Rank-ordering of known enzymes as starting points for re-engineering novel substrate activity using a convolutional neural network

Vikas Upadhyay, Veda Sheersh Boorla, Costas D. Maranas*

Department of Chemical Engineering, The Pennsylvania State University, University Park, PA, 16802

* *Correspondence to be addressed to: costas@psu.edu*

Keywords: Enzyme selection; Machine learning; Novel enzyme-substrate activity; Biosynthesis pathway design

Abstract

Retro-biosynthetic approaches have made significant advances in predicting synthesis routes of target biofuel, bio-renewable or bio-active molecules. The use of only cataloged enzymatic activities limits the discovery of new production routes. Recent retro-biosynthetic algorithms increasingly use novel conversions that require altering the substrate or cofactor specificities of existing enzymes while connecting pathways leading to a target metabolite. However, identifying and re-engineering enzymes for desired novel conversions are currently the bottlenecks in implementing such designed pathways. Herein, we present EnzRank, a convolutional neural network (CNN) based approach, to rank-order existing enzymes in terms of their suitability to undergo successful protein engineering through directed evolution or *de novo* design towards a desired specific substrate activity. We train the CNN model on 11,800 known active enzyme-substrate pairs from the BRENDA database as positive samples and data generated by scrambling these pairs as negative samples using substrate dissimilarity between an enzyme's native substrate and all other molecules present in the dataset using Tanimoto similarity score. EnzRank achieves an average recovery rate of 80.72% and 73.08% for positive and negative pairs on test data after using a 10-fold holdout method for training and cross-validation. We further developed a web-based user

interface (available at https://huggingface.co/spaces/vuu10/EnzRank) to predict enzyme-substrate activity using SMILES strings of substrates and enzyme sequence as input to allow convenient and easy-to-use access to EnzRank. In summary, this effort can aid *de novo* pathway design tools to prioritize starting enzyme re-engineering candidates for novel reactions as well as in predicting the potential secondary activity of enzymes in cell metabolism.

1. Introduction

Metabolic pathway design has enabled the synthesis of a wide range of (non-)natural chemical products, including specialty chemicals, pharmaceutical compounds, and other bio-renewables (Lin et al., 2019; Rios et al., 2021; Wang et al., 2018). Metabolic pathway assembly and construction require consideration of a range of criteria, including encoding the relevant biochemistries, biophysical understanding of enzymatic activity and specificity, understanding the thermodynamic feasibility of individual enzymatic reactions, and subsequent host selection and metabolic engineering. Several computational pathway design tools such as Path Hunter Tool (Rahman et al., 2005), MetaRoute (Blum and Kohlbacher, 2008), and optStoic (Chowdhury and Maranas, 2015) have been put forth, which automate some of the required steps in assembling pathways from a substrate(s) towards a target metabolite by querying databases of known biochemical reactions. In contrast, *de novo* pathway design tools such as BNICE (Finley et al., 2009), XTMS (Carbonell et al., 2014b), RetroPath (Carbonell et al., 2014a), Retropath2.0 (Delépine et al., 2018), and novoStoic (Kumar et al., 2018) encode reactions as molecule-agnostic rules to enable the design of pathways that include novel metabolite's reaction steps. Pathway design tools follow either (i) a graph-based description of reactions and metabolites (Wang et al., 2017), (ii) a stoichiometric matrix encoding of allowed chemistries (Kumar et al., 2018), or (iii) a retrosynthetic approach that applies generalizable reaction operators to capture possible biotransformation (Delépine et al., 2018). The advantage of retro-biosynthesis-based algorithms is that they are not limited by the chemistry and metabolite choices tabulated in reaction databases. Instead, they can instantiate novel conversions on metabolites. A novel conversion is a hypothesized reaction not cataloged or

characterized before that arises by employing a chemical reaction rule on a metabolite in a manner not seen in nature before. Novel metabolites, in turn, are hypothetical molecules that satisfy the chemical bonding rules of the novel conversion. The potential of novel conversions has already been demonstrated to produce 3-hydroxy-propionic acid (3HP) (Jessen et al., 2014) using a novel β-alanine/α-ketoglutarate aminotransferase enzyme to convert β-alanine to malonyl semialdehyde, a precursor to 3HP. In another study (Liu et al., 2021), the *de novo* production of 3-phenylpropanol was demonstrated using retro-biosynthetic techniques.

As discussed above, through the application of reaction rules, retrosynthetic algorithms can craft a multitude of possible *de novo* conversions that usually require the incorporation of enzymes that need to operate on novel metabolites (either acting on non-native substrates or accepting a non-native cofactor). However, the selection of enzyme candidates out of all known enzymes that can catalyze the needed metabolic transformation on non-native substrates poses a crucial design challenge in metabolic engineering and pathway design with novel steps (Fig. 1 pictorially illustrates this challenge). SimZyme (Pertusi et al., 2015) is a tool that rank-orders enzymes based on the substrate similarity of a given substrate to native substrates of the enzymes. Such tools can offer reasonable starting points for enzyme discovery for missing pathway steps (Carbonell et al., 2018; Pertusi et al., 2015). However, this approach relies upon the assumption that enzymes would structurally accommodate and catalytically engage with novel substrates based solely on the similarity in shape and size to their native substrates. However, enzyme-substrate reactivity depends on both substrate similarity and enzyme plasticity to accept and perform the same chemistry on different substrates. For example, the acetyl-CoA hydrolase Ach1p enzyme from *Saccharomyces cerevisiae* (Fleck and Brock, 2009) exhibits a high specificity towards succinyl-CoA. However, it shows no activity towards the chemically and size similar substrate, acetyl-CoA.

On the contrary, phosphatase enzymes from various thermophiles overexpressed in *Escherichia coli* show activity on three chemically distinct substrates glucose-6-phosphate (G6P), fructose-6-

74 phosphate (F6P), and mannose-6-phosphate (M6P). The study also shows that the phosphatase derived
75 from different thermophiles exhibits varying substrate specificities towards the above three metabolites,
76 preferring one over the other in different organisms—the phosphatase derived from *Thermotoga sp. 38H*,
77 *Thermoclostridium stercorarium*, and *Petrotoga miotherma* show the highest proportion of mannose,
78 glucose, and fructose from M6P, G6P, and F6P, respectively (Tian et al., 2022). Therefore, enzyme
79 variants from different organisms can often have widely different substrate preferences and degrees of
80 promiscuity, reinforcing that it is not only substrate similarity but also enzyme plasticity that needs to be

captured to identify promising enzyme candidates for novel conversions. A number of different factors have been used on a case-by-case basis to select enzymes based on (i) substrate similarity, (ii) enzyme family promiscuity, and (iii) accommodation of new substrate in the active site (Carbonell et al., 2018, 2014a; Kumar et al., 2018). However, no systematic and broadly applicable method exists for enzyme selection that simultaneously leverages these multiple criteria. It is important to note that even though numerous enzyme selection strategies use algorithms (Feehan et al., 2021; Goldman et al., 2022) that

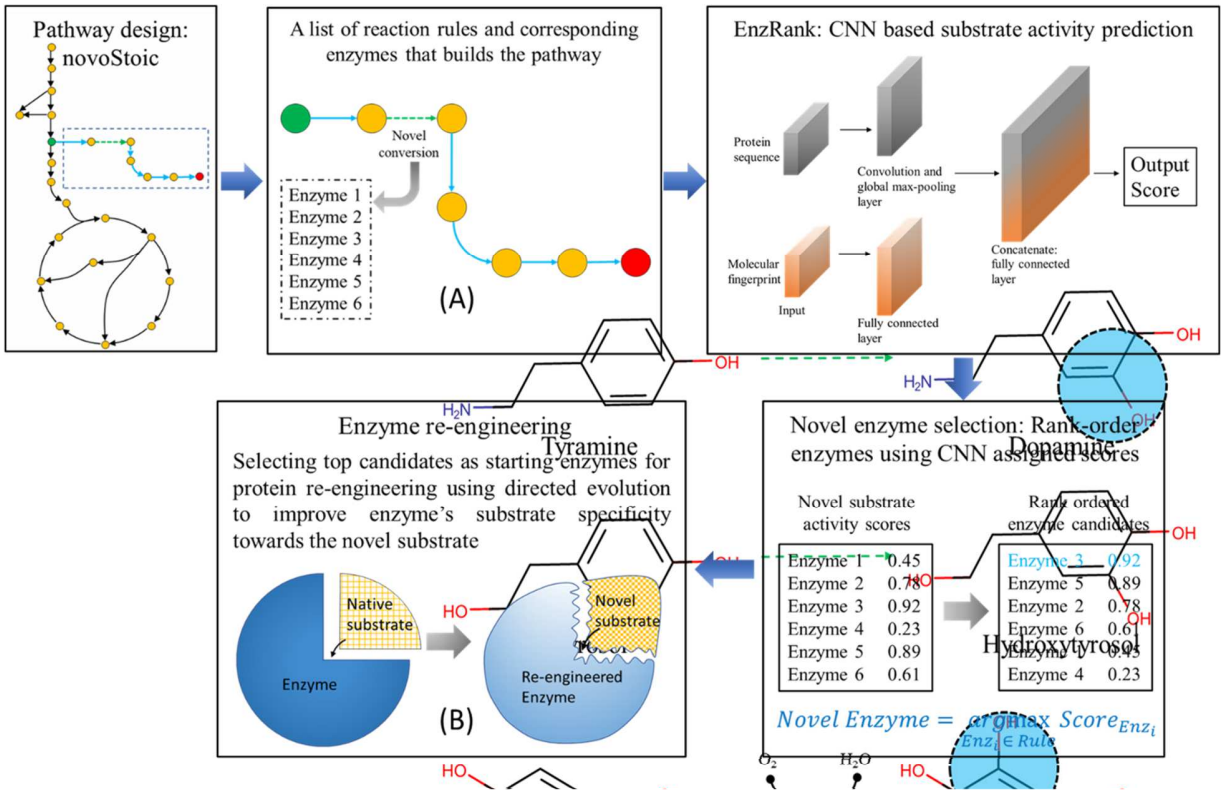


Fig. 1. Overall workflow for de novo pathway design. Retrosynthesis tools can assemble pathways with novel steps. EnzRank calculates a probability score for an enzyme to have activity on a substrate. This rank-order the enzymes with the same reaction rule can be carried out using the EnzRank scores. High scoring enzymes can be subsequently re-engineered to increase substrate specificity towards the novel substrate.

predict the EC classification of a given enzyme sequence to infer whether the enzyme can catalyze a novel substrate, this level of detail is not sufficient for our goal as all candidate enzymes would presumably be classified with the same EC number.

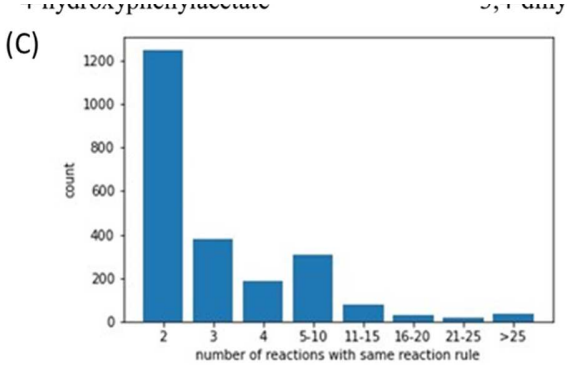


Fig. 2. Novel reactions using reaction rules, (A) Novel promiscuous hydroxylase enzyme that was engineered to work on two different substrates, tyramine and tyrosol for the synthesis of an antioxidant drug hydroxytyrosol. (B) The native reaction of the flavin-dependent monooxygenase HpaBC enzyme that acts on 4-hydroxyphenylacetate undergoes an exact metabolic transformation as the novel reactions, hence the same reaction rule (shown in the blue circle). (C) number of reactions with the same reaction rules. There are a significant number of reactions within the same rule which demonstrates the need to design an enzyme selection tool

Often the desired activity on a novel substrate is present, albeit the activity level is low (Wen et al., 2008). Protein engineering through directed evolution (DE) is a

popular strategy for ratcheting up the desired activity with the novel substrates while at the same time abolishing activity with the original substrate (Hibbert and Dalby, 2005; Porter et al., 2015; Taylor et al., 2015). For example, among many other examples, DE has been demonstrated to produce steviol glucosides using cytochrome p450 (CYPs) enzymes kaurene oxidase (KO) and kaurenoic acid hydroxylase (KAH) in yeast (Gold et al., 2018). However, often no obvious choice exists for the starting enzyme to undergo DE. For example, Fig. 2A shows the novel conversion of tyramine and tyrosol as substrates with a promiscuous hydroxylase enzyme (Chen et al., 2019). These novel conversions are inferred by applying the reaction rule implied by the known enzymatic function of a flavin-dependent monooxygenase HpaBC enzyme on 4-hydroxyphenylacetate (shown in Fig. 2B). The KEGG database contains as many as 967 distinct sequences of HpaBC that belong to multiple organisms making it non-trivial to select a single (or even a handful) of suitable enzymes which may be the best starting points for DE. In addition, depending on the specificity of the employed reaction rule, there are often undesired reactions with the same reaction rule operating on different substrates, which should not be excluded from consideration (see Fig. 2C). Therefore, there is a need for a computational tool to select and pick top enzyme candidates that are most likely to either have or are likely to acquire upon engineering activity on a given novel substrate. Herein we introduce EnzRank, a computational method that relies on training a CNN model which uses enzyme sequences and their respective substrate's molecular signature as inputs to assign a probability score that the enzyme sequence has activity on the substrate. The EnzRank calculated probability score of the enzyme-substrate pairs is used to rank-order all candidate enzymes by their potential to exhibit activity for the novel substrate.

EnzRank is inspired by the recent advances of machine learning-based approaches in predicting enzyme classes, properties, and functions with tools such as alphafold2 (Jumper et al., 2021), DeepEC (Ryu et al., 2019), ECpred (Dalkiran et al., 2018), and SDN2GO (Cai et al., 2020) (Table 1 shows a list of such tools for enzymes). Machine learning methods that use enzyme sequences as input incorporate a variety of feature encoding techniques such as UniRep (Alley et al., 2019), SeqVec (Heinzinger et al.,

2019), ESM-1b (Rives et al., 2021), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019), etc. We opted for a learnable embedding technique to learn problem-specific features for enzyme sequences along with a convolutional neural network DeepConv-DTI (I et al., 2019). EnzRank encodes the enzyme's amino acid sequence as trainable embeddings to learn mappings that connect to the respective enzyme-substrate activity.

To encode substrate structural information, we used a molecular signature-based encoding, morgan fingerprints, which uses the presence of small substructures to define a molecule and are widely used for chemical similarity (Capecchi et al., 2020). This architecture is largely inspired by DeepConv-DTI (I et al., 2019), which was successful in drug-interaction prediction, similar to the problem at hand – enzyme-substrate activity prediction. EnzRank achieves an average recovery of 80.72% for true positives and 73.08% for true negatives on blinded test data splits. EnzRank's performance on an unseen dataset was assessed by a small dataset of P450 monooxygenase enzymes achieving the recovery for positive and negative activity pairs are 75% and 95%, respectively.

Table 1. Recent machine learning-based tools for enzyme functions predictions

ML tools for enzymes	**Description**	**Method**	**Input required**	**Reference**
DeepEC	Enzyme Commission (EC) number prediction	CNN	Protein Sequence	Ryu et al., 2019
ECpred	EC number prediction	Ensemble method (SVM and kNN)	Protein Sequence	Dalkiran et al., 2018
mlDEEPre	EC number prediction	Ensemble method (CNN and RNN)	Protein Sequence	Zou et al., 2019
Proteinfer	EC number and Gene ontology prediction	CNN	Protein Sequence	Sanderson et al., 2021
SDN2GO	Protein function prediction	CNN	Protein sequence, protein domain content, and known protein-protein interactions	Cai et al., 2020

DeepFRI	Protein function prediction	Graph convolution network (GCN)	Protein sequence and structure	Gligorijević et al., 2021
SimZyme	Enzyme selection for pathway design	Substrate similarity of enzyme's native and the novel substrate	Molecular fingerprint of substrates	Pertusi et al., 2017
EnzRank	Enzyme selection for novel reactions in de novo pathway design	CNN	Protein sequence and substrate structure	**This study**

We demonstrate the application of EnzRank to rank-order enzyme variants by their likelihood to exhibit activity upon the corresponding substrate. We applied it to rank-order the enzymes for novel conversions in the pathways identified by novoStoic (Kumar et al., 2018) for 3-HP synthesis. This computational pre-processing step can potentially speed up the process of assembling *de novo* pathways.

2. Results

2.1. Performance comparison of Similarity-based and CNN models on the validation dataset

We explored two different models for enzyme-substrate activity prediction, (i) SimProd: a substrate similarity and protein sequence identity-based model, and (ii) EnzRank: a Convolutional neural network. SimProd uses Tanimoto-based substrate similarity and sequence alignment with all other protein sequences. It formalizes existing intuitive strategies for enzyme selection by relying on substrate similarity and protein sequence identity. SimProd serves as a performance floor to assess improvements afforded by the more sophisticated CNN model. Briefly, in SimProd, for a given enzyme-substrate pair, the substrate similarity and sequence alignment scores with all the known enzyme-substrate pairs in the database are multiplied. The combination with the maximum score is assigned as the final score to the examined enzyme-substrate pair. A threshold parameter was used to classify the corresponding enzyme-substrate pairs as active based on the final scores. The SimProd model devises an optimal threshold by

minimizing the mean absolute error between the predicted and assigned label for the activity (1: active and 0: inactive) over the entire training data (detailed method in Supplementary data).

In contrast, the CNN-based model uses both enzyme sequence and substrate molecular signatures as inputs to a convolutional neural network in a feed-forward fashion to output a probability score for the enzyme-substrate activity (details in the Methods section). The model uses a representation learning approach to encode features from the enzyme sequence. The enzyme sequence first passes through an embedding layer to assign a real-valued vector to each amino acid. The output from the embedding layer goes through a series of 1-D convolutional layers. The molecular fingerprints of the substrate are binary vectors of length 2,048 which are input to a full-connected layer parallelly and independently to the enzyme sequence. Finally, the enzyme and substrate outputs are combined in a fully connected layer to output a single real-valued probability between 0 and 1. The training data is passed through the model, and the binary cross-entropy loss between true labels and predicted loss is set as the training objective function. We used an 80:10:10 split of the entire dataset parsed from the BRENDA database to generate training, validation, and test datasets and perform a 10-fold cross-validation training of the CNN model. The cross-validation training yielded ten different models. We used the average recovery percentage of positive and negative predictions across the ten models as a metric to assess model performance. The positive recovery is defined as the percentage of known active enzyme-substrate pairs correctly predicted as active, also called true positive (TP). The negative recovery is the percentage of correctly predicted negative pairs in the data, true negative (TN).

While both positive and negative recovery rates for both SimProd and EnzRank are comparable (shown in Fig. 3), EnzRank dramatically outperforms SimProd over the validation dataset (especially for positive pair recovery). The average

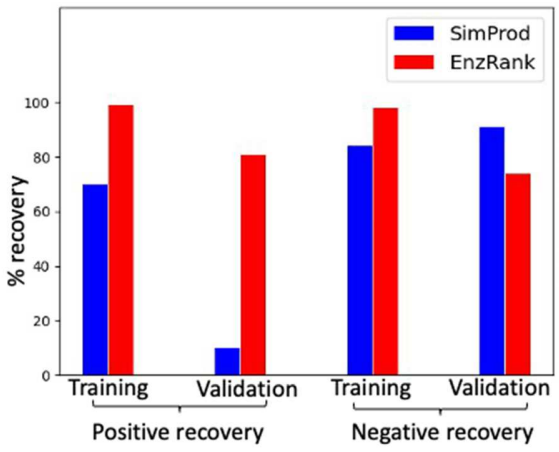


Fig. 3. The comparison of performance accuracy as % recovery of two different models (SimProd and EnzRank) tested for activity predictions. The positive and negative recovery shows the percentage of correctly predicted positive and negative pairs in the training and validation data

recovery on validation data splits for the positive datasets are 80.74% and ~10% for EnzRank and SimProd, respectively. Whereas the negative recovery on validation dataset is 73.74% and 91% for EnzRank and SimProd, respectively. This implies that, unlike SimProd, which is susceptible to overfitting, EnzRank can genuinely "learn" from patterns seen in active substrate-enzyme pairs in the dataset and correctly recognize them in the validation dataset. Note that the negative recovery for SimProd is higher than the EnzRank because the negative dataset itself was generated using substrate dissimilarity of the enzyme's native substrate and all other substrate molecules in the dataset using the Tanimoto similarity score. Therefore, SimProd is biased toward negative recovery as it uses the similarity between substrates and enzyme sequences to predict activity, thereby giving high accuracy toward the negative pairs. As our goal is to predict positive activity for novel enzyme-substrate pairs, the SimProd model will not be a viable choice.

Next, we performed hyperparameter optimization of the CNN model parameters by varying the dropout probability and the number of epochs (Supplementary data). The dropout probability is an important parameter to optimize as it can prevent model overfitting (Srivastava et al., 2014). The number of epochs is the number of times the entire training data is passed through the learning algorithm. It is associated with the generalization capacity of the neural networks and controls overfitting (Perin et al., 2021). The grid search showed an optimal performance at a dropout probability of 0.3 and 70 epochs. The CNN model recovered 98.96% and 98.11% positive and negative pairs over the training dataset and 80.74% and 73.74% over the validation dataset with the optimized parameters (see Table 2 for details).

Table 2. Performance metrics of EnzRank CNN model on ten different splits of the training, validation, and test datasets using optimized hyperparameters

Data Split # ↓	Training (% recovery)		Validation (% recovery)		Test (% recovery)	
	Positive pairs	**Negative pairs**	**Positive pairs**	**Negative pairs**	**Positive pairs**	**Negative pairs**
1	98.86	97.42	81.17	72.70	81.44	73.24
2	99.52	98.96	81.44	77.11	79.73	77.12

23	99.78	97.97	87.20	68.64	87.48	67.29
4	99.52	98.41	80.99	73.06	82.16	73.87
5	99.69	97.04	83.60	71.62	84.59	70.36
6	98.51	98.23	80.36	73.78	78.19	74.77
7	99.48	98.34	81.80	73.15	82.52	71.26
8	98.82	98.50	80.36	74.68	78.38	72.34
9	98.97	97.46	79.81	74.77	80.45	72.16
10	96.45	98.75	70.72	77.93	72.25	78.38
Average score	**98.96**	**98.11**	**80.74**	**73.74**	**80.72**	**73.08**

Moreover, it is always a good practice to test machine learning models' performance on a blinded dataset that has a different origin from the training dataset and has not been a part of training. We manually curated a dataset of cytochrome P450 monooxygenase enzymes (shown in Supplementary data), including 36 enzyme-substrate pairs with 16 active pairs (positive) and 20 inactive pairs (negative). The negative pairs were generated using small molecules (usually show activity on long-chain fatty acid with >9C (Nakayama et al., 1996)) for which the cytochrome P450 enzymes are known not to show any activity (Kitazume et al., 2002; Nakayama et al., 1996). EnzRank (CNN model) achieved strong positive (75%) and negative (95%) recovery rates comparable to validation performance. It correctly identified the lack of activity of cytochrome P450 enzymes for almost all decoy enzyme-substrate pairs (recovery 95%). EnzRank captured enzyme-substrate activity on a smaller blind dataset with a comparable recovery compared to the split dataset used for validation and testing.

2.2. Challenging EnzRank with a negative dataset composed of similar substrates to the native ones

We next assessed the recovery of native from very similar substrates by EnzRank as opposed to simply randomizing enzyme to substrate assignments in the negative dataset. To this end, we generated new test

datasets using the previous 10 random test data splits. The previous datasets contained an equal proportion of positive and negative data with 2,220 enzyme-substrate pairs (1,110 positive and 1,110 negative pairs). To generate a more "challenging" negative dataset, we paired the enzyme sequences of the positive data with the top ~10% of most similar compounds with respect to their native substrates using the Tanimoto similarity index, resulting in an average of ~200 negative data points for each protein sequence. We repeated this process for all 10 splits, resulting in a test dataset of length ~222,000 in each split.

Upon training, we found that positive pairs (i.e., native substrates) exhibited significantly higher scores compared to the near-native negative pairs, as shown in Figure 4. Supplementary data includes the same plot for all ten splits of the test data. the statistical significance of the score differences between positive and negative pairs was assessed using the Wilcoxon rank-sum test accessed through the SciPy Python package. The null hypothesis posited that the two sets of scores (positive and negative data) were drawn from the same distribution, while the alternative hypothesis ascertains that the positive pairs scores were more likely to be greater than those from negative pairs. The analysis yielded a p-value of 2.37e-22, which rejects the null hypothesis and demonstrates that the positive dataset exhibiting higher scores over the negative dataset is statistically meaningful. This proves that EnzRank can discriminate between positive and negative pairs generated by matching known positive enzyme sequence pairs with near-native substrates. Note that it is entirely possible that many of the pairs in the "negative" dataset may be in fact functional. Therefore, we observed an average recovery rate of 80.72% for positive pairs (same as shown in Table 2) but also a non-trivial recovery of 55.72% for negative pairs. This may allude to the fact that (some) of the near-native

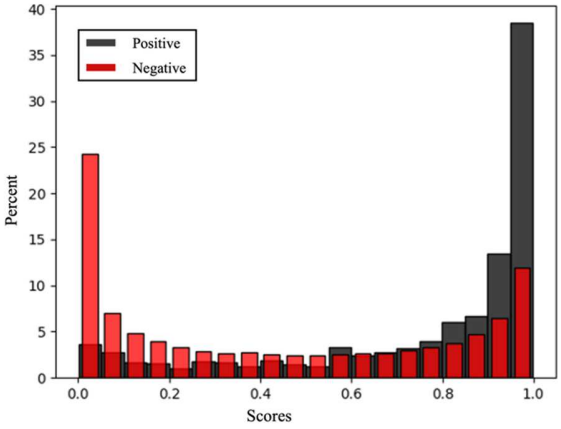


Fig. 4. Comparison of positive and negative pairs performance on one of the split test datasets generated using similar substrates. Red and black bars show distributions of EnzRank predicted scores for negative and positive enzyme-substrate pairs, respectively. Y-axis is normalized such that the sum of bars for each category sum to 100%.

substrates have activity on the paired enzymes. Note that prior models (Carbonell et al., 2018; Moriya et al., 2016) solely relied on substrate similarity and consistently ranked the most similar substrate as the top candidate. In contrast, EnzRank leverages both substrate information and local residue patterns within protein sequences to account for both substrate and sequence characteristics, surpassing the use of substrate similarity alone.

2.3. Selection of enzymes for novel reactions in pathways found for 3-Hydroxypropionic acid

We demonstrate the utility of EnzRank for pathway design by selecting prototype enzyme candidates for conversions in the novel pathways identified by novoStoic (Kumar et al., 2018) for 3-hydroxypropionic acid (3-HP) synthesis. Several biosynthetic pathways for 3-HP synthesis have already been identified using glucose, glycerol, xylose, and malonic acid as starting points (Liang et al., 2022). Here, we used pyruvate as the precursor, as glucose to pyruvate conversion pathways are well established. Both KEGG reaction database (Kanehisa et al., 2016) entries and reaction rules implied by the cataloged reactions were used. We first assessed the best overall conversion stoichiometry satisfying overall thermodynamic feasibility:

$$1\ C_3H_4O_3 + 1\ NADH + 1\ H^+ \rightarrow 1\ C_3H_6O_3 + 1\ NAD^+ \qquad \Delta_r G'^{\circ} = -32.4\ kJ/mol$$

Using KEGG ids we have:

$$C00022\ (pyruvate) + C00004\ (NADH) + H^+ \rightarrow C01013\ (3\text{-}HP) + C00003\ (NAD^+)$$

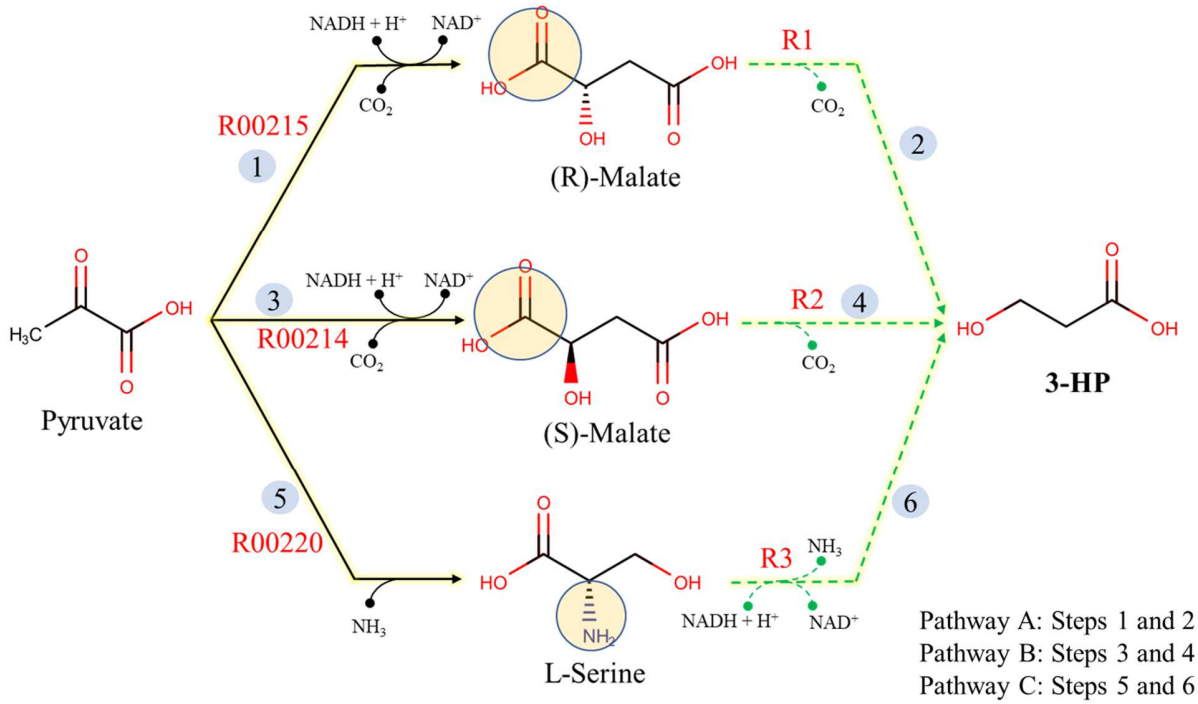
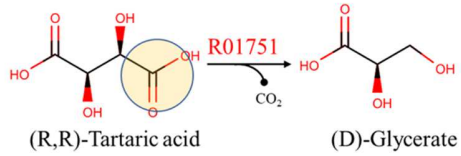
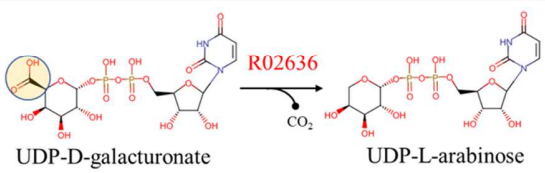
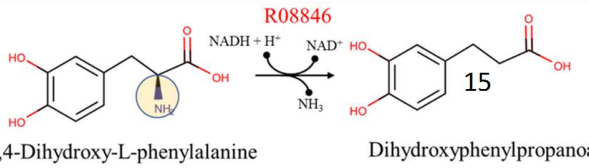


Fig. 5. Three novel pathways found for 3-HP biosynthesis; The first identified pathway A includes reaction steps 1 and 2, where pyruvate is first converted to (R)-Malate via (R)-Malate oxidoreductase in step 1, and step 2 involves a novel conversion derived from reaction rule R1 (carboxy-lyase enzyme shown in Table 3) to convert (R)-Malate to 3-HP. Pathway B consists of reaction steps 3 and 4; step 3 converts pyruvate to (S)-Malate using the (S)-Malate oxidoreductase enzyme, which then uses reaction rule R2 (carboxy-lyase enzyme, Table 3) based novel conversion to produce 3-HP as step 4. Pathway C involves steps 5 and 6, where step 5 converts pyruvate to L-serine and uses reaction rule R3, derived from ammonia-lyase enzyme based on KEGG reaction "R08846" (Table 3).

The negative standard Gibbs energy change of the overall conversion ensures that the overall stoichiometry is thermodynamically feasible under standard conditions. The next step was the identification of the intervening chemical conversion steps that conforms to the identified overall

Table 3. Reaction rules used in novel conversion for 3HP synthesis

Rule id	Reaction rule	Enzyme/homolog name	Step
R1		(R, R)-tartrate carboxy-lyase	2
R2		UDP-D-galacturonate carboxy-lyase	4
R3		3,4-dihydroxy-L-phenylalanine ammonia-lyase	6

stoichiometry using the 3,603 unique reaction rules derived from 7,053 biochemical reactions in the KEGG database. Among many identified, three pathways (see Fig. 5) were unexplored before. Existing biosynthetic pathways for 3-HP involve four or more reaction steps starting from pyruvate (Liang et al., 2022). All three pathways found by novoStoic involve only two steps by leveraging designed novel conversions. Next, we used dGPredictor (Wang et al., 2021), an automated standard Gibbs energy prediction tool, to ensure the thermodynamic feasibility of individual steps in the identified pathways. dGPredictor allows the estimation of standard Gibbs energy change for reactions, including reactions with hypothetical metabolites. The predicted $\Delta_r G'^{o}$ for reaction step 1, which is KEGG reaction R00215 is -24.15 ± 6.62 *kJ/mol,* and for the novel reaction step R1 is −11.67 ± 7.33 *kJ/mol* consistent with thermodynamic feasibility for the two reaction steps under standard conditions.

Table 4. Top 10 genes from different organism with EnzRank score > 0.995 for rule R1

Organisms
B. Composti
S. humosa
U. limnaea
B. bacterium GR16-43
A. sediminis
H. halodurans
S. rapamycinicus
C. thermophila
S. usitatus
E. billingiae

Organisms
U. croceus
C. manihotivorum
A. indistinctus
C. amurskyense
A. soli
A. finegoldii

Next, we applied EnzRank to rank order enzymes for novel reaction steps R1. We found two reactions in the KEGG database conforming to the same reaction rule as R1 (i.e., KEGG reaction R01751 (EC 4.1.1.73) and R07125 (EC 4.1.1.85)). Upon parsing all the genes for those reactions from different organisms using the KEGG database, we found 2,819 homologous sequences. We used EnzRank to rank-order them in terms of their predicted activity fitness for the novel substrate. The top ten candidates with EnzRank scores of more than 0.995 for reaction rule R1 are shown in Table 4

Similarly, for pathway B, the thermodynamic analysis using dGPredictor estimates a $\Delta_r G'^o$ of -13.79 ± 4.78 *kJ/mol* for reaction step 3 (R00214) and -22.03 ± 5.76 *kJ/mol* for step 4, (i.e., rule R2). Reactions R02636 (EC 4.1.1.67) and R01384 (EC 4.1.1.35) conform to reaction rule R2. As many as 2,994 sequences that code for homologous enzymes was found. Table 5 tabulates the top ten scoring candidates. Finally, for pathway C, the thermodynamic analysis using dGPredictor revealed that steps 5 (i.e., KEGG R00220) and 6 (i.e., rule R3) have standard free energy of change $\Delta_r G'^o$ of 29.83 ± 3.43 *kJ/mol* and -65.64 ± 4.72 *kJ/mol*, respectively. Even though the overall $\Delta_r G'^o$ for the pathway is negative, step 5 seems to be thermodynamically unfavorable; therefore, pathway C does not seem to be a viable option. In summary, we demonstrated how EnzRank could be integrated with an overall retrosynthetic workflow to rank-order enzyme sequences as appropriate candidates for the novel conversion(s).

Rhizobium sp. N731
A. sanaruensis
M. preniciosa
S. halorespirans DSM 13726

Table 5. Top 10 genes from different organisms with EnzRank score > 0.99 for rule R2

2.4. Class Activation Maps to identify protein residues positively influencing EnzRank scores

There are various approaches to interpret predictions made by a convolutional neural network in terms of which parts of input features are most important for arriving at the predictions (Selvaraju et al., 2016; Zhou et al., 2015). Such methods are widely used in computer vision, specifically in image classification, to find regions of images that lead to a positive class prediction. Here we leverage one such approach, namely, the gradient-weighted Class Activation Maps (grad-CAMs) (Selvaraju et al., 2016). Grad-CAM uses the gradient of the output score with respect to each convolutional feature map to estimate the contribution of individual residues to the final score (see details in the method section "Class Activation Map"). After calculating the grad-CAM scores, we visualize them by mapping onto their corresponding three-dimensional structures by AlphaFold-2.0 (Jumper et al., 2021). To assess the grad-CAM scores, we applied the method on enzyme-substrate complex structures available in the Protein data bank (Berman et al., 2003). The results are summarized in Fig. 6. We found that residues involved in binding exhibit

higher on average grad-CAM scores implying that the CNN model can learn structural features even though the training data did not explicitly encode any. This has been seen in other studies where training purely on sequence features identified structurally important features (Elnaggar et al., 2021; Rives et al., 2021). Not all residues with high grad-CAM scores are near the binding site pocket. Possibly other factors important in catalysis, such as interaction with partner subunits, dynamic movement of different enzyme parts, and interaction with allosteric sites could be at play. Assessing these other factors is beyond the scope of this effort.

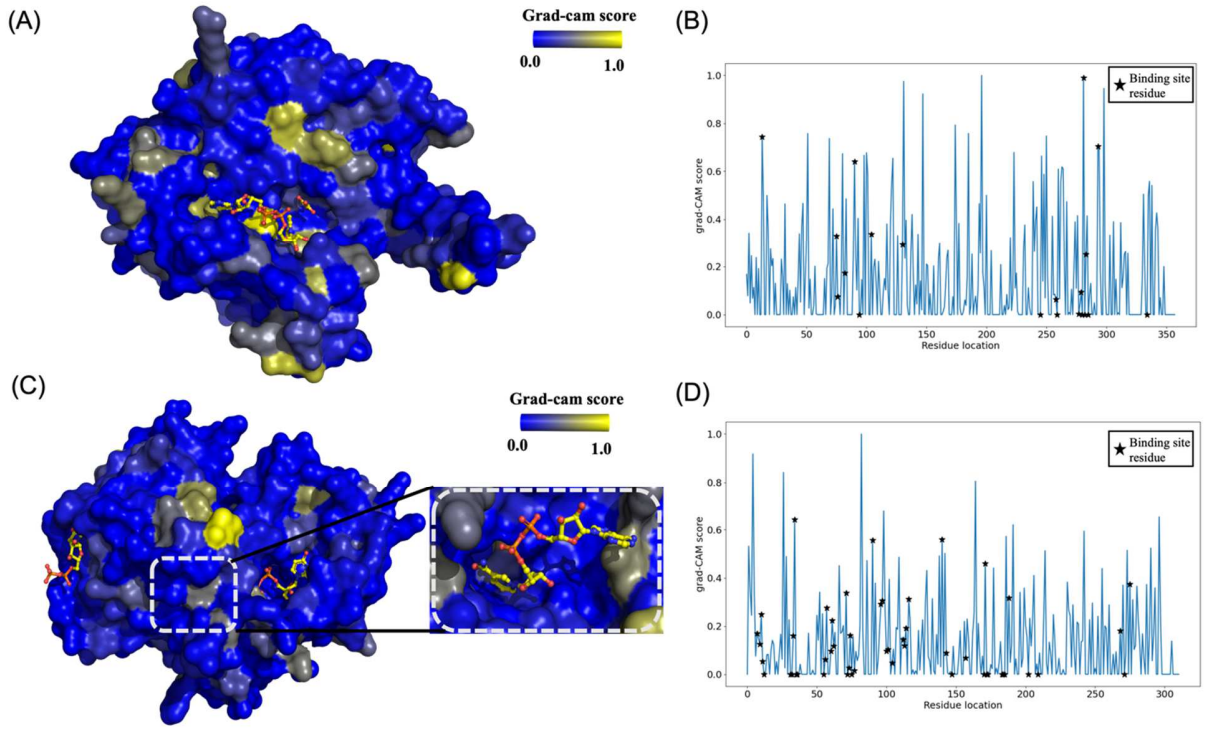


Fig. 6. Grad-CAM results for two top enzyme candidates in reaction rules R1 and R2. (A) and (C) show three-dimensional structures of the enzymes in surface representation with the homology inferred predicted substrate interaction shown in stick representation. The surface of the enzyme is color-coded based on the grad-CAM scores ranging from yellow (highest score) to gray to blue (least score). (B) and (D) show respective grad-CAM score line plots for enzymes shown in (A) and (C), respectively (stars in line plots show the binding site residues). (A) and (B) Decarboxylase enzyme in *B. composti* (Uniprot id: A0A7T5EPK7) for reaction rule R1 and corresponding grad-CAM line plot. (C) and (D) Decarboxylase enzyme in *U. croceus* (Uniprot id: A0A1D8PAW7) for reaction rule R2 and corresponding grad-CAM line plot.

2.5.User-friendly interface for predicting activity probability score using EnzRank

EnzRank input format uses the enzymes' entire protein sequence and SMILES string of the substrate as input. For example, EnzRank recognizes the phenylpyruvate tautomerase enzyme "P14174:

MPMFIVNT…" as a protein sequence and the SMILES string of substrate as "3-phenylpyruvate: [O-]C(=O)C(=O)Cc1ccccc1" as input (shown in Fig. 7). Another benefit of using SMILES string as input to EnzRank instead of chemical IDs from known databases is that it allows the inclusion of novel substrates that are not cataloged in any biochemical databases. A user-friendly interface is developed to facilitate easier access to EnzRank for rank-ordering starting enzyme candidates for novel reactions in *de novo* chemical synthesis pathways. This allows users to input multiple enzyme sequences at once and rank order EnzRank predicted activity scores with the desired novel substrate. EnzRank can also be used as a pre-processing tool to reduce the sample size of enzymes for re-engineering novel activity. We envision that the developed GUI will facilitate easy adoption of EnzRank to the broader metabolic engineering and synthetic biology community, which relies on de novo synthesis tools and enzyme re-engineering tools to improve/find novel efficient routes for biochemical synthesis.

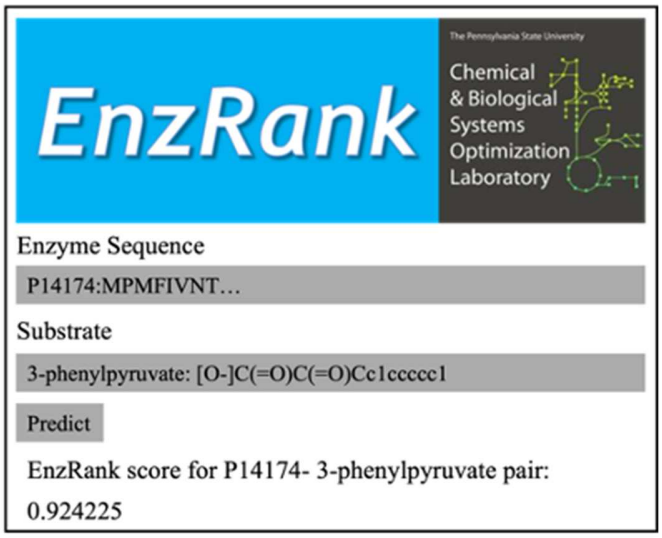


Fig. 7. A web-based graphical user interface for easier access to the EnzRank tool. The interface requires the protein sequence of the enzyme and SMILES string of the substrate as input. Next, clicking the search button outputs the scores that can be used to to rank-order enzymes for selecting a starting point for protein re-engineering.

3. Discussion

In this work, we developed a CNN model for enzyme activity prediction using only the substrates' molecular fingerprints and the enzyme sequence as inputs. We trained and validated the model using a dataset of known enzyme-substrate activities curated from the BRENDA database. We found that the CNN model (EnzRank) performs better than an enzyme/substrate similarity-based model (SimProd). The better performance of the CNN model can be attributed to the use of convolution layers that can

automatically detect local features of the enzyme sequence responsible for activity prediction. Although the CNN-based model performed well across training-validation datasets and also on a blind dataset, some challenges remain, such as the lack of a true negative dataset (experimentally validated) for training the model. Here, we used a synthetic dataset formed by substrates dissimilar to the native substrate of an enzyme as the negative data. While it is unlikely that an enzyme can be active on substrates highly dissimilar to native substrates, using such a strict negative dataset would prevent the model from learning about inactive substrates that are only marginally dissimilar to native substrates. However, one could keep enriching the training dataset by including substrates that have been experimentally tested to show no activity and progressively increase the model's fidelity. Recent developments in natural language processing (NLP) based literature mining tools (Cheng et al., 2008; Hur et al., 2009; Simon et al., 2019) could potentially be leveraged to help generate better negative datasets to aid in training future machine learning models. Various articles referring to EC class 1.1.1.1 in the BRENDA database contain limited information on the zero activity of enzyme-substrate pairs. We performed a literature survey to find substrates for EC class 1.1.1.1 that show no activity toward the enzyme (Supplementary data) to show the potential of leveraging NLP-based literature mining tools to automate the generation of the experimentally validated negative dataset as manual curation for each EC class might not be feasible.

In its current form, we anticipate that EnzRank can assist *de novo* pathway design tools such as novoStoic in selecting the starting enzyme for novel substrate activity in the identified *de novo* pathways. The current version of novoStoic uses substrate similarity between the desired substrate and the native substrate of the enzyme to rank the enzymes that can perform the exact chemical transformation in a reaction. EnzRank could uncover better enzyme targets than just using substrate similarity by accounting for both enzyme sequence information and substrate information. Along with that, using grad-CAMs also provide insights on specific residue level information that influences most toward the predicted EnzRank score for enzyme-substrate activity.

There is still scope for improvement over EnzRank rankings by including more descriptors of protein features. For example, several machine learning-based feature extraction methods (Cai et al., 2020; Devlin et al., 2019; Elnaggar et al., 2021; Liu et al., 2019; Rives et al., 2021) (such as UniRep, ProtBERT, SeqVec, ESM-1b, etc.) can be utilized along with protein sequence to build a CNN model that might further improve the performance of the activity prediction model. These feature extraction methods are pre-trained on millions of protein sequences to learn essential features from the protein sequences. EnzRank is, to our knowledge, the first computational tool to assist in rank-ordering starting enzymes to undergo directed evolution toward a new substrate. EnzRank can be used within any de novo pathway design tool that uses reaction rules to build retro-biosynthesis pathways to select the enzymes for novel reactions, as current practices only use reaction rules for novel reactions but there exist multiple enzymes with the same reaction rule that possess the challenge to pick a few candidates for protein re-engineering to alter substrate/cofactor specificity. EnzRank could also help complete organism-specific metabolic models by pinpointing possible secondary enzymatic activities of the known enzymes.

4. Methods

4.1. Dataset

Data on enzyme-substrate activities were obtained from the BRENDA database (Schomburg et al., 2002) alongside the PDB ID for the protein sequences and the common chemical names of the active substrates. The protein sequences were downloaded from UniProt (Bateman et al., 2021). We compiled a list of all substrates across all enzymes. We then queried the PubChem database (Kim et al., 2021) and the Open Parser for Systematic IUPAC nomenclature (OPSIN) database (Lowe et al., 2011) to establish uniform identities. OPSIN allows the identification of the IUPAC names (Mc Naught and Wilkinson, 2012) from the common names of the substrates obtained from the BRENDA database. We then used these IUPAC names to retrieve the simplified molecular-input line-entry system (SMILES) strings (Lunnon et al., 1988) for the substrates from the PubChem database (Kim et al., 2021). Next, the RDkit python

410 package(Landrum, 2006) (https://www.rdkit.org/) was used to generate the morgan fingerprints
411 (Landrum, 2006) of the substrates. The morgan fingerprints, also known as extended connectivity
412 fingerprints (ECFPs), are the molecular representation based on the topology of the chemical structure
413 within a specific distance (Rogers and Hahn, 2010). We used the Morgan fingerprint (Rogers and Hahn,
414 2010) to encode substrate molecules as a graph feature. Using RDkit, the molecular fingerprints of
415 substrates were generated for radius 2 utilizing the SMILES string of the molecules. In the end, each
416 substrate can be represented as a binary vector of 2,048 lengths, whose indices indicate the presence of a
417 specific chemical moiety within the molecule. A total of 3,500 enzymes sequence and 10,353 compounds
418 were parsed, resulting in 11,080 known enzyme-substrate activities. The lack of availability for negative
419 enzyme-substrate activity data in BRENDA leads us to use random enzyme-substrate pairs that are not
420 present in the parsed positive dataset. Therefore, the negative datasets were generated by first finding all
421 the enzyme-substrate pairs that are not known as active. Next, substrate similarity was used to pick
422 substrates that are dissimilar to the native substrates of the enzymes using a Tanimoto-based chemical
423 similarity score (Bajusz et al., 2015). The primary reason for using an entirely dissimilar substrate is to
424 ensure that the functional groups present in the native substrate, which might be responsible for the
425 possible activity, are absent in the substrate generated for negative activity. The Tanimoto index uses
426 molecular substructures/fingerprints to find the similarity between two chemical structures. We then pick
427 the top dissimilar substrates and build the dataset so that the positive and negative datasets are in equal
428 proportion. Thus, we generated 11,080 positive and 11,076 negative enzyme-substrate pairs, respectively.
429 Next, we performed an 80-10-10 split of the dataset to generate the training and validation, and test
430 datasets. The training dataset consists of 8,880 pairs of positive enzyme-substrate pairs with known
431 activity from the BRENDA database and 8,876 negative pairs with no known activity (pairs not present in
432 the BRENDA database and substrates dissimilar to the enzyme's native substrate, i.e., zero Tanimoto
433 similarity index). The validation and test dataset both includes 1,100 pairs of both positive and negative
434 enzyme-substrate pairs. The notion here is to use the percentage recovery of the known enzyme-substrate

activity (% positive recovery, TP) as a metric to measure the performance of the CNN model over a dataset that is different from training, validation, and test data.

4.2. Convolutional neural network for enzyme-substrate activity prediction

We adopted the convolutional neural network (CNN) architecture provided by DeepConvDTI (I et al., 2019), which was used for predicting drug-target interactions. The CNN involves convolution over the protein sequence to extract the local residue patterns within the protein sequences and a fully connected layer of the substrates using the molecular fingerprint as a feature. After processing these two layers, the model concatenated these layers and constructed a fully connected layer, which resulted in the output layer (Fig. 7). An exponential linear unit (ELU) function (Clevert et al., 2016) was used as an activation function for every CNN layer except the output layer. ELU functions have been known for speeding up learning in deep neural network models, leading to higher accuracy (Clevert et al., 2016). Here we define a function *ELU* as-

$$ELU(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha(\exp(x) - 1) & \text{if } x \leq 0 \end{cases}$$

Where α is the hyperparameter that controls the ELU function, and x is the input to the activation function. The output layer was activated using the sigmoid function to classify enzyme-substrate pairs as active or inactive. The final scores of the output layer are used to rank-order all the enzymes for novel substrate activity. The entire neural network model was implemented in the Keras python package (Chollet and others, 2015a). The detailed model summary is provided in Supplementary data.

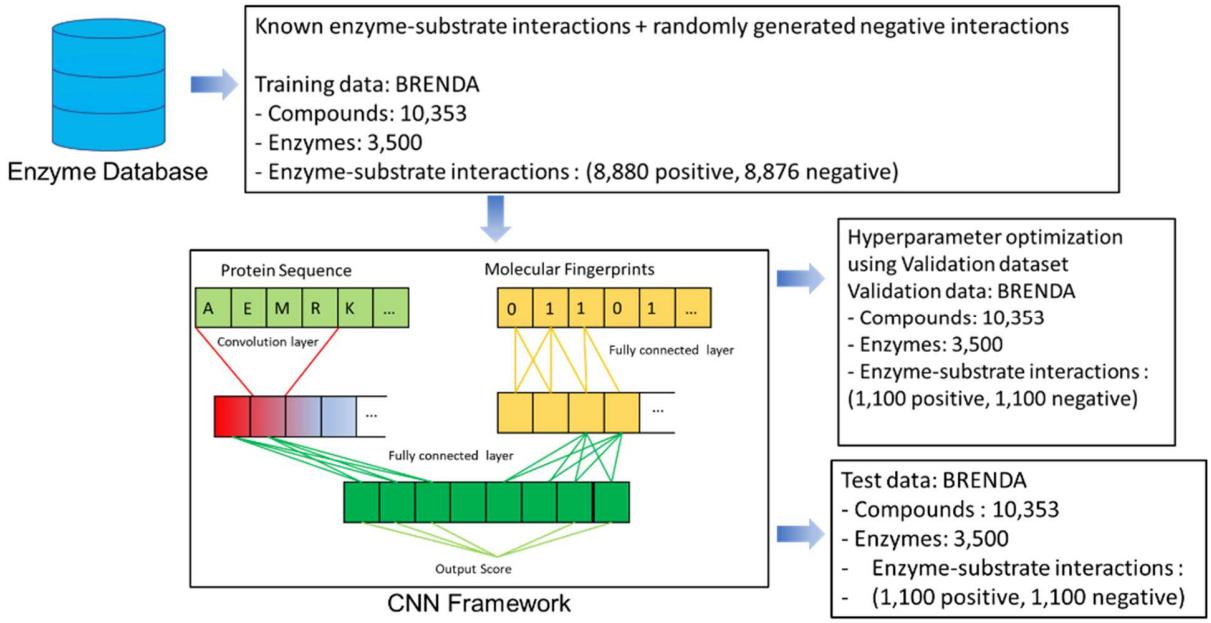


Fig. 8. Framework for the CNN model of EnzRank, uses the BRENDA database for known enzyme-substrate activity. An equal proportion of negative activity was generated using the structural similarity of the enzyme's native structures and chemical compounds in the database and considering the ones that are completely dissimilar as the hard negative dataset. Next, we used the 80:10:10 split of the dataset to train, test, and validate the model. The CNN model uses convolution over the protein sequences and the fully connected layer from the molecular fingerprints and concatenates them together in the final layer to output the final score of an enzyme to have activity on a substrate

4.3. Convolution layer using embedding vectors assigned to the protein sequences

CNN-based models are known to capture important local patterns from the entire space. Fig. 8 shows the overall schema of the convolutional layers. The model starts with an embedding layer which is a lookup table of amino acids to the corresponding embedding vector. We used a Xavier initializer (*viz.*, the 'glorot normal' function in Keras (Glorot and Bengio, 2010)) to randomly initialize the embedding vector values, which imposes normal distribution of the weights and output variance following the variance of input (Glorot and Bengio, 2010). Embedding vectors are trainable, meaning embedding vector values change to optimize the loss during the model training (Chollet and others, 2015). From the lookup table, the embedding layer for the protein sequence is constructed by querying the embedding vector values corresponding to each amino acid in the embedding layer. The length of the embedding matrix for all proteins was fixed to the maximum protein length (*i.e.*, 2,500). The margins were covered using null

labels (*i.e.*, the character $) and the corresponding embedding vectors, which would give a meaningless convolution result filtered out during the global max-pooling. The convolution on embedding layer of protein along the sequence in 1D fashion with one stride (Supplementary data), with the convolution from j^{th} to the $(j+WS)^{th}$ amino acids in sequence, which can be defined as-

$$(x + w)_j = \sum_{m=1}^{ES} \sum_{n=0}^{WS-1} w_{m,n} x_{m,j+n}$$

Convolution over the entire sequence gives (MPL–WS+1) size layer for each filter, where WS is the window size. Finally, to extract the essential local features, we conducted global max-pooling for each filter, which is defined as

$$MaxPooling_{global}\left(E_{P_k}\right) = \max((x * w)_j)$$

Where j covers all of the convolution results of the embedding matrix from the protein sequence P_k. The result is a filter-sized vector with a max-valued convolution result for each window, which does not include bias from the locations of local residue patterns and the maximum protein length. After pooling all convolution results, we concatenated them to represent the essential local interaction patterns as vector-formatted features. Finally, for the organization and abstraction of protein features, concatenated max-pooling results are fed into a fully connected layer, which constructs the latent representation of the protein sequences.

The fully connected layer over the molecular fingerprints was used as the latent representation of the substrates. The latent representation of the data contains all the necessary information needed to represent the original data point (Bishop, 2006). Finally, a fully connected layer was constructed by concatenating the layers to represent proteins and substrates to predict the activity of enzyme-substrate pairs.

4.4. Loss function estimations and weight optimization

The CNN architecture allows the flow of input to the output layer in a feed-forward method. The CNN model estimates the loss using binary cross-entropy functions defined as-

$$loss(W, x) = -\frac{1}{n}\sum_{i}^{n}[y_i \log y_i + (1 - y_i) \log(1 - y_i)]$$

Where W and x are the weights and input to the CNN and y_i and n represents the predicted output and its size (I et al., 2019). We also used L2-norm regularization along with the loss function to avoid overfitting. Specifically, the final loss function can be defined as-

$$loss_{L2}(W, x) = loss(W, x) + \beta \sum_{l=1}^{L-1} ||W^l||_2$$

Where the weights are optimized using the Adam optimizer (Kingma and Ba, 2015) in Keras, which penalizes loss to give a generalized prediction for the model.

4.5. Class Activation Map

We utilized a method based on a gradient-weighted Class Activation Map (grad-CAM) (Selvaraju et al., 2016) to identify the residue level inference on the protein sequence for predicting the enzyme-substrate activity. The goal here is to find residues that contribute most to the activity score. Grad-CAM uses a class-specific gradient information that flows through the convolutional layer of a CNN to produce a localization map of the important regions in the input. Grad-CAM is proven effective in image classification problems (Panwar et al., 2020; Selvaraju et al., 2016). We use grad-CAM to detect residues in the protein sequence that influences the most to the enzyme-substrate activity score.

In grad-CAM, we first compute the contribution of each filter, *k*, in the convolutional layers used to predict the enzyme-substrate activity label *s* by computing the gradient of the output y^s, with respect to the feature map $F_k \in \mathbb{R}^L$ of a convolutional layer over the entire sequence of length L:

$$\alpha_k^s = \sum_{i=1}^{L} \frac{\partial y^s}{\partial F_{k,i}}$$

Where α_k^s is the neuron importance weight of feature map *k* for predicting activity label *s,* estimated by adding the contribution of individual residues. Afterward, we apply the weighted combination of all feature maps in the convolution layer and follow it with a ReLU function to only obtain the residues that have a positive influence on the activity score.

$$C_{grad-CAM}^s[i] = ReLU\left(\sum_k \alpha_k^s F_{k,s}\right)$$

Where $C_{grad-CAM}^s[i]$ denotes the relative importance of residue *i* to activity label *s*. Finally, we take the average of the grad-CAM scores from individual convolutional layers to estimate the residue-level contribution of the protein sequence to the final activity score. The grad-CAM provides an advantage that it can be done as a post-processing step and does not require re-training or change in architecture of the model, making it efficient computationally and directly applicable to the model.

Author Contributions

Vikas Upadhyay: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – Original draft, visualization.

Veda Sheersh Boorla: Conceptualization, Data curation, Writing- Review and Editing

Costas D. Maranas: Supervision, Funding acquisition, Writing – Review and Editing

Data Availability

All the relevant data are within the manuscript and supplementary data. The codes are available at https://github.com/maranasgroup/EnzRank, and the web-based interface is available at https://huggingface.co/spaces/vuu10/EnzRank

Acknowledgements

This work is supported by the U.S. National Science Foundation funded Molecule Maker Lab Institute (MMLI), award number 2019897 supported by National AI Research Institutes Program of the Directorate for Computer and Information Science and Engineering (CISE), in collaboration with the Division of Chemistry (CHE) and the Division of Chemical, Bioengineering, and Environmental Transport Systems (CBET) awarded to CDM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., Church, G.M., 2019. Unified rational protein engineering with sequence-based deep representation learning. Nature Methods 2019 16:12 16, 1315–1322.

Bajusz, D., Rácz, A., Héberger, K., 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform 7, 1–13.

Bateman, A., Martin, M.J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Silva, A. Da, Denny, P., Dogan, T., Ebenezer, T.G., Fan, J., Castro, L.G., Garmiri, P., Georghiou, G., Gonzales, L., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Jokinen, P., Joshi, V., Jyothi, D., Lock, A., Lopez, R., Luciani, A., Luo, J., Lussi, Y., MacDougall, A., Madeira, F., Mahmoudy, M., Menchi, M., Mishra, A., Moulang, K., Nightingale, A., Oliveira, C.S., Pundir, S., Qi, G., Raj, S., Rice, D., Lopez, M.R., Saidi, R., Sampson, J., Sawford, T., Speretta, E., Turner, E., Tyagi, N., Vasudev, P., Volynkin, V., Warner, K., Watkins, X., Zaru, R., Zellner, H., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M.C., Bolleman, J., Boutet, E., Breuza, L., Casals-Casas, C., de Castro, E., Echioukh, K.C., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Estreicher, A., Famiglietti, M.L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Keller, G., Kerhornou, A., Lara, V., Le Mercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T.B., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Pozzato, M., Pruess, M., Rivoire, C., Sigrist, C., Sonesson, K.,

Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Wu, C.H., Arighi, C.N., Arminski, L., Chen, C., Chen, Y., Garavelli, J.S., Huang, H., Laiho, K., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Q., Wang, Y., Yeh, L.S., Zhang, J., 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49, D480–D489.

Berman, H., Henrick, K., Nakamura, H., 2003. Announcing the worldwide Protein Data Bank. Nat Struct Mol Biol 10, 980.

Bishop, C., 2006. Pattern Recognition and Machine Learning. springer.

Blum, T., Kohlbacher, O., 2008. MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization. BIOINFORMATICS APPLICATIONS NOTE 24, 2108–2109.

Cai, Y., Wang, J., Deng, L., 2020. SDN2GO: An integrated deep learning model for protein function prediction. Front Bioeng Biotechnol 8, 391.

Capecchi, A., Probst, D., Reymond, J.L., 2020. One molecular fingerprint to rule them all: Drugs, biomolecules, and the metabolome. J Cheminform 12, 1–15.

Carbonell, P., Parutto, P., Baudier, C., Junot, C., Faulon, J.L., 2014a. Retropath: Automated pipeline for embedded metabolic circuits. ACS Synth Biol 3, 565–577.

Carbonell, P., Parutto, P., Herisson, J., Pandit, S.B., Faulon, J.L., 2014b. XTMS: pathway design in an eXTended metabolic space. Nucleic Acids Res 42.

Carbonell, P., Wong, J., Swainston, N., Takano, E., Turner, N.J., Scrutton, N.S., Kell, D.B., Breitling, R., Faulon, J.L., 2018. Selenzyme: Enzyme selection tool for pathway design. Bioinformatics 34, 2153–2154.

Chen, W., Yao, J., Meng, J., Han, W., Tao, Y., Chen, Y., Guo, Y., Shi, G., He, Y., Jin, J.M., Tang, S.Y., 2019. Promiscuous enzymatic activity-aided multiple-pathway network design for metabolic flux rearrangement in hydroxytyrosol biosynthesis. Nature Communications 2019 10:1 10, 1–12.

Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., Wishart, D.S., 2008. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Res 36.

Chollet, F., others, 2015. Keras.

Chowdhury, A., Maranas, C.D., 2015. Designing overall stoichiometric conversions and intervening metabolic reactions. Sci Rep 5, 16009.

Clevert, D.A., Unterthiner, T., Hochreiter, S., 2016. Fast and accurate deep network learning by exponential linear units (ELUs). 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings.

Dalkiran, A., Rifaioglu, A.S., Martin, M.J., Cetin-Atalay, R., Atalay, V., Doğan, T., 2018. ECPred: A tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. BMC Bioinformatics 19, 1–13.

Delépine, B., Duigou, T., Carbonell, P., Faulon, J.L., 2018. RetroPath2.0: A retrosynthesis workflow for metabolic engineers. Metab Eng 45, 158–170.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1, 4171–4186.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., Rost, B., 2021. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. IEEE TRANS PATTERN ANALYSIS & MACHINE INTELLIGENCE 14.

Feehan, R., Montezano, D., Slusky, J.S.G., 2021. Machine learning for enzyme engineering, selection and design. Protein Engineering, Design and Selection 34, 1–10.

Finley, S.D., Broadbelt, L.J., Hatzimanikatis, V., 2009. Computational framework for predictive biodegradation. Biotechnol Bioeng 104, 1086–1097.

Fleck, C.B., Brock, M., 2009. Re-characterisation of Saccharomyces cerevisiae Ach1p: Fungal CoA-transferases are involved in acetic acid detoxification. Fungal Genetics and Biology 46, 473–485.

Gligorijević, V., Renfrew, P.D., Kosciolek, T., Leman, J.K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B.C., Fisk, I.M., Vlamakis, H., Xavier, R.J., Knight, R., Cho, K., Bonneau, R., 2021. Structure-based protein function prediction using graph convolutional networks. Nature Communications 2021 12:1 12, 1–14.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. Journal of Machine Learning Research 9, 249–256.

Gold, N.D., Fossati, E., Hansen, C.C., DIfalco, M., Douchin, V., Martin, V.J.J., 2018. A Combinatorial Approach To Study Cytochrome P450 Enzymes for De Novo Production of Steviol Glucosides in Baker's Yeast. ACS Synth Biol 7, 2918–2929.

Goldman, S., Das, R., Yang, K.K., Coley, C.W., 2022. Machine learning modeling of family wide enzyme-substrate specificity screens. PLoS Comput Biol 18, e1009853.

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., Rost, B., 2019. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics 20, 1–17.

Hibbert, E.G., Dalby, P.A., 2005. Directed evolution strategies for improved enzymatic performance. Microb Cell Fact 4, 1–6.

Hur, J., Schuyler, A.D., States, D.J., Feldman, E.L., 2009. SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. Bioinformatics 25, 838.

I, L., J, K., H, N., 2019. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput Biol 15.

Jessen, H.J., Liao, H.H., Gort, S.J., Selifonova, O. v, 2014. Beta-alanine/alpha-ketoglutarate aminotransferase for 3-hydroxypropionic acid production. US Patent 8,889,391.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E.,

Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 2021 596:7873 596, 583–589.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K., 2016. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45, 353–361.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E., 2021. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 49, D1388–D1395.

Kingma, D.P., Ba, J.L., 2015. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

Kitazume, T., Tanaka, A., Takaya, N., Nakamura, A., Matsuyama, S., Suzuki, T., Shoun, H., 2002. Kinetic analysis of hydroxylation of saturated fatty acids by recombinant P450foxy produced by an Escherichia coli expression system. Eur J Biochem 269, 2075–2082.

Kumar, A., Wang, L., Ng, C.Y., Maranas, C.D., 2018. Pathway design using de novo steps through uncharted biochemical spaces. Nat Commun 9, 1–15.

Landrum, G., 2006. RDKit: Open-source cheminformatics [WWW Document].

Liang, B., Sun, G., Zhang, X., Nie, Q., Zhao, Y., Yang, J., 2022. Recent advances, challenges and metabolic engineering strategies in the biosynthesis of 3-hydroxypropionic acid. Biotechnol Bioeng 119, 2639–2668.

Lin, G.M., Warden-Rothman, R., Voigt, C.A., 2019. Retrosynthetic design of metabolic pathways to chemicals not found in nature. Curr Opin Syst Biol.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Liu, Z., Zhang, X., Lei, D., Qiao, B., Zhao, G.-R., 2021. Metabolic engineering of Escherichia coli for de novo production of 3-phenylpropanol via retrobiosynthesis approach. Microbial Cell Factories 2021 20:1 20, 1–15.

Lowe, D.M., Corbett, P.T., Murray-Rust, P., Glen, R.C., 2011. Chemical name to structure: OPSIN, an open source solution. J Chem Inf Model 51, 739–753.

Lunnon, W.F., Brunvoll, J., Cyvin, S.J., Cyvin, B.N., Balaban, A.T., 1988. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. J Chem Inf Comput Sci 28, 31–36.

Mc Naught, a. D., Wilkinson, a, 2012. Compendium of Chemical Terminology-Gold Book. Iupac 1670.

Moriya, Y., Yamada, T., Okuda, S., Nakagawa, Z., Kotera, M., Tokimatsu, T., Kanehisa, M., Goto, S., 2016. Identification of Enzyme Genes Using Chemical Structure Alignments of Substrate-Product Pairs. J Chem Inf Model 56, 510–516.

Nakayama, N., Takemae, A., Shoun, H., 1996. Cytochrome P450foxy, a Catalytically Self-Sufficient Fatty Acid Hydroxylase of the Fungus Fusarium oxysporum1. The Journal of Biochemistry 119, 435–440.

Panwar, H., Gupta, P.K., Siddiqui, M.K., Morales-Menendez, R., Bhardwaj, P., Singh, V., 2020. A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. Chaos Solitons Fractals 140, 110190.

Perin, G., Buhan, I., Picek, S., 2021. Learning When to Stop: A Mutual Information Approach to Prevent Overfitting in Profiled Side-Channel Analysis. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 53–81.

Pertusi, D.A., Moura, M.E., Jeffryes, J.G., Prabhu, S., Walters Biggs, B., Tyo, K.E.J., 2017. Predicting novel substrates for enzymes with minimal experimental effort with active learning. Metab Eng 44, 171–181.

Pertusi, D.A., Stine, A.E., Broadbelt, L.J., Tyo, K.E.J., 2015. Efficient searching and annotation of metabolic networks using chemical similarity. Bioinformatics 31, 1016–1024.

Porter, J.L., Boon, P.L.S., Murray, T.P., Huber, T., Collyer, C.A., Ollis, D.L., 2015. Directed evolution of new and improved enzyme functions using an evolutionary intermediate and multidirectional search. ACS Chem Biol 10, 611–621.

Rahman, S.A., Advani, P., Schunk, R., Schrader, R., Schomburg, D., 2005. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). BIOINFORMATICS ORIGINAL PAPER 21, 1189–1193.

Rios, J., Lebeau, J., Yang, T., Li, S., Lynch, M.D., 2021. A critical review on the progress and challenges to a more sustainable, cost competitive synthesis of adipic acid. Green Chemistry 23, 3172–3190.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., Fergus, R., 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A 118.

Rogers, D., Hahn, M., 2010. Extended-connectivity fingerprints. J Chem Inf Model 50, 742–754.

Ryu, J.Y., Kim, H.U., Lee, S.Y., 2019. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci U S A 116, 13996–14001.

Sanderson, T., Bileschi, M.L., Belanger, D., Colwell, L.J., 2021. ProteInfer: deep networks for protein functional inference. bioRxiv 2021.09.20.461077.

Schomburg, I., Chang, A., Schomburg, D., 2002. BRENDA, enzyme data and metabolic information, Nucleic Acids Research.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2016. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization.

Simon, C., Davidsen, K., Hansen, C., Seymour, E., Barnkob, M.B., Olsen, L.R., 2019. BioReader: A text mining tool for performing classification of biomedical literature. BMC Bioinformatics 19, 165–170.

Srivastava, N., Hinton, G., Krizhevsky, A., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 15, 1929–1958.

Taylor, J.L., Price, J.E., Toney, M.D., 2015. Directed evolution of the substrate specificity of dialkylglycine decarboxylase. Biochim Biophys Acta 1854, 146.

Tian, C., Yang, J., Liu, C., Chen, P., Zhang, T., Men, Y., Ma, H., Sun, Y., Ma, Y., 2022. Engineering substrate specificity of HAD phosphatases and multienzyme systems development for the thermodynamic-driven manufacturing sugars. Nature Communications 2022 13:1 13, 1–13.

Wang, L., Dash, S., Ng, C.Y., Maranas, C.D., 2017. A review of computational tools for design and reconstruction of metabolic pathways. Synth Syst Biotechnol 2, 243–252.

Wang, L., Ng, C.Y., Dash, S., Maranas, C.D., 2018. Exploring the combinatorial space of complete pathways to chemicals. Biochem Soc Trans.

Wang, L., Upadhyay, V., Maranas, C.D., 2021. dGPredictor: Automated fragmentation method for metabolic reaction free energy prediction and de novo pathway design. PLoS Comput Biol 17, e1009448.

Wen, F., McLachlan, M., Zhao, H., 2008. Directed Evolution: Novel and Improved Enzymes. Wiley Encyclopedia of Chemical Biology 1–10.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2015. Learning Deep Features for Discriminative Localization.

Zou, Z., Tian, S., Gao, X., Li, Y., 2019. mlDEEPre: Multi-functional enzyme function prediction with hierarchical multi-label deep learning. Front Genet 10, 714.