distribution of substrate and nearby active site residues and is invariant to the translation and reflection of cartesian coordinates. The classifier adopts cross-entropy in its loss function for binary classification of reactive versus unreactive substrate-enzyme complexes. The regressor of EnzyKR takes input from both the classifier embedding and substrate-enzyme interaction maps. Unlike the distance map encoded by the classifier, the interaction map stacks the matrices of atomic distances between each atom on the substrates and the atoms of the enzyme residues (i.e., C_{α} and C_{β}).

The regressor leverages a cross-attention module to encode a representation matrix that concatenates the embedding of the classifier with the substrate-enzyme interaction map. The representation matrix is fed into a two-layer residual block to extract features from the cross-attention embeddings. These features are then used to predict the $\Delta G^{\ddagger}$ value of a hydrolase-substrate complex through a two-layer multiple-layer perceptron (MLP) neural network.

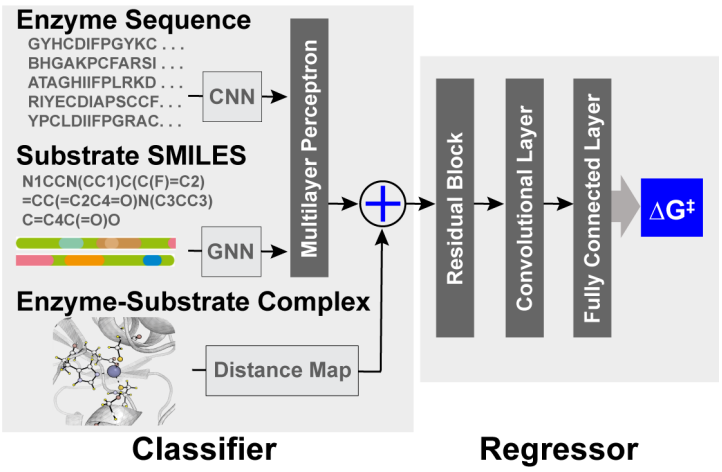

Figure 1. The architecture of EnzyKR. The classifier takes in substrate-enzyme distance maps, enzyme multiple sequence alignment, and substrate SMILES strings to determine whether the hydrolase-substrate complex is reactive or not. The embeddings generated from the classifier are passed to the regressor along with the substrate-enzyme interaction map to predict the activation-free energy. CNN refers to convolutional neural network. GNN refers to graph neural network.

Compared to existing deep learning models that predict k_{cat} or $\Delta G^{\ddagger}$ for enzyme catalysis,[11, 13, 14] the novelty of EnzyKR architecture manifests in three aspects. First, EnzyKR explicitly encodes chirality involved in the interactions between hydrolase and substrate enantiomers in the form of distance map and interaction map for the classifier and regressor, respectively. In comparison, existing predictive models for enzyme kinetics do not include features that describe spatial relationships between enzyme and substrate atoms. Second, EnzyKR uses a cross-attention mechanism to extract important features from hydrolase sequence, substrate SMILES strings, and the interaction map. This allows the model to effectively identify the most relevant encoded features for downstream prediction tasks. Third, EnzyKR employs a GNN to encode the substrate's topology, which is likely to encode atomic connectivity more effectively than one-hot embedding. Notably, new encoding strategies for molecular structures have been developed that preserve chiral

information, such as ChIRo[24] and SELFIES[25]. These methods present as potential alternatives for the future development of EnzyKR.

3.2 The Training and Test Dataset of EnzyKR

The dataset used for training and testing EnzyKR includes 224 hydrolase-substrate enantiomer complexes curated from 13 enzyme commission subclasses under the category of hydrolases (left, Figure 2). The most populated subclasses are 3.1 (e.g., esterases and lipases) and 3.2 (e.g., amylase), which have 63 and 56 members, respectively. The distribution of $\Delta G^{\ddagger}$ values (i.e., converted from k_{cat} using Eyring's equation, eq1) ranges from 5.0 to 23.0 kcal/mol, with an average of 16.4 kcal/mol (right, Figure 2).

$$\Delta G^{\ddagger} = -RT\ln\left(\frac{k_{cat}h}{k_B T}\right) \qquad \text{eq. 1}$$

In this equation, R is the gas constant, T is temperature, h is the Planck constant, and k_B is the Boltzmann constant. The hydrolase with the lowest $\Delta G^{\ddagger}$ is 3',5'-cyclic-AMP phosphodiesterase (i.e., EC = 3.1.4.53), which hydrolyzes the second messenger 3',5'-cyclic AMP (cAMP), and the one with the highest $\Delta G^{\ddagger}$ is acylaminoacyl-peptidase (i.e., EC = 3.4.19.1), which cleaves an N-acetyl or N-formyl amino acid from the N-terminus of a polypeptide. A large proportion of the curated data (i.e., 83.5%) has an activation free energy between 12.2 and 21.2 kcal/mol. The wide distribution of $\Delta G^{\ddagger}$ values reflects the diversity of catalytic performance of hydrolases. We partitioned the dataset into training and test sets based on hydrolase sequence identity. Among the 224 hydrolase-enantiomer complexes, we selected 20 complexes whose sequence identities are less than 85% from each other in the test set, leaving the remaining 204 complexes in the training set. A lower hydrolase sequence identity in the test set creates a more challenging task for the generalizability of EnzyKR.

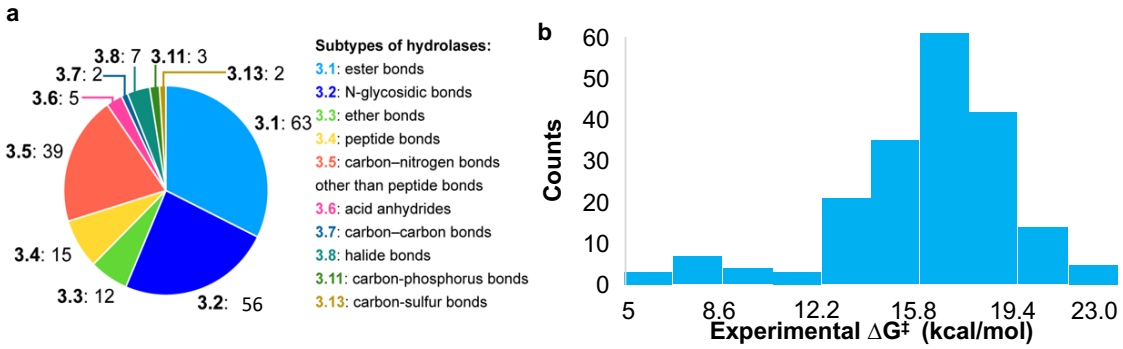

Figure 2. Statistics of the curated dataset used for developing EnzyKR. (a) Distribution of enzyme commission (EC) subtypes for the hydrolases used in this work. The specific hydrolase subtypes as well as their EC numbers (up to the second digit) are labeled on the right-hand side of the pie chart. (b) Distribution of activation free energy, $\Delta G^{\ddagger}$ for a total of 224 hydrolase-substrate enantiomer complexes, in which $\Delta G^{\ddagger}$ values are converted from k_{cat} using Eyring's equation shown in eq1. The bin size is 1.8 kcal/mol.

3.3 The Performance of EnzyKR

The performance of EnzyKR was evaluated using both the training and test sets. To assess the classification ability of EnzyKR's classifier component, we employed the area under the curve (AUC) metric. The classifier of EnzyKR achieves an AUC of 0.87. Reduction in AUC was observed upon removal of enzyme sequence or substrate SMILES strings from the input features. Replacing GNN by CNN for encoding the SMILES strings also decreases the AUC (Supporting Information, Figure S1). We used both Pearson and Spearman correlations to examine the regressor of EnzyKR for its ability to predict the value and rank of activation free energies of different hydrolase-substrate complexes. The parity plot for the training set (204 data points) shows a decent linear correlation with a Pearson R of 0.91, Spearman R of 0.86, and a mean absolute error (MAE) of 0.8 kcal/mol. On the test set, EnzyKR achieves a Pearson R of 0.66,

Spearman R of 0.70, and MAE of 1.5 kcal/mol. For both training and test sets, the value of Spearman R resembles that of Pearson R, which indicates that EnzyKR balances the regression of target values or ranking without overfitting. The drop of EnzyKR performance on the test set is likely due to the small sample size. However, we should note that curating high-quality structure-sequence-kinetics dataset is intrinsically challenging. In our integrated structure-kinetics database IntEnzyDB,[20] the total number of hydrolase-substrate pairs is only 355, where the hydrolase mutants and unstructured substrate (e.g., cellulose) have been removed for the development of EnzyKR.

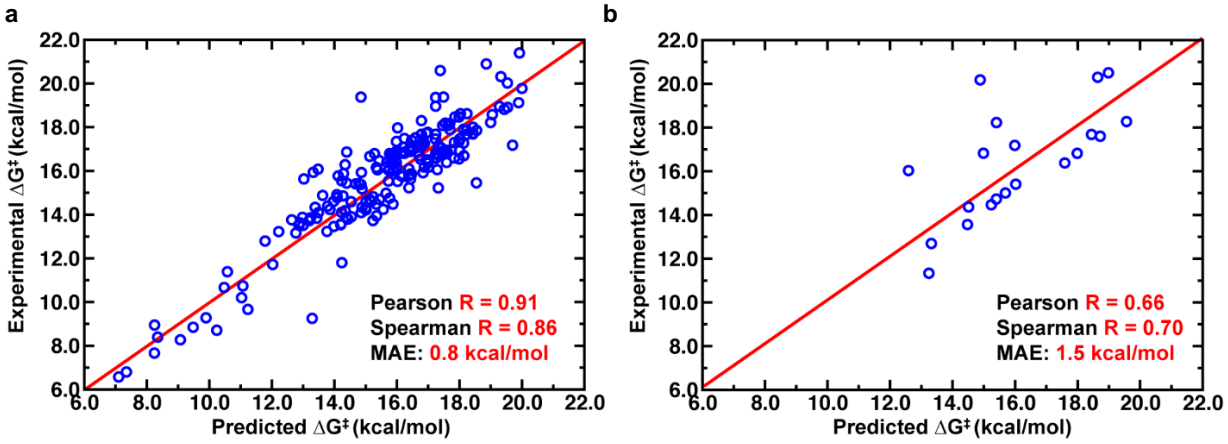

Figure 3. The performance of EnzyKR on the training set and test set. (a) EnzyKR was trained on a dataset comprising 204 substrate-enzyme complexes and achieved a Pearson correlation of 0.91, a Spearman correlation of 0.86, and a mean absolute error (MAE) of 0.83 kcal/mol. (b) To evaluate the model's performance on unseen data, it was tested on a pre-splitted dataset of 20 substrate-enzyme complexes, resulting in a Pearson correlation of 0.66, a Spearman correlation of 0.70, and an MAE of 1.48 kcal/mol. The red line indicates the prediction values are equal to the experimental values.

To assess the contribution of various features to EnzyKR, we benchmarked the impact of removing various features on the predictive accuracy of the regressor (Supporting Information,

Table S1). With the removal of interaction map from the input, we observed a decrease in both Pearson and Spearman R values to 0.53 and 0.51, respectively, as well as an increase in MAE to 2.2 kcal/mol. Similarly, with the removal of the classifier embedding from the classifier output layer, we observed a significant decrease in Pearson and Spearman R values to 0.58 and 0.61, respectively, and an increase in MAE to 2.0 kcal/mol. In a subsequent analysis, we removed the SMILES strings of substrates from the input of the classifier. This led to a drop in both Pearson and Spearman R values to 0.6 and an increase in MAE to 2.0 kcal/mol. These results suggest that both the interaction map and substrate SMILES strings are essential features that contribute positively to the predictive accuracy of the EnzyKR regressor.

Furthermore, we compared the performance of EnzyKR against two predictors: DLKcat,[14] a deep learning k_{cat} predictor, and a compound-protein interaction (CPI) model[26] that predicts the substrate-enzyme binding affinity K_d. Using the same hydrolase training set curated in this study (i.e., 204 data points), we retrained DLKcat and CPI models based on the code reported in their original publications, and then evaluated the predictive performance of both models on the same test set. The results show that the retrained DLKcat model exhibits a Pearson R of 0.64, a Spearman R of 0.63, and an MAE of 1.7 kcal/mol, and the CPI model displays a Pearson R of 0.63, a Spearman R of 0.65, and an MAE of 1.8 kcal/mol (Supporting information, Table S2). In comparison, EnzyKR performs better in accuracy (especially for Spearman R) than DLKcat and the CPI model in predicting activation free energies. This is likely due to EnzyKR's incorporation of the interaction map, which involves greater information density and thus helps to enhance the learning efficiency of the model.

3.4 Prediction of Enantiomeric Excess Values.

We further tested EnzyKR's ability to distinguish the reactivity difference between enantiomers. Specifically, we assessed its performance in predicting the outcomes of hydrolytic kinetic resolution. We curated a new test set that consists of 18 hydrolytic reactions catalyzed by fluoroacetate dehalogenase RPA1163[6] and halohydrin HheC[23], due to the high stereoselectivity in these two reactions (Figure 4). Under the conditions of 60°C and pH 7.0, RPA1163 selectively catalyzes the defluorination of (*S*)-2-fluoro-2-phenylacetic acid and its derivatives, while leaving the (*R*)-enantiomer untouched. On the other hand, at pH 6.5 and 35°C, Hhec catalyzes the ring-opening reaction of (*R*)-spiro-epoxyoxindoles and its derivatives, while not reacting with the *S*-enantiomer. In both reactions, the enantiomeric excess values are greater than 95%.

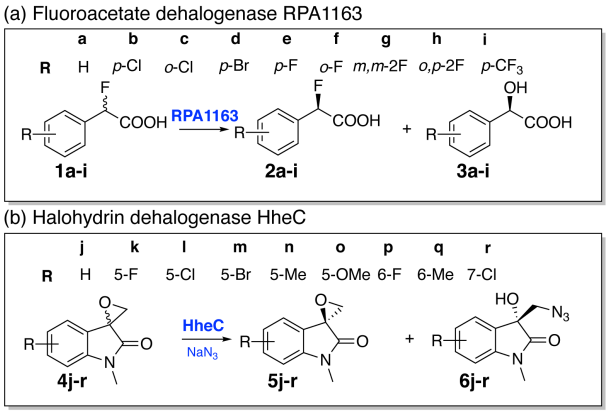

Figure 4. The test set of kinetic resolution prediction for EnzyKR. The test set was constructed by 18 enantioselective hydrolytic reactions derived from two hydrolases. (a) The fluoroacetate dehalogenase, RPA1163, catalyzes the C–F bond hydrolysis in 9 fluoroacetic acid derivatives labeled using a to i. (b) The halohydrin dehalogenase, HheC, catalyzes the stereoselective epoxide ring-opening in 9 spiro-epoxyoxindoles derivatives labeled using j to r.

To predict the *ee* value using EnzyKR, we first employed ChemDraw 22.0 to construct the SMILES strings and structural files (i.e., .sdf file) for the substrate enantiomers. Next, we employed RosettaDock to construct the hydrolase-enantiomer complexes. Taking the hydrolase-

enantiomer complex, enzyme sequence, and substrate SMILES string as input, EnzyKR predicts the $\Delta G^{\ddagger}$ values for both *R*- and *S*- enantiomers, which are denoted as $\Delta G_R^{\ddagger}$ and $\Delta G_S^{\ddagger}$, respectively. Finally, the predicted $\Delta G_R^{\ddagger}$ and $\Delta G_S^{\ddagger}$ values are plugged into eq2 to obtain *ee*%, which ranges from–100% to 100%. Notably, a positive *ee*% value indicates that the *S*-configuration is favored.

$$ee\% = \frac{1-e^{-\left(\Delta G_R^{\ddagger}-\Delta G_S^{\ddagger}\right)}}{1+e^{-\left(\Delta G_R^{\ddagger}-\Delta G_S^{\ddagger}\right)}} \qquad \text{eq. 2}$$

Figure 5 shows the *ee*% values predicted by EnzyKR(red) and DLKcat (grey), along with the reference experimental values (black). EnzyKR correctly predicts the favored enantiomer and outperforms DLKcat in 13 out of 18 reactions (i.e., 1a-e, 1g-i, 4j, 4m-n, 4p, and 4r), which occupies >70% of the test cases. For 4o, EnzyKR identifies the favored enantiomer but DLKcat performs better in quantitative accuracy. For 1f, 4k, 4i, and 4q, EnzyKR failed to identify the favored enantiomer. In more than half of the test cases, DLKcat predicts an *ee*% value lower than 50%. The overall predictive performance of DLKcat appears to be similar to a random guess. This is likely caused by the missing of chirality information in the input features. Although the SMILES string annotates chirality, chirality is not learned by the model in a physically meaningful fashion. In contrast, EnzyKR employs the atomic distance map and interaction map to differentiate substrate chirality, allowing the model to effectively learn the enantiomeric preference of hydrolases.

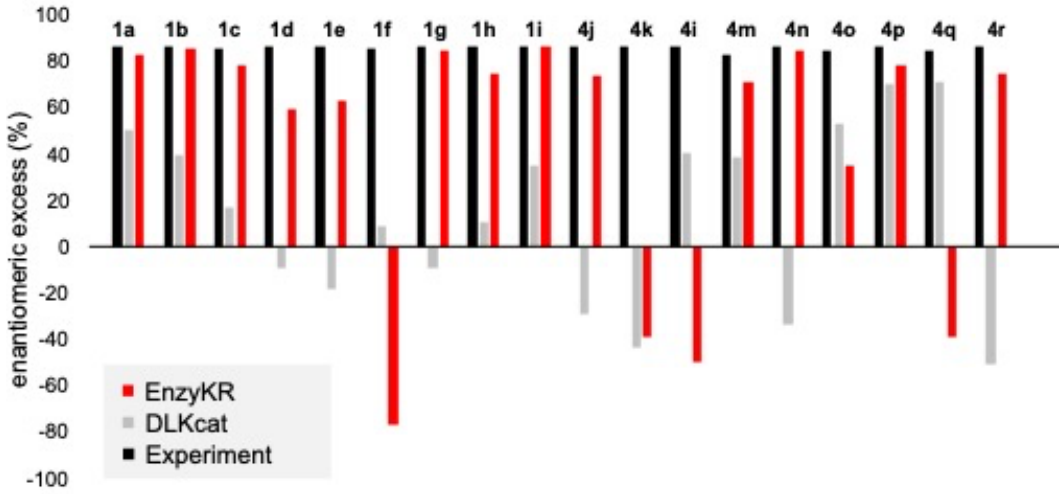

Figure 5. The predicted enantiomeric excess (*ee*%) values of EnzyKR (red) and the baseline model DLKcat (grey) for 18 enantiomer pairs in hydrolase-catalyzed kinetic resolution. The labels of the derivatives are consistent with those used in Figure 4. The reference experimental *ee*% value is shown in black.

Despite a greater predictive accuracy of EnzyKR than the baseline model DLKcat, we should note the limitation of EnzyKR in the data size and in the representation of chirality. In our future works, we plan to further improve EnzyKR in two aspects. First, we will expand the training dataset by incorporating a diverse set of substrate-enzyme complexes involved in different types of catalytic reactions, such as oxidase, reductase, and transferase, among others. This approach will allow EnzyKR to learn from a wider range of catalytic scaffolds and improve its generalizability. Second, we plan to employ equivariant neural networks (EGNN)[27] or E(n)-transformers to explicitly encode the Euclidean coordinates for the enzyme-substrate complexes. By incorporating additional geometric features, we expect that the new architecture will enable EnzyKR to better represent the chiral interactions between enzymes and substrates, leading to a further enhanced predictive performance.

4. Conclusions

Here we reported the development of EnzyKR as a deep learning model specialized in predicting the activation free energies of hydrolase-substrate complexes in a chirality-resolved manner. The model was trained on 204 data points and tested on 20 data points, where the structure and function data for hydrolase-substrate pairs have been collected from IntEnzyDB. EnzyKR comprises two components: a classifier and a regressor. The classifier is responsible for distinguishing reactive hydrolase-enantiomer complexes from unreactive binding poses, which yields a area under the curve value of 0.87. The regressor was designed to predict the hydrolytic activation free energy for the reactive complexes. On the training set, the EnzyKR regressor exhibits a strong linear correlation, with a Pearson correlation coefficient R of 0.91, a Spearman correlation coefficient of 0.86, and a mean absolute error (MAE) of 0.8 kcal/mol. On the test set, EnzyKR achieves a Pearson R of 0.66, a Spearman R of 0.70, and an MAE of 1.5 kcal/mol. Furthermore, EnzyKR was tested on a kinetic resolution task involving 18 hydrolytic reactions catalyzed by fluoroacetate dehalogenase RPA11636 and halohydrin HheC. Notably, EnzyKR accurately predicted the favored enantiomer in 13 out of 18 reactions, which significantly surpasses the performance of DLKcat, a former k_{cat} predictor model that does not embed substrate chirality. EnzyKR provides a computational tool for guiding the selection of hydrolase scaffolds for stereoselective synthesis.

ASSOCIATED CONTENT

Data and Software Availability. The raw data of IntEnzyDB can be obtained from http://intenzydb.accre.vanderbilt.edu. The source code of EnzyKR can be adopted from http://github.com/ChemBioHTP/EnzyKR.

Supporting Information. The performance of EnzyKR classifier; the method used to obtain substrate 3D structure; the method used to obtain the substrate-enzyme complexes; the benchmark results of EnzyKR features; and the comparison between the EnzyKR and other models (PDF)
The csv file of kinetics curated from IntEnzyDB; the pdb dataset of the original docked structure complexes (ZIP)

AUTHOR INFORMATION

Corresponding Author

*Email: zhongyue.yang@vanderbilt.edu phone: 615-343-9849

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This research was supported by the startup grant from Vanderbilt University. Z. J. Yang, X. Ran, Y. Jiang, and Q. Shao are supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM146982. Z. J. Yang thanks the sponsorship from Rosetta Commons Seed Grant Award and the Dean's Faculty Fellowship in the College of Arts and Science at Vanderbilt. This work used SDSC Dell Cluster with AMD Rome HDR IB at Expanse from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants BIO200057.[28]

References

(1) Reetz, M. T. Witnessing the birth of directed evolution of stereoselective enzymes as catalysts in organic chemistry. *Advanced Synthesis & Catalysis* **2022**, *364* (19), 3326-3335.

(2) Pinheiro, M. P.; Rios, N. S.; Fonseca, T. d. S.; Bezerra, F. d. A.; Rodríguez-Castellón, E.; Fernandez-Lafuente, R.; Carlos de Mattos, M.; Dos Santos, J. C.; Gonçalves, L. R. Kinetic resolution of drug intermediates catalyzed by lipase B from Candida antarctica immobilized on immobead-350. *Biotechnology Progress* **2018**, *34* (4), 878-889.

(3) Lee, J.; Oh, Y.; Choi, Y. K.; Choi, E.; Kim, K.; Park, J.; Kim, M.-J. Dynamic kinetic resolution of diarylmethanols with an activated lipoprotein lipase. *ACS Catalysis* **2015**, *5* (2), 683-689.

(4) Bassegoda, A.; Nguyen, G. S.; Schmidt, M.; Kourist, R.; Diaz, P.; Bornscheuer, U. T. Rational protein design of Paenibacillus barcinonensis esterase EstA for kinetic resolution of tertiary alcohols. *ChemCatChem* **2010**, *2* (8), 962-967.

(5) Bornscheuer, U. T.; Kazlauskas, R. J. *Hydrolases in organic synthesis: regio-and stereoselective biotransformations*; John Wiley & Sons, 2006.

(6) Zhang, H.; Tian, S.; Yue, Y.; Li, M.; Tong, W.; Xu, G.; Chen, B.; Ma, M.; Li, Y.; Wang, J.-b. Semirational design of fluoroacetate dehalogenase RPA1163 for kinetic resolution of α-fluorocarboxylic acids on a gram scale. *ACS Catalysis* **2020**, *10* (5), 3143-3151.

(7) Saini, P.; Sareen, D. An overview on the enhancement of enantioselectivity and stability of microbial epoxide hydrolases. *Molecular biotechnology* **2017**, *59*, 98-116.

(8) Qu, G.; Li, A.; Acevedo-Rocha, C. G.; Sun, Z.; Reetz, M. T. The crucial role of methodology development in directed evolution of selective enzymes. *Angewandte Chemie International Edition* **2020**, *59* (32), 13204-13231.

(9) Kazlauskas, R. J.; Weissfloch, A. N.; Rappaport, A. T.; Cuccia, L. A. A rule to predict which enantiomer of a secondary alcohol reacts faster in reactions catalyzed by cholesterol esterase, lipase from Pseudomonas cepacia, and lipase from Candida rugosa. *The Journal of Organic Chemistry* **1991**, *56* (8), 2656-2665.

(10) Tomić, S.; Kojić-Prodić, B. A quantitative model for predicting enzyme enantioselectivity: application to Burkholderia cepacia lipase and 3-(aryloxy)-1, 2-propanediol derivatives. *Journal of Molecular Graphics and Modelling* **2002**, *21* (3), 241-252.

(11) Jiang, Y.; Ran, X.; Yang, Z. J. Data-driven enzyme engineering to identify function-enhancing enzymes. *Protein Engineering, Design and Selection* **2022**, gzac009.

(12) Cadet, F.; Fontaine, N.; Li, G.; Sanchis, J.; Ng Fuk Chong, M.; Pandjaitan, R.; Vetrivel, I.; Offmann, B.; Reetz, M. T. A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. *Scientific reports* **2018**, *8* (1), 16757.

(13) Heckmann, D.; Lloyd, C. J.; Mih, N.; Ha, Y.; Zielinski, D. C.; Haiman, Z. B.; Desouki, A. A.; Lercher, M. J.; Palsson, B. O. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nature communications* **2018**, *9* (1), 5252.

(14) 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.

(15) Suzek, B. E.; Wang, Y.; Huang, H.; McGarvey, P. B.; Wu, C. H.; Consortium, U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **2015**, *31* (6), 926-932.

(16) Eddy, S. HMMER user's guide. *Department of Genetics, Washington University School of Medicine* **1992**, *2* (1), 13.
(17) 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.
(18) Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **2013**, *8*.
(19) Yan, B.; Ran, X.; Jiang, Y.; Torrence, S. K.; Yuan, L.; Shao, Q.; Yang, Z. J. Rate-Perturbing Single Amino Acid Mutation for Hydrolases: A Statistical Profiling. *The Journal of Physical Chemistry B* **2021**, *125* (38), 10682-10691.
(20) 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.
(21) DeLuca, S.; Khar, K.; Meiler, J. Fully flexible docking of medium sized ligand libraries with RosettaLigand. *PLOS one* **2015**, *10* (7), e0132508.
(22) Mendenhall, J.; Brown, B. P.; Kothiwale, S.; Meiler, J. BCL:: Conf: improved open-source knowledge-based conformation sampling using the crystallography open database. *Journal of chemical information and modeling* **2020**, *61* (1), 189-201.
(23) Zhang, F.-R.; Wan, N.-W.; Ma, J.-M.; Cui, B.-D.; Han, W.-Y.; Chen, Y.-Z. Enzymatic Kinetic Resolution of Bulky Spiro-Epoxyoxindoles via Halohydrin Dehalogenase-Catalyzed Enantio-and Regioselective Azidolysis. *ACS Catalysis* **2021**, *11* (15), 9066-9072.
(24) Adams, K.; Pattanaik, L.; Coley, C. W. Learning 3d representations of molecular chirality with invariance to bond rotations. *arXiv preprint arXiv:2110.04383* **2021**.
(25) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **2020**, *1* (4), 045024.
(26) 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.
(27) Satorras, V. G.; Hoogeboom, E.; Welling, M. E (n) equivariant graph neural networks. In *International conference on machine learning*, 2021; PMLR: pp 9323-9332.
(28) Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D.; et al. XSEDE: Accelerating Scientific Discovery. *Comput. Sci. Eng.* **2014**, *16* (5), 62-74. DOI: 10.1109/MCSE.2014.80.