CLR_ESP: Improved enzyme-substrate pair prediction using contrastive learning

Zhenjiao Du[1], Weiming Fu[2], Xiaolong Guo[2], Doina Caragea[3], Yonghui Li[1,*]

[1]Department of Grain Science and Industry, Kansas State University, Manhattan, KS 66506, USA

[2]Department of Electrical and Computer Engineering, Kansas State University, Manhattan, KS 66506, USA

[3]Department of Computer Science, Kansas State University, Manhattan, KS 66506, USA

*Corresponding author: Dr. Yonghui Li, 785-532-4061, yonghui@ksu.edu

Abstract

To reduce the cost of experimental characterization of the potential substrates for enzymes, machine learning prediction model offer an alternative solution. Pretrained language models, as powerful approaches for protein and molecule representation, have been employed in the development of enzyme-substrate prediction models, achieving promising performance. In addition to continuing improvements in language models, effectively fusing encoders to handle multimodal prediction tasks is critical for further enhancing model performance using available representation methods. Here, we present CLR_ESP, a multimodal classifier that integrates protein and chemistry language models with a newly designed contrastive learning strategy for predicting enzyme-substrate pairs. Our best model achieved SOTA performance with an accuracy of 94.70% on independent test data while requiring fewer computational resources and training data. It also confirmed our hypothesis that embeddings of positive pairs are closer to each other in high-dimension space, while negative pairs exhibit the opposite trend. The proposed architecture is expected to be further applied to enhance performance in additional multimodality prediction tasks in biology. A user-friendly web server of CLR_ESP is established and freely accessible at https://78k6imn5wp.us-east-1.awsapprunner.com/.

Keywords: Large language model; multimodal model; protein-compound interaction; contrastive learning

Introduction

Most enzymes are proteins capable of catalyzing a wide range of reactions within living organisms or under mild conditions in vitro with up to over a million-fold compared to spontaneous rates (Kim et al., 2023; Kroll, Ranjan, Engqvist, et al., 2023). Moreover, enzymes typically exhibit promiscuity, facilitating multiple reactions that may include physiologically irrelevant or potentially harmful processes (Copley, 2017; Nobeli et al., 2009). A comprehensive mapping of enzyme-substrate relationships will provide crucial guidance for future research in medicine, pharmaceuticals, bioengineering, and agriculture (Kroll, Ranjan, & Lercher, 2023; D. Zhang et al., 2023, 2024). However, it is prohibitively expensive to experimentally determine the catalytic interactions between molecules and enzymes. According to the UniProt Knowledgebase, approximately 10.8 million entries are related to enzymes, yet only 0.6% of these sequences have high-quality annotations of catalyzed reactions that are manually curated (The UniProt Consortium, 2021). There is a pressing need for high-throughput methods to address the scarcity of experimental validation in enzyme-substrate relationships.

Since the introduction of transformer architecture (Vaswani et al., 2017), various machine learning tasks have been revolutionized by leveraging pretrained large language models (LLMs) for molecular and protein representation, achieving significant advancements (Ahmad et al., 2022; Akdis, 2021; Du, Ding, et al., 2023; Lin et al., 2023; Ross et al., 2022, 2022; Q. Zhang et al., 2024). Molecules and proteins can be represented as strings similar to natural language, using vocabularies such as the simplified molecular-input line-entry system (SMILES) for molecules and single-letter amino acid codes for proteins. Besides, protein and molecular databases (e.g., UniProt and ZINC) contained billions of protein sequences and molecular data, allowing to pretrain protein/chemistry LLMs. Based on their architecture, LLMs can be classified into encoder-only models (e.g., BERT - bidirectional encoder representations from transformers), decoder-only models (e.g., GPT - generative pre-trained transformers), and encoder-decoder models (e.g., T5 - text-to-text translation transformer). Among these, encoder-only architectures have gained popularity for remarkable performance in various downstream classification and regression tasks through transfer learning or finetuning (Ahmad et al., 2022; Du, Ding, et al., 2023; Kroll, Ranjan, & Lercher, 2023; Ross et al., 2022; Thumuluri et al., 2022; Q. Zhang et al., 2024). Recently, machine learning approaches for predicting compound-protein interaction (CPI)

have increasingly adopted these encoder-only models (Nguyen et al., 2023, 2024; Song et al., 2023; Wang et al., 2022), and enzyme-substrate pair prediction task is a subset of the CPI prediction task (Kroll, Ranjan, Engqvist, et al., 2023; Kroll, Ranjan, & Lercher, 2023; Mou et al., 2021; Yang et al., 2018). Compared to traditional approaches using physicochemical parameters of molecules, enzyme sequences, or enzyme active site descriptors, employing pretrained BERT style encoders allows for better representation of proteins and molecules due to the domain knowledge learned via self-supervised learning from enormous datasets (Du, Ding, et al., 2023; Kroll, Ranjan, Engqvist, et al., 2023; Lim et al., 2021; Mou et al., 2021; Yang et al., 2018; Q. Zhang et al., 2024).

Enzyme substrate pairs prediction model is a multimodal prediction task, where both two different "language" systems (i.e., SMILES and amino acid sequences) were involved during the embedding process. How to effectively leverage the enzyme embeddings and molecular embeddings generated from corresponding encoders become a crucial point in model performance enhancement. The most popular strategy is to simply concatenate the protein and molecule embeddings and this strategy and its modifications has been a common practice in the CPI discovery community (Lee et al., 2019; Nguyen et al., 2024; Wang et al., 2022). So far, the SOTA performance in enzyme-substrate pair prediction was achieved with the simple concatenation strategy in the study of Kroll et al., where a task-specific fine-tuned protein language model (PLM) (i.e., ESM-1b) and a pretrained domain-specific graph neural network (GNN) was used for embeddings and obtained a accuracy of 91.5% (Kroll, Ranjan, Engqvist, et al., 2023). Similar strategy was also used in the study of Song et al., where Kronecker products of protein and molecule embeddings as additional features was concatenated with the original embeddings to enhance CPI prediction performance (Song et al., 2023). Beyond simple concatenation, cross-attention blocks have been employed to integrate local and high-level information from multiple protein and molecule encoders, capturing the relationship between protein and molecule representations for fused representation in CPI prediction (Nguyen et al., 2023, 2024). Recently, Kroll et al. introduced a multimodal BERT model for embedding protein-molecule complex, where the protein sequences and SMILES was input to multimodal BERT model. They achieved SOTA performance across four datasets, including drug-target interactions, protein-small molecule interactions, enzyme-substrate Michaelis constants (KM),

and substrate identification for enzymes, by concatenating PLM-generated protein embeddings, chemically learned molecule (CLM)-generated molecule embeddings, and multimodal encoder-generated protein-molecule embeddings (Kroll, Ranjan, & Lercher, 2023). To maximize the prediction performance, these models either employed more encoders to enrich embedding protein and molecule or fused embeddings to enhance enzyme-substrate complex representations. However, these methods often require complex fusion architecture design and substantial computational resources, which impeded the deployment and application to different application scenarios.

Inspired by the success of CLIP, where two independent encoders for images and texts were jointly trained to predict 400 million correct image-text pairs (Radford et al., 2021), we assumed that the enzyme and the substrate in a correct pair should be closer after being projected into a high dimension space while they should be distinctly separated after projection if unrelated. This principle had been explored under single modality setting. To predict the peptide-receptor complex pair, ESM-1b and ESM-MSA-1b were connected with an independently encoders respectively for peptide embeddings and protein receptors. The additional peptide encoder and the protein receptor encoder were jointly trained to capture inter-embedding similarities in a CLIP style model architecture (Bhat et al., 2023; Palepu et al., 2022). Instead of training additional encoders, a project head was designed to refine the embeddings with less computation resource consumption. Such an strategy was employed in single modality task of enzyme commission (EC) number prediction. Yu et al. utilized frozen ESM-1b for enzyme embeddings and integrated a simple projection head with a contrastive loss function to refine embeddings (Yu et al., 2023). This study achieved SOTA performance among various benchmark datasets. Specifically, the model exhibited remarkable performance in rare EC numbers. To our best knowledge, our model was the first one employing the CLIP idea and the projection head design idea from popular contrastive learning model architectures (e.g., SimCLR, BYOL, MOCO, etc.) to solve the multimodality prediction task (i.e., enzyme substrate pair prediction) (Chen et al., 2020; Grill et al., 2020; He et al., 2020; Radford et al., 2021; Richemond et al., 2020).

Our goal was to develop an advanced machine learning module tailored for empowering existing LLMs to predict enzyme-substrate relationships across diverse proteins and molecules, offering a practical tool to streamline experimental efforts and boost laboratory efficiency. In this study, we

presented a simplified yet highly effective model architecture employing a contrastive learning strategy. The architecture exhibited extraordinary performance enhancement in knowledge fusion under multimodal context with limited data. Specifically, we utilize ESM-2 for enzyme embeddings and MolFormer for molecule embeddings. Instead of fine-tuning these encoders, adding additional encoders, or adding more features (e.g., extended-connectivity fingerprints), we designed a projection head layer for the two encoders to refine and project embeddings into the same high-dimensional space (Fig. 1). Our models achieve superior performance compared to previous studies, with simplified architecture and reduced computational demands during both training and inference phases.

Methods

Dataset

To ensure a fair comparison with previous studies, the datasets used in this study were sourced from existing studies (Kroll, Ranjan, Engqvist, et al., 2023; Kroll, Ranjan, & Lercher, 2023). The positive enzyme-substrate pairs were all from Gene Ontology (GO) annotation database, where entries have different levels of evidence: experimental, phylogenetically-inferred, computational analysis, author statement, curator statement, and electronic evidence. We utilized two datasets: one based on experimental evidence and another on phylogenetic evidence. To challenge the model to distinguish true from false substrates, negative pairs were generated by randomly sampling three small molecules highly similar to the true substrates for the same enzyme sequences (Kroll, Ranjan, Engqvist, et al., 2023).

The experimental evidence-based dataset, originally split into training, validation, and test sets in the original study (Kroll, Ranjan, Engqvist, et al., 2023; Kroll, Ranjan, & Lercher, 2023), contained 50,093 enzyme-substrate pairs in the training set, 5,422 pairs in the validation set, and 13,336 pairs in the test set. In contrast, the phylogenetic evidence-based dataset comprised a total of 765,635 pairs. These two datasets were downloaded from https://github.com/AlexanderKroll/ESP and https://github.com/AlexanderKroll/ProSmith, respectively.

Because of the computational demands of processing long protein sequences, we excluded some positive or negative enzyme-substrate pairs whose sequence embeddings exceeded hardware capacity (NVIDIA A100 GPU). Specifically, for ESM-2 models with output dimensions of 480 and 1280, sequences longer than 8000 were removed. Sixteen pairs were removed from the training dataset and no pair was removed from the test dataset. For the ESM-2 model with a 2560-dimensional output, sequences longer than 5500 were excluded. Correspondingly, 212 pairs were removed from the training dataset. Notably, no sequence in the validation and test datasets was removed.

Calculating enzyme representation

In this study, enzymes were represented numerically using ESM2 models(Lin et al., 2023; Rives et al., 2021). ESM is a PLM project, initiated by Meta Fundamental AI Research (FAIR) in 2019 (https://github.com/facebookresearch/esm), which includes 19 pretrained PLMs with various dimensional output embeddings based on modified Bidirectional Encoder Representations from Transformers (BERT) architecture. Those PLMs were trained on large protein sequence datasets (e.g., UR50/D 2021_04) through self-supervised learning. In this study, we employed three ESM-2 models, including the 480-dimensional output, 1280-dimensional output, and 2560-dimensional output, denoted as ESM-2-480, ESM-2-1280, and ESM-2-2560, respectively. The details about each PLM are listed in Supplementary Table 1. The amino acid one letter code of each enzyme was loaded into the PLM for embeddings. For an enzyme with L amino acids, the output of the ESM-2-n model for the enzymes is a $(L+1)*n$ matrix, where n represents the output dimension of the ESM-2 model. The first row is the representation of the [sos] token, which is the indicator of “start of sentence” during training process. After that, each row in the matrix represented the corresponding amino acid from N-terminal to C-terminal with the consideration of the amino acid itself and the context of the entire sequence. We used an average pooling operation to unify the output dimension of enzymes with different lengths. Finally, any enzyme sequence loaded into the ESM-2-n for embeddings would result in a $1*n$ dimensional vector as its representation.

Calculating molecule representation

MoLFormer was selected for numerical representation generation of molecules (Ross et al., 2022). MoLFormer was developed by IBM Research in 2022 and trained on canonical SMILES sequences of 1 billion molecules from ZINC database and 111 million molecules from PubChem with employment of rotary positional embeddings and linear attention mechanism. There was only one model available, called "MoLFormer-XL-both-10pct" trained with 10% of both datasets, at https://huggingface.co/ibm/MoLFormer-XL-both-10pct. Although the dataset size was smaller, this model demonstrated performance comparable to the full-size model. Therefore, we used this model for the molecule embeddings. We will use "MoLFormer" to refer this model in this paper. MoLFormer has a BERT-like architecture, similar to that of ESM models. To encode a canonical SMILES, the SMILES sequence was first converted into a numerical matrix, and then an average pooling operation was conducted to unify the feature dimension of SMILES sequences of different length into a 1*768 dimensional vector.

Encoding models

ESM-2 models represent an upgrade over ESM-1b used in previous studies (Kroll, Ranjan, Engqvist, et al., 2023; Kroll, Ranjan, & Lercher, 2023). They were trained on a new dataset (UR50/D 2021_04), while ESM-1b model was trained on UR50/S 2018_03 (Lin et al., 2023; Rives et al., 2021). Besides, the ESM-2 models employed rotary position embedding to replace the learned position embedding, which allowed it to embed any length of protein sequences, while ESM-1b can only embed sequences shorter than 1023 amino acids (Lin et al., 2023; Rives et al., 2021). As for MoLFormer, it was trained on a larger dataset than the one used in previous study (i.e., ChemBERTa-2) and employed more advanced algorithms (i.e., rotary position embedding and linear attention mechanism) (Kroll, Ranjan, & Lercher, 2023). It exhibited better performance in one regression task (Lipophilicity dataset) and three classification tasks (BACE, BBBP, and ClinTox datasets) during downstream task performance evaluation (Ahmad et al., 2022; Ross et al., 2022).

Model architecture design

The model architecture comprised two independent modules with similar inner architecture design for enzyme sequence input and molecule SMILES input, respectively (Fig. 1). The

enzyme module utilized an ESM model as the encoder for enzyme sequence embeddings, followed by average pooling to unify the embedding dimensions. These embeddings were then processed through a projection head to refine the dimension from 480/1280/2560 to 128. Similarly, the molecule module employed MoLFormer as the encoder, followed by average pooling and a projection head to refine the dimension from 768 to 128. Every enzyme substrate pair was finally transformer into two 128 dimensional vectors, where one represented the enzymes and the other represented the molecule. Cosine similarity between the two vectors was calculated to finally as the predicted label of the pair. The truth label of the positive enzyme-substrate pairs was 1 and the true label of the negative pair was 0. A mean squared error (MSE) loss function was employed for the loss calculation as following:

$$Loss = \frac{1}{n}\sum_{i=1}^{n}(Sim_i - Label_i)^2$$

, where i was the number of data points; Sim_i was the cosine similarity of the 128 dimensional vectors of enzyme and substrate in the single pair; the $Label_i$ was the true label.

The projection head mainly referred to the projection heads in contrastive learning studies with a few modifications (Chen et al., 2020; Grill et al., 2020; Richemond et al., 2020). The two projection heads for enzyme embeddings and molecular embeddings did not share weights and were trained independently through the MSE loss function except for the second batch norm layer. While the contrastive learning idea was inspired from CLIP (Radford et al., 2021), there was no huge dataset available for training, similar to the one that CLIP used to train the image and text encoders from scratch. Thus, in this study, we leveraged two pretrained encoders for embeddings to save the computational resources required for LLMs' training and bypass the need for a huge enzyme-substrate pair dataset. At the same time, we employed the knowledge/expert-based negative dataset to address the challenges of scarcity in negative data points. The encoders were frozen and two independent projection heads were trained jointly as the learnable adaptor to maximize the model performance.

For each projection head, the number of neurons in the first fully connected layer corresponded to the encoder's output dimension after average pooling. A batch normalization layer followed

the fully connected layer before the ReLU activation function. Subsequently, the second fully connected layer (bottleneck layer) projected the input into a 128-dimensional vector, employing batch normalization and L2 normalization. The 128 dimension was selected mainly referring to SimCLR (Chen et al., 2020). The neurons in the first fully connected layer varied based on the encoder selected for model development. For instance, with the optimal combination of ESM-2-2560 and MoLFormer, the first fully connected layers had 2560 and 768 neurons, respectively. Following projection into a 128-dimensional vector, both enzyme and molecule vectors shared the same batch normalization layer, assuming they were at the same high-dimensional space.

Model Training

To expedite training and reduce computation needs, enzyme and molecule embeddings were pre-generated and saved for projection head training, bypassing the iterative generation of embeddings during training process. Adam optimizer was employed with default parameters. The batch size was 16 for the training on the experimental evidence-based datasets and 512 for the training on the phylogenetic evidence-based dataset. A relatively small batch size for experimental evidence-based datasets resulted from the small size of the dataset and the better capability of capturing the nuances, while a larger batch size was selected for the larger phylogenetic evidence-based dataset and captured the general features among positive/negative pairs.

The base model, CLR_ESP-exp, trained on the experimental evidence-based dataset, underwent 500 epochs, with the best-performing model checkpoint saved based on validation dataset performance. Training typically took 2-3 hours using a Tesla T4 GPU on Google Colab. The model CLR_ESP-XL, trained on both the phylogenetic and experimental evidence-based datasets, underwent similar processes, beginning with training on the phylogenetic evidence-based dataset for 500 epochs, followed by an additional 30 epochs on the experimental evidence-based dataset. The epoch numbers were selected based on learning curves (Supplementary Figure S1). The model, after training on a phylogenetic evidence-based dataset, was noted as CLR_ESP-phylo. At each state, best model checkpoint was saved based on validation dataset performance during this extended training process.

Software

The neural network models were all implemented with Python and trained using the PyTorch library. The datasets and codes used to generate the results of this paper are available from https://github.com/dzjxzyd/CLR_ESP.

A user-friendly web server was deployed at Amazon Web Services (AWS) App Runner. The website was designed with html and css scripts, and the model deployment was achieved with Flask (2.2.2). Due to the constraint of computation resources, the CLR_ESP-XL with ESM-2-1280 and MolFormer was deployed for enzyme-substrate prediction. The web server supports large-scale processing, which allows users to upload their unknown enzyme-substrate pairs through xls or xlsx formats. The detailed scripts for web server development are available at https://github.com/dzjxzyd/pLM4CPP_server.

Results and Discussion

Model performance trained on experimental evidence-based datasets only

The dataset was divided into training, validation, and test sets as the original paper (Kroll, Ranjan, & Lercher, 2023). Three different ESM-2 models and MoLFormer were employed as encoders for enzyme and molecule representations, respectively. CLR_ESP-exp was exclusively trained on the provided experimental evidence-based datasets. Although our training were slightly smaller compared to previous models, all models, including ours, were evaluated on the same dataset, ensuring a fair final performance comparison.

To assess the superiority of our contrastive learning strategy in enhancing performance, we also implemented three counterpart baseline models with the popular concatenation strategy. We employed a simple two-layer neural network as the classifier to learn from the training data, details of which can be found in our GitHub repository. The results from all 6 models are shown in Table 1. It can be seen that all the models based on our contrastive learning strategy outperformed their counterparts, especially the CLR_ESP-exp with ESM-2-480 and MoLFormer, where the accuracy was 6.7% higher than the model with ESM-2-480 and MoLFormer for embeddings and simple concatenation strategy. This underscores the significant advantage of our

proposed contrastive learning strategy over the widely used concatenation strategy. A frozen BERT-based encoder paired with learnable neural networks as a projection head and contrastive loss function has previously shown success in single-modality conditions EC number prediction (Yu et al., 2023). In this paper, we made some modifications to adapt this idea in multimodal prediction task by employing two independent projection heads and an MSE loss function referring to CLIP model. The promising results validated our hypothesis that positive enzyme-substrate pairs tend to be closer in high-dimensional space, while negative pairs are more distant. Moreover, due to the small size of the projection heads, our approach required a relatively modest dataset size, making it well-suited for biological applications with limited data availability.

Additionally, a positive correlation between model performance and the size of ESM models/output dimensions was observed in both the contrastive learning and concatenation strategy groups. This trend aligns with findings from our previous studies (Du, Xu, et al., 2023), where larger model sizes/output dimensions encoded more information into the embeddings, resulting in improved overall performance.

The CLR_ESP-exp models achieved superior performance compared to existing SOTA models, with accuracy, and area under the curve, and Matthew's correlation coefficient (MCC) ranging from 92.21% to 93.57%, 0.9468 to 0.9594, and 0.7945 to 0.8314. In Kroll et al.'s study, models combining ESM-1b or fine-tuned ESM-1b with a pretrained GNN on the same datasets outperformed models using ESM-1b and traditional extended-connectivity fingerprints (ECFP) for enzyme and substrate representation, which exhibited the superiority of pretrained model GNN over ECFP-based features. The performance of ESM-1b and GNN model achieved 88.8% accuracy, which was quite close to the one (ACC = 88.29%) of the CLR_ESP-exp composed of ESM-2-1280 & MoLFormer based models with simple concatenation strategy. The enzyme encoders in both models had the same model size (650 million parameters) and output dimension (1280 dimensions), were some modifications in ESM-2 regarding the training set and model architectures. As for the molecule encoders, MoLFormer was pretrained for general molecule representation, while GNN was pertained on a domain specific prediction task (i.e., production of Michaelis constants K_M of enzyme-substrate pairs) which was closer to the application scenario (i.e., enzyme-substrate pair prediction) (Kroll, Ranjan, Engqvist, et al., 2023). Besides,

the pretrained GNN was still learnable during the training task on the enzyme-substrate pair dataset, and additional hyperparameter optimization for the gradient-boosting classifier was used, which further enhanced the performance. Those reason together contributed to slightly better performance of the ESM-1b & GNN based model over our simple concatenation models (ESM-2-1280 & MoLFormer), though we employed an advanced encoders for enzymes. When employing the contrastive learning strategy, CLR_ESP-exp with ESM-2-1280 & MoLFormer achieved 92.94% accuracy which was 4.7% higher than the model with ESM-1b & GNN. The second model from Kroll et al., employed a task specific fine-tuned ESM-1b model which was finetuned on 200634 enzyme substrate pairs with phylogenetically inferred evidence. The additional finetuning process enhanced the representation power of ESM-1b and increased the accuracy from 88.8 % to 91.5 %. However, compared to our model with different architectures, the performance was inferior to ours, where no additional fine-tuning process was needed, highlighting the effectiveness of our proposed architectures for enzyme-substrate prediction tasks.

Ablation study on the best-performed model

To further explore the function of each layer in the projection head, we conducted an ablation study on the best-performing CLR_ESP-exp model using ESM-2-2560 and MoLFormer (as discussed in the previous section). We systematically investigated the effects of the first dense layer, ReLU activation function, bottleneck layer dimension, batch normalization layer, L2 normalization layer, and loss function. Table 2 summarizes the results, indicating that the cross removal of the first dense layer and entropy loss function had a detrimental effect on model performance. Additionally, the bottleneck size, ReLU activation, and batch normalization layers contributed marginally to performance improvement, whereas the L2 normalization layer had negligible impact.

Pretraining the model on phylogenetic evidence-based data further enhances performance

In this section, we further enhanced model performance by training on a larger dataset combining phylogenetic evidence-based and experimental evidence-based datasets. Assuming the phylogenetic evidence-based dataset contained rich relationship information between enzymes and substrates, we directly trained models on this dataset for 500 epochs and subsequently the

model underwent further training on the experimental evidence-based dataset for 30 epochs to optimize performance. Following the approach from our reference paper (Kroll, Ranjan, & Lercher, 2023), we trained and evaluated the model performance on the same datasets for fair performance comparison. The results were shown in Table 3.

Notably, all the models trained on phylogenetic evidence-based dataset only (CLR_ESP-phylo) outperformed all the models from Table 1 (CLR_ESP-exp), including models employing ESM-1b and ChemBERTa. The phylogenetic evidence-based dataset contained 764,449 data points, which was around 15 times that of the experimental evidence-based dataset. We can infer that the number of learnable parameters were not yet saturated under current datasets and can be expected to further enhance performance with a larger dataset.

Similarly, larger encoder-based models exhibited better performance, with CLR_ESP-phylo and CLR_ESP-XL using ESM-2-2560 achieving accuracies of 93.98% and 94.77%, respectively, compared to 93.93% and 94.56% from models using ESM-2-1280. Besides, an improvement in was also observed after the models were further trained on the experimental evidence-based dataset. This is not surprising, as the data points in training dataset from experimental evidence-based dataset were much closer to the data points in test dataset, and thus allowing the model to perform better in test dataset. There was also a performance improvement between the models with ESM-2-1280 and MoLFormer and the models with ESM-1b and ChemBERTa-2 under the same contrastive learning strategy. It mainly resulted from the two encoders. ESM-2-1280 and MoLFormer were trained with a larger or updated datasets and advanced architectures (e.g., rotary position embeddings, linear attention mechanism) and achieved better performance than that of ESM-1b and ChemBERTa-2 in downstream prediction tasks (Ahmad et al., 2022; Lin et al., 2023; Ross et al., 2022). It is worth noting that ESM-1b can only embed the enzyme sequence shorter than 1024 amino acids or truncated enzymes sequences with the first 1023 amino acids, which caused the information loss and inferior performance.

Compared to the previous SOTA model, ProSmith designed by Kroll et al., our model trained and evaluated on the same datasets achieved superior performance (Kroll, Ranjan, & Lercher, 2023). In the study of ProSmith, besides employing ESM1b and ChemBERTa-2 for enzyme and molecule embeddings, they introduced an additional multimodal BERT model trained on the

combined text of enzyme sequences and molecular SMILES. The BERT model was pretrained on 1,039,565 data points from ligand target affinity dataset (Kroll, Ranjan, & Lercher, 2023). Therefore, ProSmith employed three encoders for embeddings and connected them with a gradient boost model for prediction. In contrast, our model (CLR_ESP-XL) achieved 94.77% accuracy using only two pretrained encoders without additional fine-tuning.

In order to explore the contribution to the performance of our proposed architecture , we employed ESM1b and ChemBERTa-2 for enzyme and molecule embeddings and trained the model with our architecture with modifications according to the input dimensions. The performance was slightly better than that of ProSmith. The main contribution to our promising performance was attributed to our simple projection heads, effectively fusing embeddings from two encoders for prediction tasks, comparable to the effect of an additional pretrained encoder and well-optimized gradient boost classifier. The main part that contributed to the promising performance was the two simple projection heads, in which the embeddings from two encoders were effectively fused for the prediction task and can be equal to the contribution of additional pretrained encoders and a gradient boost classifier. At the same time, the ESM1b model was based on learned position embeddings, and it cannot embed the enzymes whose sequence length is longer than 1023 amino acids (Rives et al., 2021). Therefore, our training dataset was smaller than the one used in ProSmith, but our performance was still comparable. Such a comparison further demonstrated the superiority of the proposed architecture.

Characterizing CLR_ESP-XL performance in unseen enzymes and small molecules

In order to further characterize the model performance in rarely seen enzymes, we first split the test dataset into three subgroups: datapoints with enzymes with a maximal sequence identity to training data between 0 and 40%, between 40% and 60%, and between 60% and 80% based on CD-HIT (Li & Godzik, 2006). The results were shown in Supplementary Table S2. It was observed that CLR_ESP-XL models achieved higher performance in enzyme-substrate pairs with higher sequence identity. Specifically, the accuracy, AUC, and MCC were 96.85%, 0.9871, 0.9198, respectively for enzymes with 60-80% sequence identity. The model still performed well

for enzymes with 0-40% sequence identity with an accuracy of 93.07%, AUC of 0.9476, and MCC of 0.8185. It is worth noting that our model also exhibited better performance in the three subsets compared to the performance of ESP (Supplementary Table S2) (Kroll, Ranjan, Engqvist, et al., 2023). Besides, our performance is also highly comparable to the performance of ProSmith in the three subsets (Kroll, Ranjan, & Lercher, 2023). In summary, while using similar enzymes, the more similarity between the candidate enzymes and the enzymes in our dataset, the higher confident prediction the model can make.

CLR_ESP-XL models can express uncertainty

Internally, our model can also provide the cosine similarity score instead of only the positive or negative prediction results to interpret how confident the model it is regarding its prediction. In this paper, we set 0.5 as the threshold for output results, where a cosine similarity score between an enzyme and a small molecule was predicted as a positive pair if the score is higher than 0.5, and otherwise they are a negative pair. The cosine similarity score was also output at our webserver at https://78k6imn5wp.us-east-1.awsapprunner.com for single reaction prediction or large scale prediction.

To provide a more detailed view of prediction accuracies, Fig. 2 displayed the distributions of true (blue) and false (orange) predictions for our test set across prediction scores. Most of the correct predictions had a score either close to 0 or close to 1, which means CLR_ESP-XL made prediction with high confidence. Conversely, the false prediction is distributed quite evenly across prediction scores. The prediction for the data points with a predicted score ranging from 0.4 to 0.6 are likely to be wrongly predicted. Therefore, for the practical usage of CLR_ESP-XL, the input pairs with a cosine similarity score between 0.4 – 0.6 should be considered uncertain and carefully used for decision making.

Overall, our proposed contrastive learning strategy achieved SOTA performance, relying on two frozen encoders and two simple projection heads with an MSE loss function. The best model (CLR_ESP-XL with ESM-2-2560 and MoLFormer) achieved SOTA performance with accuracy, AUC, and MCC of 94.77%, 0.9653 and 0.8628, respectively. Under limited data availability conditions (experimental evidence-based dataset only), our model CLR_ESP-exp with ESM-2-

2560 and MoLFormer also achieved the SOTA performance with accuracy, AUC, and MCC of 93.57%, 0.9594 and 0.8314, respectively. Furthermore, our models can cope with proteins of any length while ProSmith truncates all the enzymes into 1023 amino acids because of the constrains from ESM-1b models, which would confuse the model when two different enzymes share the same first 1023 residues from N-C terminal. Bedsides, our strategy does not need additional pretrained encoders as well as the dataset for pretraining, while in ProSmith model development, a highly correlated dataset (ligand-target affinity) was employed to pretrain the multimodal BERT encoders before employing it under the enzyme-substrate complex embedding scenario. Our model (CLR_ESP-XL with ESM-2-2560 and MoLFormer) has great potential to be used for future enzyme-substrate pair prediction to map the reliable relationship between enzymes and their substrates. The contrastive learning strategy proposed in this study exhibits great potential in another multimodal prediction task.

References:

Ahmad, W., Simon, E., Chithrananda, S., Grand, G., & Ramsundar, B. (2022). *ChemBERTa-2: Towards Chemical Foundation Models* (arXiv:2209.01712). arXiv. https://doi.org/10.48550/arXiv.2209.01712

Akdis, C. A. (2021). Does the epithelial barrier hypothesis explain the increase in allergy, autoimmunity and other chronic conditions? *Nature Reviews Immunology*, *21*(11), Article 11. https://doi.org/10.1038/s41577-021-00538-7

Bhat, S., Palepu, K., Yudistyra, V., Hong, L., Kavirayuni, V. S., Chen, T., Zhao, L., Wang, T., Vincoff, S., & Chatterjee, P. (2023). De Novo *Generation and Prioritization of Target-Binding Peptide Motifs from Sequence Alone*. https://doi.org/10.1101/2023.06.26.546591

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A Simple Framework for Contrastive Learning of Visual Representations* (arXiv:2002.05709). arXiv. https://doi.org/10.48550/arXiv.2002.05709

Copley, S. D. (2017). Shining a light on enzyme promiscuity. *Current Opinion in Structural Biology*, *47*, 167–175. https://doi.org/10.1016/j.sbi.2017.11.001

Du, Z., Ding, X., Xu, Y., & Li, Y. (2023). UniDL4BioPep: A universal deep learning architecture for binary classification in peptide bioactivity. *Briefings in Bioinformatics*, 1–10. https://doi.org/10.1093/bib/bbad135

Du, Z., Xu, Y., Liu, C., & Li, Y. (2023). pLM4Alg: Protein Language Model-Based Predictors for Allergenic Proteins and Peptides. *Journal of Agricultural and Food Chemistry*, acs.jafc.3c07143. https://doi.org/10.1021/acs.jafc.3c07143

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., & Valko, M. (2020). *Bootstrap your own latent: A new approach to self-supervised Learning* (arXiv:2006.07733). arXiv. https://doi.org/10.48550/arXiv.2006.07733

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). *Momentum Contrast for Unsupervised Visual Representation Learning* (arXiv:1911.05722). arXiv. https://doi.org/10.48550/arXiv.1911.05722

Kim, G. B., Kim, J. Y., Lee, J. A., Norsigian, C. J., Palsson, B. O., & Lee, S. Y. (2023). Functional annotation of enzyme-encoding genes using deep learning with transformer layers. *Nature Communications*, *14*(1), 7370. https://doi.org/10.1038/s41467-023-43216-z

Kroll, A., Ranjan, S., Engqvist, M. K. M., & Lercher, M. J. (2023). A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nature Communications*, *14*(1), 2787. https://doi.org/10.1038/s41467-023-38347-2

Kroll, A., Ranjan, S., & Lercher, M. J. (2023). *Drug-target interaction prediction using a multi-modal transformer network demonstrates high generalizability to unseen proteins* (p. 2023.08.21.554147). bioRxiv. https://doi.org/10.1101/2023.08.21.554147

Lee, I., Keum, J., & Nam, H. (2019). DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLOS Computational Biology*, *15*(6), e1007129. https://doi.org/10.1371/journal.pcbi.1007129

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lim, S., Lu, Y., Cho, C. Y., Sung, I., Kim, J., Kim, Y., Park, S., & Kim, S. (2021). A review on compound-protein interaction prediction methods: Data, format, representation and model. *Computational and Structural Biotechnology Journal*, *19*, 1541–1556. https://doi.org/10.1016/j.csbj.2021.03.004

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574

Mou, Z., Eakes, J., Cooper, C. J., Foster, C. M., Standaert, R. F., Podar, M., Doktycz, M. J., & Parks, J. M. (2021). Machine learning-based prediction of enzyme substrate scope: Application to bacterial nitrilases. *Proteins: Structure, Function, and Bioinformatics*, *89*(3), 336–347. https://doi.org/10.1002/prot.26019

Nguyen, N.-Q., Jang, G., Kim, H., & Kang, J. (2023). Perceiver CPI: A nested cross-attention network for compound–protein interaction prediction. *Bioinformatics*, *39*(1), btac731. https://doi.org/10.1093/bioinformatics/btac731

Nguyen, N.-Q., Park, S., Gim, M., & Kang, J. (2024). MulinforCPI: Enhancing precision of compound–protein interaction prediction through novel perspectives on multi-level information integration. *Briefings in Bioinformatics*, *25*(1), bbad484. https://doi.org/10.1093/bib/bbad484

Nobeli, I., Favia, A. D., & Thornton, J. M. (2009). Protein promiscuity and its implications for biotechnology. *Nature Biotechnology*, *27*(2), 157–167. https://doi.org/10.1038/nbt1519

Palepu, K., Ponnapati, M., Bhat, S., Tysinger, E., Stan, T., Brixi, G., Koseki, S. R. T., & Chatterjee, P. (2022). *Design of Peptide-Based Protein Degraders via Contrastive Deep Learning*. https://doi.org/10.1101/2022.05.23.493169

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision* (arXiv:2103.00020). arXiv. https://doi.org/10.48550/arXiv.2103.00020

Richemond, P. H., Grill, J.-B., Altché, F., Tallec, C., Strub, F., Brock, A., Smith, S., De, S., Pascanu, R., Piot, B., & Valko, M. (2020). *BYOL works even without batch statistics* (arXiv:2010.10241). arXiv. https://doi.org/10.48550/arXiv.2010.10241

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. *Proceedings of the National Academy of Sciences*, *118*(15), e2016239118. https://doi.org/10.1073/pnas.2016239118

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., & Das, P. (2022). Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, *4*(12), Article 12. https://doi.org/10.1038/s42256-022-00580-7

Song, N., Dong, R., Pu, Y., Wang, E., Xu, J., & Guo, F. (2023). PMF-CPI: Assessing drug selectivity with a pretrained multi-functional model for compound–protein interactions. *Journal of Cheminformatics*, *15*(1), 97. https://doi.org/10.1186/s13321-023-00767-z

The UniProt Consortium. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Thumuluri, V., Martiny, H.-M., Almagro Armenteros, J. J., Salomon, J., Nielsen, H., & Johansen, A. R. (2022). NetSolP: Predicting protein solubility in *Escherichia coli* using language models. *Bioinformatics*, *38*(4), 941–946. https://doi.org/10.1093/bioinformatics/btab801

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wang, X., Liu, J., Zhang, C., & Wang, S. (2022). SSGraphCPI: A Novel Model for Predicting Compound-Protein Interactions Based on Deep Learning. *International Journal of Molecular Sciences*, *23*(7), Article 7. https://doi.org/10.3390/ijms23073780

Yang, M., Fehl, C., Lees, K. V., Lim, E.-K., Offen, W. A., Davies, G. J., Bowles, D. J., Davidson, M. G., Roberts, S. J., & Davis, B. G. (2018). Functional and informatics analysis enables glycosyltransferase activity prediction. *Nature Chemical Biology*, *14*(12), 1109–1117. https://doi.org/10.1038/s41589-018-0154-9

Yu, T., Cui, H., Li, J. C., Luo, Y., Jiang, G., & Zhao, H. (2023). Enzyme function prediction using contrastive learning. *Science*, *379*(6639), 1358–1363. https://doi.org/10.1126/science.adf2465

Zhang, D., Jia, C., Sun, D., Gao, C., Fu, D., Cai, P., & Hu, Q.-N. (2023). Data-Driven Prediction of Molecular Biotransformations in Food Fermentation. *Journal of Agricultural and Food Chemistry*, *71*(22), 8488–8496. https://doi.org/10.1021/acs.jafc.3c01172

Zhang, D., Xing, H., Liu, D., Han, M., Cai, P., Lin, H., Tian, Y., Guo, Y., Sun, B., Le, Y., Tian, Y., Wu, A., & Hu, Q.-N. (2024). Discovery of Toxin-Degrading Enzymes with Positive Unlabeled Deep Learning. *ACS Catalysis*, 3336–3348. https://doi.org/10.1021/acscatal.3c04461

Zhang, Q., Ding, K., Lyv, T., Wang, X., Yin, Q., Zhang, Y., Yu, J., Wang, Y., Li, X., Xiang, Z., Zhuang, X., Wang, Z., Qin, M., Zhang, M., Zhang, J., Cui, J., Xu, R., Chen, H., Fan, X., … Chen, H. (2024). *Scientific Large Language Models: A Survey on Biological & Chemical Domains* (arXiv:2401.14656). arXiv. https://doi.org/10.48550/arXiv.2401.14656

Table 1 Performance of our model (CLR_ESP-exp) and the previous models trained on the experimental evidence-based dataset only

Model strategy	Encoder for enzymes	Encoder for molecules	Dataset (train, validation, test)	ACC	AUC	MCC	Additional notes
Contrastive learning strategy	ESM-2-480	MoLFormer	50077, 5422, 13336	92.21%	0.9468	0.7945	Our model
	ESM-2-1280	MoLFormer	50077, 5422, 13336	92.94%	0.9558	0.8146	Our model
	ESM-2-2560	MoLFormer	49881, 5422, 13336	**93.57%**	**0.9594**	**0.8314**	Our model
Simple concatenation strategy	ESM-2-480	MoLFormer	50077, 5422, 13336	86.42%	0.9078	0.6420	Our baseline
	ESM-2-1280	MoLFormer	50077, 5422, 13336	88.29%	0.9255	0.6928	Our baseline
	ESM-2-2560	MoLFormer	49881, 5422, 13336	89.64%	0.9357	0.7272	Our baseline
Simple concatenation strategy*	ESM-1b	GNN	50093, 5422, 13336	88.8%	0.94	0.72	ESP (Kroll, Ranjan, Engqvist, et al., 2023)
	ESM-1b$_{ts}$	GNN	50093, 5422, 13336	91.5%	0.956	0.78	ESP (Kroll, Ranjan, Engqvist, et al., 2023)

Note:

Original training, validation, and test dataset sizes are 50093, 5422, and 13336.

GNN: Graph neural networks.

The GNN was pretrained on predicting the Michaelis constants K_M of enzyme-substrate pairs.

*: this result is from previous model: ESP by Kroll et al.

ESM-1b$_{ts}$: It is a task specific fine-tuned ESM-1b model, where ESM-1b model was fine-tuned on 200634 enzyme substrate pairs with phylogenetically inferred evidence

Table 2 Ablation study on ESM-2-2560 and MoLFormer based model

Bottleneck size	First dense layer	ReLU	BatchNorm	L2 normalization	Loss function	ACC	AUC	MCC
32	✓	✓	✓	✓	MSE	93.09%	0.9541	0.8192
64	✓	✓	✓	✓	MSE	93.02%	0.9563	0.8177
128	✓	✓	✓	✓	MSE	**93.57%**	**0.9594**	**0.8314**
256	✓	✓	✓	✓	MSE	93.38%	0.9574	0.8268
512	✓	✓	✓	✓	MSE	93.16%	0.9562	0.8218
128	×	✓	✓	✓	MSE	74.31%	0.8405	0.1310
128	✓	×	✓	✓	MSE	92.21%	0.9499	0.7939
128	✓	✓	×	✓	MSE	92.01%	0.9448	0.7901
128	✓	✓	✓	×	MSE	93.42%	0.9591	0.8278
128	✓	✓	×	×	MSE	91.97%	0.9447	0.7888
128	×	×	✓	✓	MSE	73.90%	0.6867	0.0856
128	✓	✓	✓	✓	CE	74.47%	0.8217	0.4467

Note:

MSE: mean square error; CE: cross entropy

To calculate the cross entropy, we employed the sigmoid function to get the probability based on similarity valued and then calculated the cross entropy loss between the probability and the label.

Table 3 Performance of our models (CLR_ESP-phylo and CLR_ESP-XL) and previous model trained on both experimental evidence-based and phylogenetic evidence-based datasets

Encoder for enzymes	Encoder for molecules	ACC	AUC	MCC	Additional notes
ESM-2-2560	MoLFormer	93.98%	0.9567	0.8419	Our model*
ESM-2-2560	MoLFormer	**94.77%**	0.9653	**0.8628**	Our model**
ESM-2-1280	MoLFormer	93.93%	0.9534	0.8404	Our model*
ESM-2-1280	MoLFormer	94.56%	0.9635	0.8572	Our model**
ESM-1b	ChemBERTa-2	93.88%	0.9517	0.8391	Our model*
ESM-1b	ChemBERTa-2	94.3%	0.9618	0.8519	Our model**
ESM-1b	ChemBERTa-2	94.2%	**0.972**	0.85	ProSmith (Kroll, Ranjan, & Lercher, 2023)

Note:

* : the model was trained on the phylogenetic evidence-based dataset only;

** : the model was trained on the phylogenetic evidence-based dataset first and continued to be trained on the experimental evidence-based dataset

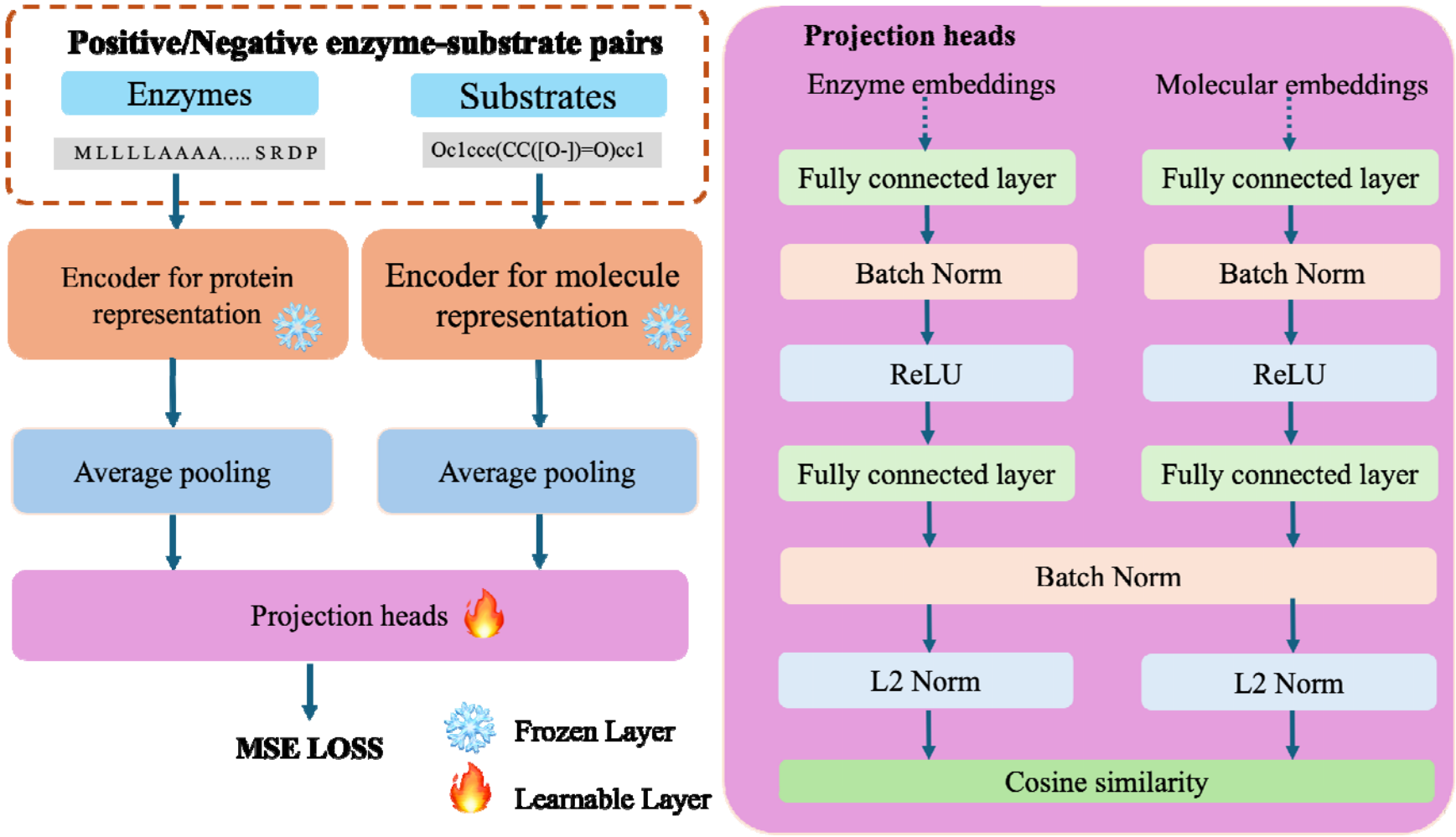


Fig. 1 Model architecture for enzyme-substrate pair prediction model

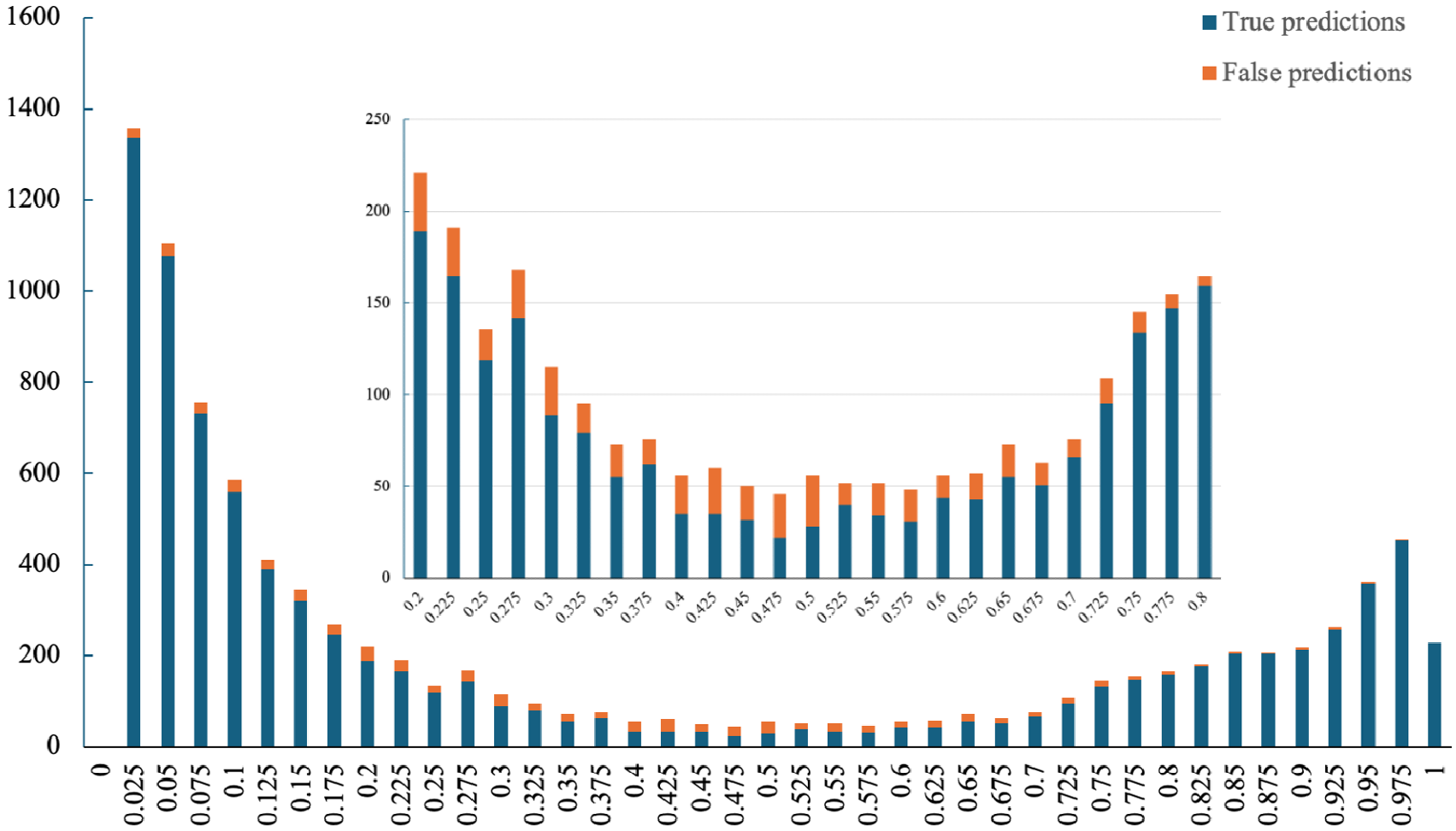


Fig. 2 Prediction scores around 0.5 indicate model uncertainty. Stacked histogram bars display the prediction score distributions of true predictions (blue) and false predictions (red). The inset shows a blow-up of the interval [0.2, 0.8]. Source data are provided as a Source Data file.

Supplementary documents

Table S1 Selected pretrained models

Pretrained model name	Number of layers	Number of parameters	Output dimensions	Dataset
ESM-2-2560	36	3 billion	2560	UR50/D 2021_04
ESM-2-1280	33	650 million	1280	UR50/D 2021_04
ESM-2-480	12	35 million	480	UR50/D 2021_04
ESM-1b	33	650 million	1280	UR50/S 2018_03

Table S2 The performance of CLR_ESP-XL in unseen enzymes

Maximal sequence identity	ACC	AUC	MCC
CLR_ESP-XL with ESM-2-2560 & MolFormer			
0-40%	93.07%	0.9476	0.8185
40-60%	96.12%	0.9752	0.9001
60-80%	96.85%	0.9871	0.9198
CLR_ESP-XL with ESM-2-1280 & MolFormer			
0-40%	91.46%	0.9334	0.7752
40-60%	95.49%	0.9731	0.8842
60-80%	96.31%	0.9864	0.9066
ESP			
0-40%	89%	0.93	0.72
40-60%	93%	0.97	0.83
60-80%	95%	0.99	0.88

Note: The performance of ProSmith was only exhibited in figure instead of table, and thus we did not have the exact values.

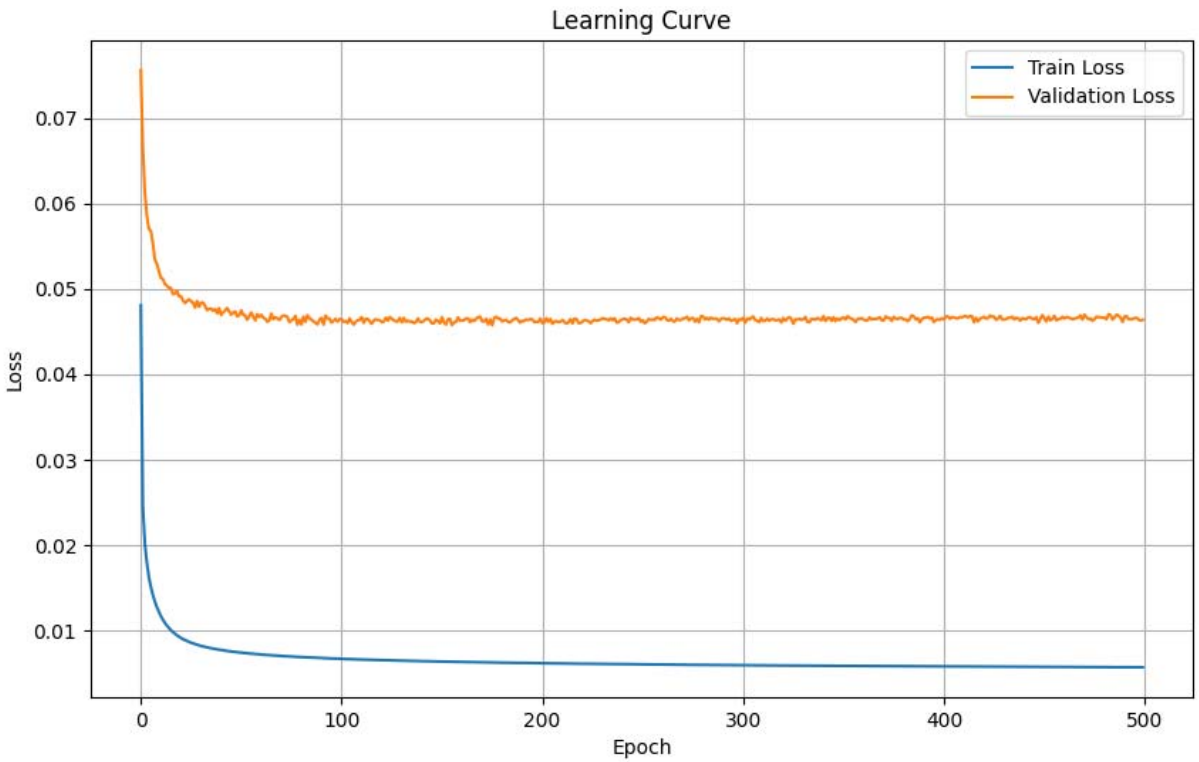
Table S3 The performance of CLR_ESP-XL in unseen small molecules

Number of positive samples with identical metabolite in the training set	Size of Subset	ACC	AUC	MCC
CLR_ESP-XL with ESM-2-2560 & MolFormer				
0	640	80.93%	0.6942	0.2966
1	498	92.57%	0.9445	0.7654
2	431	93.27%	0.8935	0.7880
3	586	96.07%	0.9657	0.8758
4	351	95.44%	0.9634	0.8631
5	414	93.48%	0.9486	0.7976
6	278	94.96%	0.9597	0.8547
7	313	94.57%	0.9731	0.8460
8	222	92.79%	0.9724	0.7938
9	317	94.32%	0.9575	0.8471
10	157	96.18%	0.9900	0.8954
>10	8876	95.86%	0.9783	0.8995
CLR_ESP-XL with ESM-2-1280 & MolFormer				
0	640	76.72%	0.6637	0.1343

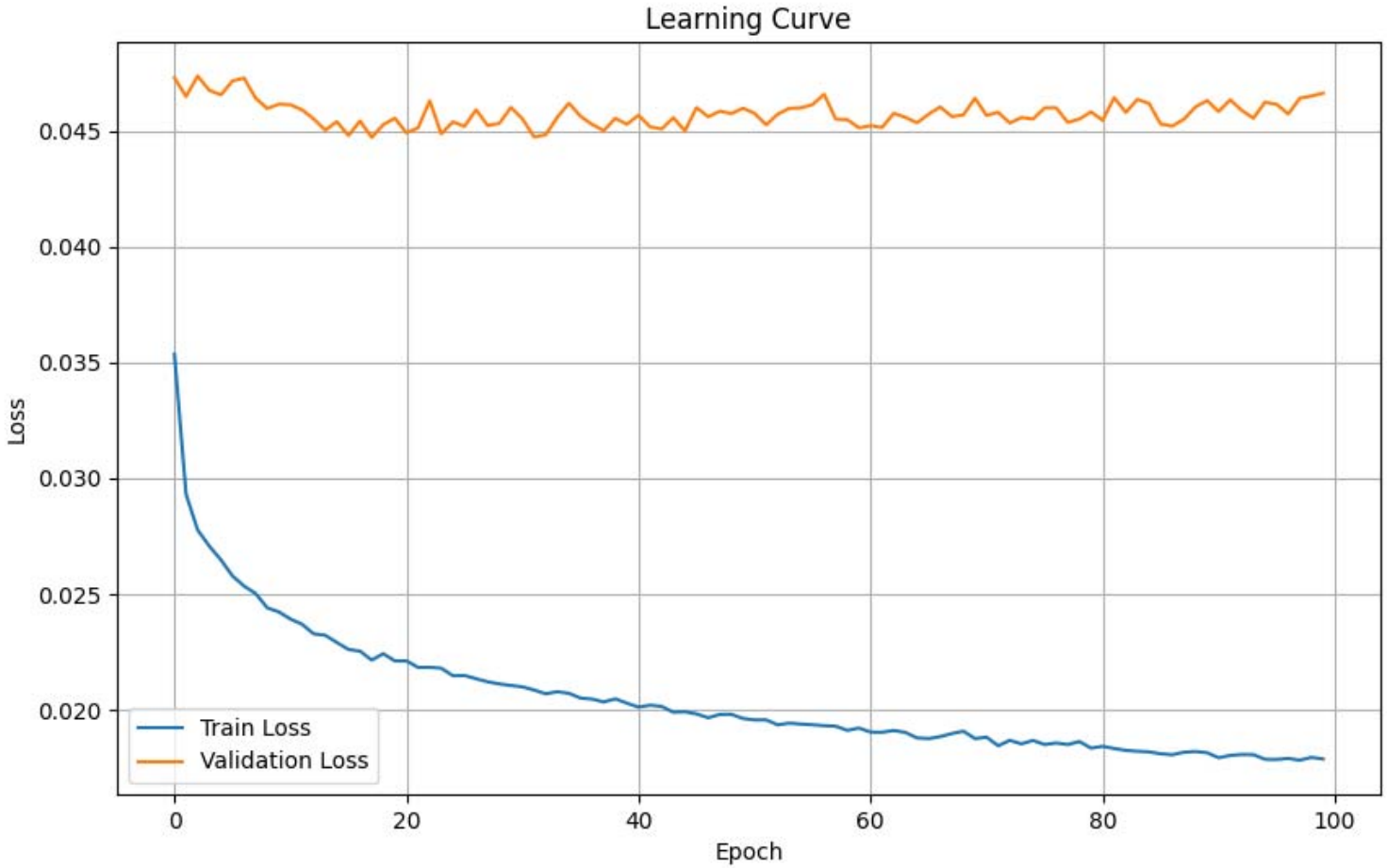
1	498	87.95%	0.8870	0.6198
2	431	88.86%	0.8738	0.6503
3	586	93.86%	0.9584	0.8080
4	351	92.02%	0.9397	0.7660
5	414	92.03%	0.9432	0.7504
6	278	92.45%	0.9580	0.7833
7	313	94.89%	0.9625	0.8543
8	222	91.44%	0.9432	0.7512
9	317	94.01%	0.9704	0.8371
10	157	95.54%	0.9737	0.8775
>10	8876	95.30%	0.9733	0.8855

Figure S1 Learning curves of the model in the phylogenetic evidence-based dataset (a) and the experimental evidence-based dataset (b)

(a)


Learning Curve
Train Loss
Validation Loss
Loss
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
100
200
300
400
500
Epoch

(b)


Learning Curve
Loss
Epoch
Train Loss
Validation Loss
0.045
0.040
0.035
0.030
0.025
0.020
0
20
40
60
80
100