JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

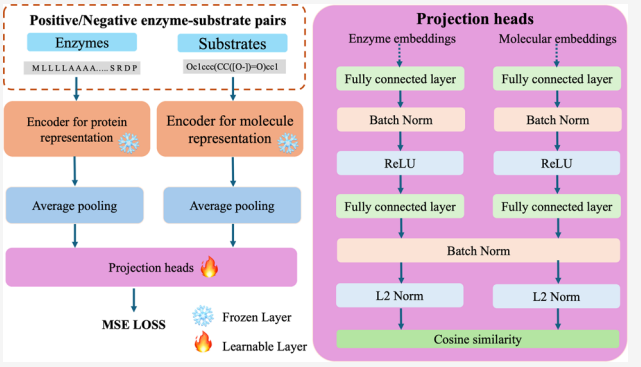
FusionESP: Improved Enzyme−Substrate Pair Prediction by Fusing Protein and Chemical Knowledge

Zhenjiao Du, Weimin Fu, Xiaolong Guo, Doina Caragea, and Yonghui Li*

Cite This: https://doi.org/10.1021/acs.jcim.4c02357 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: To reduce the cost of the experimental characterization of the potential substrates for enzymes, machine learning prediction models 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 by using available representation methods. Here, we present FusionESP, a multimodal architecture that integrates protein and chemistry language models with two independent projection heads and a contrastive learning strategy for predicting enzyme−substrate pairs. Our best model achieved state-of-the-art performance with an accuracy of 94.77% on independent test data and exhibited better generalization capacity while requiring fewer computational resources and training data, compared to previous studies of a fine-tuned encoder or employing more encoders. It also confirmed our hypothesis that embeddings of positive pairs are closer to each other in a high-dimension space, while negative pairs exhibit the opposite trend. Our ablation studies showed that the projection heads played a crucial role in performance enhancement, while the contrastive learning strategy further improved the projection heads' capacity in classification tasks. 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 FusionESP is established and freely accessible at https://rqkjkgpsyu.us-east-1.awsapprunner.com/.



1. 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.[1,2] Moreover, enzymes typically exhibit promiscuity, facilitating multiple reactions that may include physiologically irrelevant or potentially harmful processes.[3,4] A comprehensive mapping of enzyme−substrate relationships will provide crucial guidance for future research in medicine, pharmaceuticals, bioengineering, and agriculture.[5−7] 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.[8] There is a pressing need for high-throughput methods to address the scarcity of experimental validation in enzyme−substrate relationships.

Machine learning approaches have shown promising performance in various compound−protein interaction (CPI) prediction tasks.[9−20] As a subset of CPI prediction, enzyme−substrate pair prediction task is also a multimodal prediction task, where two different "language" systems [e.g., amino acid sequences of proteins and simplified molecular-input line-entry system (SMILES) of small molecules] are involved.[9,14] The first challenge lies in effectively representing small molecules and protein sequences as numerical vectors. Recent studies have employed advanced neural network-based approaches for automate representation learning, eliminating the need for manual feature selection.[9,16,21−23] These methods can be broadly categorized into sequence-based and graph-based machine learning approaches.

Sequence-based approaches represent proteins and small molecules in a 1D text format (e.g., SMILES for molecules and single-letter amino acid codes for proteins) by leveraging natural language processing techniques. For instance, methods like DeepDTA,[11] GraphDTA,[13] CSI,[10] and Perceiver CPI[18] utilized 1D convolutional neural networks (1D-CNNs) to

Received: December 17, 2024
Revised: February 25, 2025
Accepted: February 26, 2025

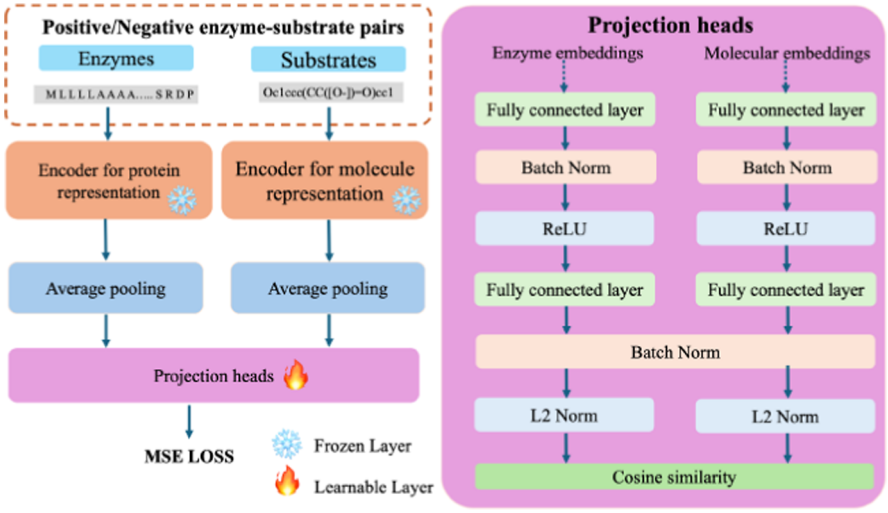


Figure 1. Model architecture for the enzyme−substrate pair prediction model.

extract features directly from protein sequences. Graph-based approaches, on the other hand, model the structural properties of molecules and proteins as graphs. Specifically, atoms or amino acid residues serve as nodes, while bonds or spatial interactions define the edges. Studies such as GraphDTA,[13] PMF-CPI,[19] CSI,[10] and Perceiver CPI[18] leveraged graph neural networks (GNNs) to learn molecular representations through 2D structures.

With the introduction of transformer architecture,[24] various pretrained large language models (LLMs) for protein and molecular representation have been released.[12,25−30] Pretrained on billions of protein sequences and small molecules, these LLMs have gained popularity for remarkable performance in representation learning and generating across various downstream classification and regression tasks through transfer learning or fine-tuning.[2,5,12,15,25,27,29−33] In the study of Kroll et al., a promising performance (accuracy = 91.5%) was achieved for enzyme−substrate pair prediction task by employing a task-specific fine-tuned protein language model (PLM) (i.e., ESM-1b) and a pretrained domain-specific GNN.[2] Subsequently, Kroll et al. introduced a multimodal BERT model for embedding a protein-molecule complex, where the protein sequences and SMILES were input into a multimodal BERT model. They achieved SOTA performance across four data sets, including drug−target interactions, protein−small molecule interactions, enzyme−substrate Michaelis constants (K_M), and substrate identification for enzymes, by concatenating PLM-generated protein embeddings, chemical language model (CLM)-generated molecule embeddings, and multimodal encoder-generated protein−molecule embeddings.[5]

Effectively leveraging comprehensive enzyme and molecular embeddings generated by their respective encoders is critical for enhancing model performance. The most popular strategy is to simply concatenate protein and molecule embeddings, a method widely adopted and modified in the CPI discovery community.[11,13,17,20,34] Beyond that, in the study of Song et al., Kronecker products of protein and molecule embeddings as additional features were concatenated with the original embeddings to enhance CPI prediction performance.[19] Attention mechanisms have been employed to integrate multiview embeddings of drug molecules and proteins for better representation learning.[17,18] Specifically, Perceiver CPI sequentially utilized a cross-attention block, a self-attention block, and another cross-attention block, to integrate multiple views of the molecules, refine molecular representations, and capture the semantic relevance between proteins and molecules.[18] Besides, MGNDTI employed gated linear units (GLUs) to filter nonimportant features from three encoders and an extra mean square error (MSE) loss function to enhance the representation learning from molecular graphs and SMILES sequences, as well as element-wise manipulation for knowledge fusion before concatenation,[15] while MMCL-CPI adopted 2D CNN blocks to extract a fused representation from two stacked embedding matrixes.[35] To maximize the prediction performance, these models employed either more encoders to enrich embedding protein and molecule or fused embeddings to enhance enzyme−substrate complex representations. However, these methods often require either complex fusion architecture design and substantial computational resources or additional data sets for fine-tuning, which impeded the deployment and application to different scenarios.

Inspired by the success of contrastive language-image pretraining (CLIP), where two independent encoders for images and texts were jointly trained to predict 400 million correct image−text pairs,[36] we hypothesized that enzyme and substrate in a correct pair should be closer in a high dimensional space after projection, while unrelated components should be distinctly separated. This concept has been explored in both single modality settings (e.g., protein−peptide interaction)[37,38] and multimodality settings (e.g., CPI and drug−target interaction).[39−42] Recent studies, such as DrugCLIP[39] and ConPLex,[41] either followed the original CLIP architecture with a few modifications to fine-tune the protein encoder and the molecule encoder for better representation learning[39] or freeze the encoders and trained the model with projection heads.[41] The latter can save more computation resources compared to the consumption of large encoder fine-tuning in the former. This technical route was also found in the study of Yu et al. (2023), where a project head was designed to refine embeddings from ESM-1b for a single modality task of enzyme commission (EC) number prediction and achieved state-of-the-art (SOTA) performance across various benchmark data sets, particularly excelling in rare EC number predictions.[43]

Based on these advances, we proposed FusionESP, leveraging the concept of contrastive learning to tackle

https://doi.org/10.1021/acs.jcim.4c02357
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

multimodal prediction tasks in the enzyme substrate prediction task. Our objective of this work was to develop an advanced machine learning module tailored for empowering existing LLMs to predict enzyme−substrate relationships across a wide range of proteins and molecules. This module aims to serve as a practical tool to streamline experimental processes and boost laboratory efficiency. In this study, we designed a simplified yet highly effective model architecture that employs a contrastive learning strategy. The architecture demonstrated a remarkable performance enhancement in knowledge fusion within a multimodal context, even with limited data. Specifically, we utilized ESM-2 for enzyme embeddings and MolFormer for molecule embeddings. Rather than fine-tuning these encoders, adding additional encoders, or incorporating more features [e.g., extended-connectivity fingerprints (ECFPs)], we designed projection layers for both encoders to refine and align the embeddings in the same high-dimensional space (Figure 1). Our model outperforms previous approaches with a simplified architecture and reduced computational demands during both training and inference phases.

2. METHODS

2.1. Data Set. To ensure a fair comparison with previous studies, the data sets used in this study were sourced from existing studies.[2,5] Detailed data set construction approach is available in the study of Kroll et al.[2] In brief, the positive enzyme−substrate pairs were extracted from the Gene Ontology (GO) annotation database, where entries have different levels of evidence: experimental, phylogenetically inferred, computational analysis, author statement, curator statement, and electronic evidence. The data set construction was based on experimental evidence and phylogenetic evidence. For irreversible enzymatic reactions, only reactants explicitly identified as substrates were included. Substrates that could not be mapped to identifiers in KEGG, ChEBI, or PubChem were excluded. 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.[2] Specifically, the FingerprintsSimilarity function from the RDKit package was used for the pairwise similarity calculation among small molecules based on molecular fingerprints. Random sampling was conducted from the molecules with similarity scores ranging from 0.7 to 0.95. If eligible molecules were not enough, the lower bound was reduced in steps of 0.2 until enough small molecules could be sampled.

The experimental evidence-based data set, originally split into training, validation, and test sets in the original study,[2,5] contained 50,093 enzyme−substrate pairs in the training set, 5422 pairs in the validation set, and 13,336 pairs in the test set. In contrast, the phylogenetic evidence-based data set comprised a total of 765,635 pairs. These two data sets 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 the hardware capacity (NVIDIA A100 GPU). Specifically, for ESM-2 models with output dimensions of 480 and 1280, sequences longer than 8000 nm were removed. Sixteen pairs were removed from the training data set, and no pair was removed from the test data set. For the ESM-2 model with a 2560-dimensional output, sequences longer than 5500 were excluded. Correspondingly, 212 pairs were removed from the training data set. Notably, no sequence in the validation and test data sets was removed.

2.2. Calculating Enzyme Representation. In this study, enzymes were represented numerically using ESM2 models.[28,44] 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 a modified bidirectional encoder representation from transformers (BERT) architecture. Those PLMs were trained on large protein sequence data sets (e.g., UR50/D 2021_04) through self-supervised learning. In this study, we employed three ESM-2 models, including esm2_t12_35M_UR50D with 35 million parameters and 480-dimensional output, esm2_t33_650M_UR50D with 650 million parameters and 1280-dimensional output, and esm2_t36_3B_UR50D with 3 billion parameters and 2560-dimensional output, denoted as ESM-2-35M, ESM-2-650M, and ESM-2-3B, respectively. The details about each PLM are listed in Supporting Information Table S1. 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-2-650M model for the enzymes is a $(L + 1) \times 1280$ matrix, where 1280 represents the output dimension of the ESM-2-650M model. The first row is the representation of the [sos] token, which is the indicator of "start of sentence" during the training process. After that, each row in the matrix represented the corresponding amino acid from the N-terminus to the C-terminus 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 dimensions 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.

2.3. Calculating Molecule Representation. MoLFormer was selected for numerical representation generation of molecules.[29] MoLFormer was developed by IBM Research in 2022 and trained on canonical SMILES sequences of 1 billion molecules from the ZINC database and 111 million molecules from PubChem with employment of rotary positional embeddings and a linear attention mechanism. There was only one model available, called "MoLFormer-XL-both-10pct" trained with 10% of both data sets, at https://huggingface.co/ibm/MoLFormer-XL-both-10pct. Although the data set size was smaller, this model demonstrated performance comparable to the full-size model. Therefore, we used this model for molecule embeddings. We will use "MoLFormer" to refer to 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.

2.4. Encoding Models. ESM-2 models represent an upgrade over ESM-1b used in previous studies.[2,5] They were trained on a new data set (UR50/D 2021_04), while the ESM-1b model was trained on UR50/S 2018_03.[28,44] 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.[28,44] As for

https://doi.org/10.1021/acs.jcim.4c02357
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

MoLFormer, it was trained on a larger data set 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).[5] It exhibited better performance in one regression task (Lipophilicity data set) and three classification tasks (BACE, BBBP, and ClinTox data sets) during downstream task performance evaluation.[25,29]

2.5. Model Architecture Design. The model architecture comprised two independent modules with similar inner architecture design for enzyme sequence input and molecule SMILES input (Figure 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. Each enzyme−substrate pair was finally transformed into two 128 dimensional vectors, where one represented the enzymes and the other represented the molecule. Cosine similarity between the two refined embeddings was calculated finally as the interaction portability of the pair. The value of the positive enzyme−substrate pairs was 1, and the true value of the negative pair was 0. A MSE loss function was employed for the loss calculation as follows

$$\text{loss} = \frac{1}{n}\sum_{i=1}^{n}(\text{Sim}_i - \text{Label}_i)^2$$

where n is the number of data points; Sim_i is the cosine similarity of the 128-dimensional vectors of enzyme and substrate in the single pair; and Label_i is the true label.

The projection head mainly referred to the projection heads in contrastive learning studies with a few modifications.[45−47] The two projection heads for enzyme embeddings and molecular embeddings did not share weights and were trained simultaneously and independently through the MSE loss function, except for the second batch norm layer. While the contrastive learning idea was inspired from CLIP,[36] there was no huge data set 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 data set. At the same time, we employed the knowledge/expert-based negative data set 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.[45] 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-2-3B and MoLFormer, the first fully connected layers had 2560 and 768 neurons for enzyme and molecule embeddings, respectively. Following projection into a 128-dimensional vector, both enzyme and molecule vectors shared the same batch normalization layer, assuming that they were in the same high-dimensional space.

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

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

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

2.8. Web Server. A user-friendly web server was deployed with Amazon Web Services (AWS) app runner. The Web site was designed with html and css scripts, and the model deployment was achieved with Flask (2.2.2). Due to the constraint of computational resources, FusionESP-XL with ESM-2-650M and MolFormer was deployed for enzyme−substrate prediction. The web server supports large-scale processing, which allows users to upload their peptide information through xls or xlsx formats. The detailed scripts for web server development are available at https://github.com/dzjxzyd/FusionESP_server_1280.

3. RESULTS AND DISCUSSION

3.1. Contrastive Learning Strategy Exhibits Superiority Over Simple Concatenation Strategy. The data set was divided into training, validation, and test sets as in the original paper.[5] Three different ESM-2 models and MoLFormer were employed as encoders for enzyme and molecule representations, respectively. FusionESP-exp was exclusively trained on the provided experimental evidence-based data sets. Although our training data set was slightly smaller compared to those of previous models, all models, including ours, were evaluated on the same test data set, ensuring a fair final performance comparison.

https://doi.org/10.1021/acs.jcim.4c02357
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Performance of Our Model (FusionESP-exp) and the Previous Models Trained on the Experimental Evidence-Based Data Set Only[a]

model strategy	encoder for enzymes	encoder for molecules	data set (train, validation, test)	ACC (%)	AUC	MCC	additional notes
contrastive learning strategy	ESM-2-35M	MoLFormer	50077, 5422, 13336	92.21	0.9468	0.7945	our model
	ESM-2-650M	MoLFormer	50077, 5422, 13336	92.94	0.9558	0.8146	our model
	ESM-2-3B	MoLFormer	49881, 5422, 13336	**93.57**	**0.9594**	**0.8314**	our model
simple concatenation strategy	ESM-2-35M	MoLFormer	50077, 5422, 13336	86.42	0.9078	0.6420	our baseline
	ESM-2-650M	MoLFormer	50077, 5422, 13336	87.59	0.9188	0.6755	our baseline
	ESM-2-3B	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[1]
	ESM-1bts	GNN	50093, 5422, 13336	91.5	0.956	0.78	ESP[1]

[a]Note: original training, validation, and test data set 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. *Results were retrieved from previous studies by Kroll et al.[2] 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.

To assess the superiority of our contrastive learning strategy in enhancing model 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, the details of which can be found in our GitHub repository. The results from all 6 models are shown in Table 1. All of the models based on our contrastive learning strategy outperformed their counterparts, especially for FusionESP-exp with ESM-2-35M and MoLFormer, where the accuracy was 5.79% higher than that of the model based on the simple concatenation strategy with ESM-2-35M and MoLFormer for embeddings. This underscores the significant advantage of our proposed contrastive learning strategy over the widely used concatenation strategy. Among the baseline models, the embeddings of enzymes and small molecules were directly concatenated and processed through fully connected layers. Negative pairs were generated by randomly selecting three small molecules with a high degree of similarity to the true substrates of the same enzyme sequence. Consequently, for a given enzyme, the embeddings of small molecules in the negative pairs closely resemble those in the positive pairs. Since the enzyme embeddings remained identical across the four pairs, the only source of variation in the concatenated embeddings came from the small molecule representations. This design introduced a potential issue: the learnable two-layer neural network could be struggling with concatenated embeddings with large duplicate portions between positive and negative pairs, reducing its ability to effectively distinguish between them. However, our model architecture mitigated this issue by maintaining separate, independent projection heads for enzyme and small molecule embeddings until the final loss calculation. This decoupled structure ensured that each projection head focused exclusively on their respective inputs and learned distinct patterns specific to enzymes and small molecules rather than being influenced by their combined representations. Under the above circumstance, this design amplifies the differences between similar small molecules during backpropagation, enhancing the model's ability to differentiate them effectively.

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 for EC number prediction.[43] Frozen encoders for embeddings were also employed for DTI prediction in the study of ConPLex, where a PLM (ProtBERT) and Morgan fingerprint were used for protein and small molecule embeddings, respectively, while the learnable projection layers were shallower than that of ours.[41] Such a model architecture also exhibited promising performance among public data sets (e.g., DAVIS, BIOSNAP, and BindingDB) as well as unseen drugs and proteins.[41] In this study, we employed two independent projection heads and an MSE loss function referring to the CLIP model for multimodal prediction tasks. The promising results validated our hypothesis that positive enzyme−substrate pairs tend to be closer in the 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 data set size, making it well-suited for biological applications with limited data availability.

Additionally, a positive correlation between model performance and the size of the ESM models/output dimensions was observed in both the contrastive learning and concatenation strategy groups. This trend aligns with findings from our previous studies,[48] where larger model sizes/output dimensions encoded more information into the embeddings, resulting in improved overall performance.

3.2. Model Performance Comparison with SOTA Performance Trained Only on Experimental Evidence-Based Data Set. The FusionESP-exp models achieved superior performance compared to existing SOTA models, with accuracy, area under the curve, and Matthews correlation coefficient (MCC) ranging from 92.21% to 93.57%, 0.9468 to 0.9594, and 0.7945 to 0.8314, respectively (Table 1). In Kroll et al.'s study, models combining ESM-1b or fine-tuned ESM-1b with a pretrained GNN on the same data sets outperformed models using ESM-1b and traditional ECFPs for enzyme and substrate representation, which exhibited the superiority of pretrained model GNN over ECFP-based features.[2] The performance of ESM-1b and GNN model achieved 88.8% accuracy, which was close to the one (ACC = 87.59%) of the FusionESP-exp composed of ESM-2-650M and MoLFormer base models with simple concatenation strategy. The enzyme encoders in both models had the same model size (650 million parameters) and output dimension (1280 dimensions), with some modifications in ESM-2 regarding the training set and model architecture. 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 our application scenario (i.e.,

https://doi.org/10.1021/acs.jcim.4c02357
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Ablation Study on FusionESP-exp with ESM-2-3B and MoLFormer[a]

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
256	×	√	√	√	MSE	73.63	0.6517	N/A
512	×	√	√	√	MSE	73.63	0.6517	N/A
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

[a]Note: MSE: mean square error; CE: cross entropy; N/A: not available. *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.

enzyme−substrate pair prediction).[2] Besides, the pretrained GNN was still learnable during the training task on the enzyme−substrate pair data set, and additional hyperparameter optimization for the gradient-boosting classifier was used to enhance the performance. Those reasons together contributed to the slightly better performance of the ESM-1b and GNN-based model over our simple concatenation models (ESM-2-650M and MoLFormer), though we employed more advanced encoders for enzymes. When employing the contrastive learning strategy, FusionESP-exp with ESM-2-650M and MoLFormer achieved 92.94% accuracy, which was 4.14% higher than that of the model with ESM-1b and GNN. The second model from Kroll et al. employed a task specific fine-tuned ESM-1b model, which was fine-tuned on 200,634 enzyme−substrate pairs with phylogenetically inferred evidence.[2] The additional fine-tuning process enhanced the representation power of ESM-1b and increased the accuracy from 88.8% to 91.5%. However, compared to our models with contrastive learning strategy, the performance was inferior to ours by approximately 0.78−2.26%, where no additional data set, ESM-1b fine-tuning, and GNN fine-tuning were needed, highlighting the effectiveness of our proposed architecture for enzyme−substrate prediction tasks.

3.3. Ablation Study on the Best-Performing FusionESP Model. To further explore the function of each element in the projection head, we conducted an ablation study on the best-performing FusionESP-exp model using ESM-2-3B and MoLFormer with an accuracy of 93.57% (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. The results, summarized in Table 2, indicated that the removal of the first dense layer or the replacement of MSE loss function with the entropy-based loss function led to a substantial decline in performance. Notably, given the data set's 1:3 ratio of positive to negative pairs, models lacking these components exhibited severe training instability, rendering them effectively untrainable. Additionally, the bottleneck size and batch normalization layers contributed marginally to performance improvement, whereas the L2 normalization layer had a negligible impact.

One of the most notable findings was the importance of the first dense layer, which functioned as an identity mapping layer, maintaining the same input−output dimensionality for enzyme and small molecule embeddings. This architectural choice, originally inspired by SimCLR, is not a simple identity mapping but rather serves as a mechanism for feature reweighting, representation learning, and nonlinearity introduction, facilitating a smooth transition before compression in the subsequent layer.[45] Our observation indicated that, unless the first dense layer was removed, the model can still learned from the data set, though the performance remained suboptimal, while the removal of nonlinearity had moderate impact. Although the second dense layer remains learnable, it cannot compensate the absence of the first dense layer, even with modest feature compression, and the absence of the first dense layer substantially impairs the model's ability to effectively adapt to the designated downstream task, highlighting its importance in structuring the contrastive representation space.

The performance based on cross entropy is in expectation. In the original CLIP architecture, the key idea of contrastive learning is to maximize the similarity between correct image−text pairs and minimize the similarity between incorrect pairs, where the similarity is a continuous value instead of a discrete label/category.[36] Therefore, the introduction of cross entropy for loss calculation cannot capture the loss very well for model training and caused poor performance.

3.4. Ablation Study on the Best-Performing FusionESP Model without Contrastive Learning Strategy. In the previous section, we systematically examined the role of each element within the projection head. To further assess the contribution of the contrastive learning strategy to model performance, we investigated whether a concatenation-based approach, employing a similar projection module, could achieve comparable results. To address this, we designed two concatenation-based strategies derived from FusionESP. In the first strategy, embeddings from enzyme (2560-dimensional) and small molecule (768-dimensional) language models were directly concatenated and fed into a single projection head identical to that in FusionESP. The output of the projection head was then passed through a sigmoid-activated output layer for cross-entropy loss calculation. In the second strategy, refined embeddings from two independent projection heads in FusionESP were concatenated before being fed into either an output layer or an additional dense layer (third layer)

https://doi.org/10.1021/acs.jcim.4c02357
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 3. Ablation Study on FusionESP-exp (ESM-2-3B and MoLFormer) without Contrastive Learning[a]

bottleneck size	first dense layer size	third dense layer	ReLU	BatchNorm	L2 normalization	loss function	ACC (%)	AUC	MCC
First Concatenation Strategy									
128	×	×	√	√	√	CE	91.57	0.9546	0.7849
128	3328	×	√	√	√	CE	92.23	0.9602	0.7989
128	256	×	√	√	√	CE	92.09	0.9583	0.7965
128	128	×	√	√	√	CE	92.05	0.9568	0.7955
128	128	128	√	√	√	CE	91.72	0.9585	0.7860
Second Concatenation Strategy									
√	√	×	√	√	√	CE	73.84	0.5697	0.0757
√	√	32	√	√	√	CE	91.01	0.9494	0.7690
√	√	128	√	√	√	CE	90.50	0.9487	0.7588
√	√	256	√	√	√	CE	90.98	0.9483	0.7689

[a]Note: CE: cross entropy. First concatenation strategy: concatenate the embeddings from the protein language model and the chemical language model and load the concatenated embeddings into a single projection head and an output layer subsequently. Second concatenation strategy: concatenate the two refined embeddings from two independent projection heads for an output layer or third dense layer before the output layer.

Table 4. Performance of Our Models (FusionESP-phylo and FusionESP-XL) and Previous Model Trained on Both Phylogenetic Evidence-Based and Experimental Evidence-Based Data Sets

encoder for enzymes	encoder for molecules	ACC (%)	AUC	MCC	additional notes
ESM-2-3B	MoLFormer	93.98	0.9567	0.8419	our model (fusionESP-phylo)[a]
ESM-2-3B	MoLFormer	**94.77**	0.9653	**0.8628**	our model (fusionESP-XL)[b]
ESM-2-650M	MoLFormer	93.93	0.9534	0.8404	our model (fusionESP-phylo)[a]
ESM-2-650M	MoLFormer	94.56	0.9635	0.8572	our model (fusionESP-XL)[b]
ESM-1b	chemBERTa-2	93.88	0.9517	0.8391	our model (fusionESP-phylo)[a]
ESM-1b	chemBERTa-2	94.3	0.9618	0.8519	our model (fusionESP-XL)[b]
ESM-1b	chemBERTa-2	94.2	**0.972**	0.85	ProSmith[2]

[a]Note: the model (FusionESP-phylo) was trained on the phylogenetic evidence-based data set only. [b]The model (FusionESP-XL) was trained on the phylogenetic evidence-based data set first and continued to be trained on the experimental evidence-based data set, corresponding model.

preceding the output layer, followed by a sigmoid activation function for the loss computation.

The results, summarized in Table 3, reveal that the projection module from SimCLR remains effective, even outside the contrastive learning context. Specifically, in the first strategy, concatenating the embeddings before projection yielded notably better performance than concatenating them after the two learnable projection heads. Additionally, varying the output dimension of the first dense layer had only a modest impact. The original identity mapping design in the first dense layer (as used in FusionESP's projection head) achieved the best performance. Removing the first dense layer or introducing a third dense layer led to performance degradation, while in the contrastive learning setting, the model lost the ability to learn from data. For the second strategy, directly concatenating the refined 128-dimensional embeddings from the two independent projection heads before the output layer severely impaired the model's learning ability. However, introducing a third dense layer as an adapter significantly improved the performance, and its output dimension had minimal impact on the final performance.

In summary, the projection head design plays a crucial role in enhancing the prediction performance compared to a simple multilayer perceptron classifier, as shown in Table 1. Moreover, the contrastive learning strategy further improved the model's predictive capability.

3.5. Training the Model with Including Phylogenetic Evidence-Based Data Further Enhances Its Performance. In this section, we further enhanced model performance by training on a larger data set combining phylogenetic evidence-based and experimental evidence-based data sets. Assuming that the phylogenetic evidence-based data set contained rich relationship information between enzymes and substrates, we directly trained models on this data set for 500 epochs, and subsequently, the model underwent further training on the experimental evidence-based data set for 30 epochs to optimize performance. Following the approach from a reference (ref 5, we trained and evaluated the model performance on the same data sets for fair performance comparison. The results are shown in Table 4.

Notably, all the models in Table 4 trained on phylogenetic evidence-based data set only (FusionESP-phylo) outperformed all the models trained on experimental evidence-based data set from Table 1 (FusionESP-exp), including models employing ESM-1b and ChemBERTa-2. The phylogenetic evidence-based data set contained 764,449 data points, which was around 15 times larger than the experimental evidence-based data set. We can infer that the number of learnable parameters was not yet saturated under current data sets and can be expected to further enhance performance with a larger data set.

Similarly, larger encoder-based models exhibited better performance, with FusionESP-phylo and FusionESP-XL using ESM-2-3B, achieving accuracy of 93.98% and 94.77%, respectively, compared to 93.93% and 94.56% from models using ESM-2-650M. Besides, improvement was also observed after the models were further trained on the experimental evidence-based data set. This is not surprising as the data points in the training data set from the experimental evidence-based data set were much closer to the data points in the test data set, thus allowing the model to perform better in the test data set. There was also a performance improvement between the models with ESM-2-650M and MoLFormer and those with

https://doi.org/10.1021/acs.jcim.4c02357
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

ESM-1b and ChemBERTa-2 under the same contrastive learning strategy. It mainly resulted from the two encoders. ESM-2-650M and MoLFormer were trained with larger or updated data sets 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.[25,28,29] It is worth noting that ESM-1b can embed enzyme sequences shorter than 1024 amino acids or truncated enzyme sequences with the first 1023 amino acids, which caused some information loss and inferior performance.

3.6. Projection Heads Outperform an Additional Multimodal BERT. Compared to the previous SOTA model (ProSmith designed by Kroll et al.), our model trained and evaluated on the same data sets achieved superior performance[5] (Table 4). 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 the enzyme sequence and SMILES. The BERT model was pretrained on 1,039,565 data points from ligand target affinity data set.[5] Therefore, ProSmith employed three encoders for embeddings and connected them with an optimized gradient boost model for prediction, achieving an accuracy of 94.2%.[5] In contrast, our model (FusionESP-XL) achieved 94.77% accuracy using only two pretrained encoders without additional pretraining of a new multimodal BERT model. With our contrastive learning strategy, FusionESP-XL with ESM-1b and ChemBERTa-2 employed only two encoders, same as that of ProtSmith without the multimodal BERT, and achieved slightly higher accuracy.

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. Our promising performance was attributed to our simple projection heads, effectively fusing embeddings from two encoders for prediction tasks, marginally superior to the effect of an additional multimodal BERT encoder and a well-optimized 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.[44] Therefore, those enzyme−substrate pairs whose sequence length was longer than 1023 were removed. Though our training data set was smaller than the one used in ProSmith, our performance was still comparable. Such a comparison further demonstrated the superiority of our proposed architecture.

3.7. Evaluating FusionESP-XL Performance in Unseen Enzymes and Small Molecules. In order to further characterize the model performance in rarely seen enzymes, we split the test data set into three subgroups: data points 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.[49] The results are presented in Table 5. It was observed that FusionESP-XL model with ESM-2-3B and MolFormer achieved higher performance in enzyme−substrate pairs with higher sequence identity. Specifically, the accuracy, AUC, and MCC were 96.85%, 0.9871, and 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 by approximately 1.95−4.57% in accuracy.[2] Though trained on the same experimental evidence-based data set, FusionESP-exp with ESM-2-3B and MolFormer still outperformed the ESP model across all three sub-data sets. When compared with ProtSmith, our model (FusionESP-XL with ESM-2-3B and MolFormer) outperformed in the most challenging sub-data set (0−40% maximal sequence identity), with the MCC increasing from 0.78 to 0.8185. This model also surpassed ProtSmith in the 40−60% maximal sequence identity sub-data set in terms of MCC, while the MCC was slightly lower than ProtSmith in the 60−80% maximal sequence identity sub-data set.

Table 5. Performance of FusionESP-XL and FusionESP-exp with ESM-2-3B and MolFormer in Rarely Seen and Unseen Enzymes[a]

maximal sequence identity	ACC (%)	AUC	MCC
FusionESP-XL with ESM-2-3B & 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
FusionESP-exp with ESM-2-3B and 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

[a]Note: the performance of ProSmith was only exhibited in figure instead of table, and thus we did not have the exact values. *The performance of ESP model was based on the task-specific fine-tuned ESM-1b model.

Furthermore, we investigated the predictive capabilities of our models for small molecules with different frequencies in training data set (Table 6). Similar to that in maximal sequence identity, FusionESP-XL achieved better performance compared to FusionESP-exp across the 12 sub-data sets. Besides, the rarely seen small molecules tend to be wrongly predicted by the model, especially in unseen small molecules with an ACC of 80.93%. Compared to the performance in unseen small molecules in ProtSmith (MCC = 0.29) and ESP (MCC = 0.00), our model (FusionESP-XL with ESM-2-3B and MolFormer) achieved slightly higher performance with a MCC of 0.2966. When it came to rarely seen small molecules (only present once in training data set), our model's performance was increased from 80.93 to 92.57%. Compared to ESP (MCC = 0.28) and ProtSmith (MCC = 0.69), our model achieved much higher performance (MCC = 0.7654).

The performance of our model in rarely seen and unseen enzymes and small molecules indicated that our model has better generalization capability. The projection heads can not only contribute to the increase in overall performance but also enable the model to perform well in unseen data points.

To further explore the model performance under different maximal sequence identity and frequency of small molecules in training data set, we further explore the model's performance under different frequencies of small molecules in each sequence similarity subgroup (Figure 2). It was observed

https://doi.org/10.1021/acs.jcim.4c02357
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 6. Performance of FusionESP-XL and FusionESP-exp with ESM-2-3B and MolFormer in Rarely Seen and Unseen Small Molecules

number of positive samples with identical metabolite in the training set	size of Subset	ACC (%)	AUC	MCC
FusionESP-XL with ESM-2-3B and 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
FusionESP-exp with ESM-2-3B and 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

that within the same maximal sequence identity subgroup, performance improved as the frequency of small molecules in the training data set increased. For example, in the 0−40% maximal sequence identity subgroup, the accuracy increased from 76.13% for unseen small molecules to 90.72% for those with a frequency ranging from 1 to 10 and further to 94.26% for frequencies greater than 10. Similarly, the prediction performance for unseen and rarely seen small molecules also improved as the maximal sequence identity increased. Particularly, the accuracy for those unseen small molecules was increased from 76.13% with 0−40% maximal sequence identity to 85.09% for enzymes with 40−60% maximal sequence identity and 93.07% for enzymes with 60−80% maximal sequence identity. Predicting enzyme−substrate pairs was particularly challenging when the enzyme exhibited 0−40% sequence identity and the small molecule was unseen. However, when the enzyme had a higher maximal sequence identity or the small molecules were present in the training data set, the model was able to make significantly more reliable predictions.

3.8. FusionESP-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 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, it is a negative pair. The cosine similarity score can be also provided as output at our Web server at https://rqkjkgpsyu.us-east-1.awsapprunner.com/ for single reaction prediction or large-scale prediction.

To provide a more detailed assessment of prediction accuracies, Figure 3 displays the distributions of true (blue) and false (orange) predictions within our test data set across various prediction scores. Most correct predictions had scores either close to 0 or close to 1, indicating that FusionESP-XL made predictions with high confidence. In contrast, false predictions were distributed more evenly across the prediction score range. The predictions for the data points with scores

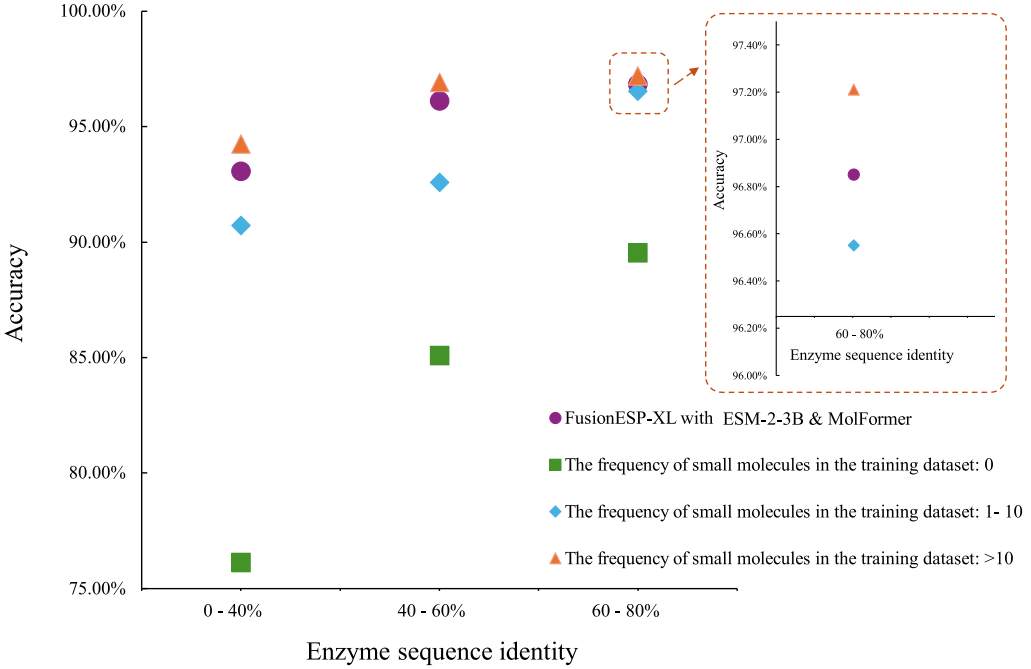


Figure 2. Prediction performance of FusionESP-XL with ESM-2-3B and MolFormer. Note: we divided the test data set into subsets with different levels of enzyme sequence identity and frequency of small molecules in the training data set. Source data are provided as a Source Data file.

https://doi.org/10.1021/acs.jcim.4c02357
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

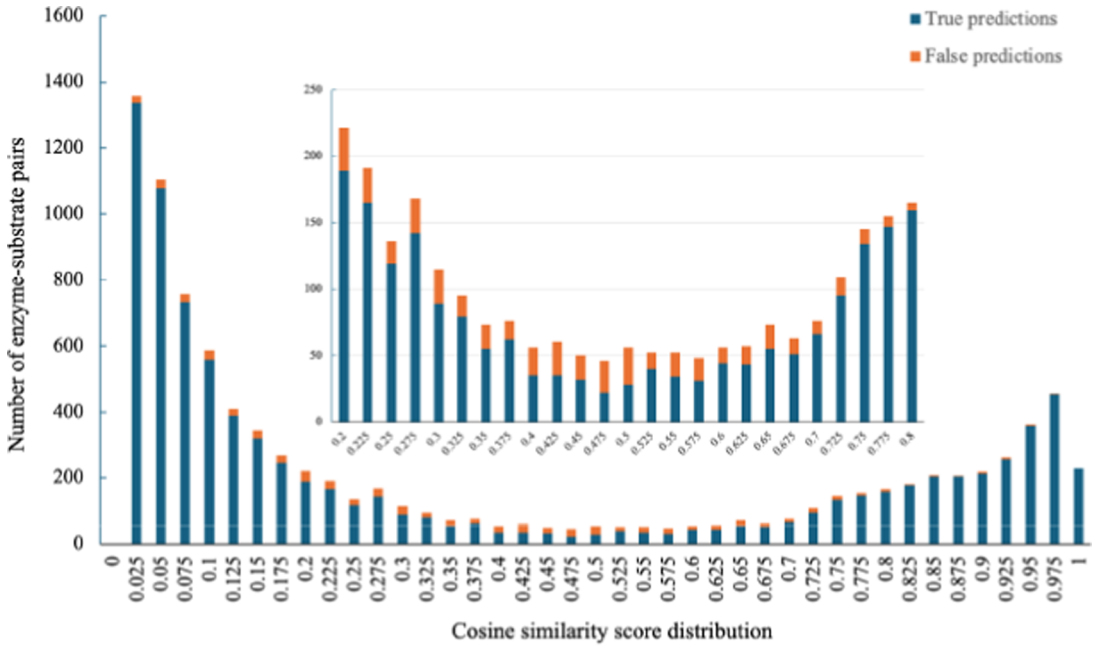


Figure 3. Prediction scores around 0.5 indicate model uncertainty. Note: stacked histogram bars display the prediction score distributions of true predictions (blue) and false predictions (red). The inset shows a blow-up of the interval [0.2, 0.8]. Scores are predicted by FusionESP-XL with ESM-2-3B and MoLFormer.

between 0.4 and 0.6 were more likely to be incorrectly predicted. Therefore, for practical usage of FusionESP-XL, input pairs with cosine similarity score within the 0.4 to 0.6 range should be treated as uncertain and used cautiously in decision making.

4. CONCLUSIONS

Overall, our proposed contrastive learning strategy (FusionESP) achieved SOTA performance by leveraging two frozen encoders and two simple projection heads with an MSE loss function. The best model, FusionESP-XL, which utilized ESM-2−3B and MoLFormer, achieved SOTA performance with accuracy, AUC, and MCC of 94.77%, 0.9653, and 0.8628, respectively. Even under limited data availability (using an experimental evidence-based data set only), our model FusionESP-exp with ESM-2-3B and MoLFormer also achieved SOTA performance with accuracy, AUC, and MCC of 93.57%, 0.9594, and 0.8314, respectively. Furthermore, our models can handle proteins of any length, unlike ProSmith, which truncates all enzymes into 1023 amino acids due to constraints from the ESM-1b models. This truncation can lead to confusion when two different enzymes share the same first 1023 residues from the N- to C-terminus. Moreover, our strategy does not require extra pretrained encoders or data sets for pretraining but exhibited better generalization capability, whereas ProSmith's development relied on a highly correlated data set (ligand-target affinity) to pretrain the multimodal BERT encoders before applying them to enzyme−substrate complex embedding. Our model (FusionESP-XL with ESM-2-3B and MoLFormer) shows great potential for future enzyme−substrate pair prediction, mapping the reliable relationship between enzymes and their substrates.

In the detailed ablation studies, the projection head was found as an effective module in both contrastive learning and concatenation contexts. Compared with the popular concatenation strategy (two neural network layers), the strategy with our projection head in this study can further enhance the model's performance. When combined with a contrastive learning strategy, independent projection heads exhibited even better performance. We also found that the completeness of the projection heads was important to maximize its capacity, and modifications to the projection head undermined its performance in both contrastive learning and concatenation strategies.

The contrastive learning strategy proposed in this study shows significant potential for other multimodal prediction tasks. However, it is worth noting that our model was built on pretrained LLMs and had inherent limitations, for example, the relatively large computational resources required to generate embeddings for long protein sequences. Furthermore, the learning process of FusionESP relies entirely on known information to map enzymes and small molecules into the same high-dimensional space. While our model has made notable progress compared to previous studies, this reliance may limit its generalizability to unseen data points. As a plug-and-play module, we believe that FusionESP can be easily compatible with other encoding approaches for performance enhancement. Future work will focus on combining FusionESP with lightweight yet powerful encoding strategies tailored for molecules and proteins, aiming to enhance efficiency without compromising performance.

■ ASSOCIATED CONTENT

Data Availability Statement

All source codes used in this publication are freely available for academic use under an MIT license at https://github.com/dzjxzyd/FusionESP. The data sets can be download at https://zenodo.org/records/13891018.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c02357.

https://doi.org/10.1021/acs.jcim.4c02357
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Learning curves of the model on the phylogenetic evidence-based data set and the experimental evidence-based data set, the pretrained PLM used in this study, and more detailed performance parameters of our models (PDF)

AUTHOR INFORMATION

Corresponding Author

Yonghui Li − *Department of Grain Science and Industry, Kansas State University, Manhattan, Kansas 66506, United States;* orcid.org/0000-0003-4320-0806; Phone: 785-532-4061; Email: yonghui@ksu.edu

Authors

Zhenjiao Du − *Department of Grain Science and Industry, Kansas State University, Manhattan, Kansas 66506, United States;* orcid.org/0000-0002-8492-4328

Weimin Fu − *Department of Electrical and Computer Engineering, Kansas State University, Manhattan, Kansas 66506, United States;* orcid.org/0000-0002-9623-6522

Xiaolong Guo − *Department of Electrical and Computer Engineering, Kansas State University, Manhattan, Kansas 66506, United States*

Doina Caragea − *Department of Computer Science, Kansas State University, Manhattan, Kansas 66506, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c02357

Author Contributions

Z.D.: Conceptualization, methodology, validation, formal analysis, investigation, visualization, writing—original draft, and writing—review and editing. W.F.: Investigation and writing—review and editing. X.G.: Investigation, resources, writing—review and editing, and funding acquisition. D.C.: Investigation, resources, and writing—review and editing. Y.L.: Conceptualization, methodology, formal analysis, investigation, resources, writing—review and editing, supervision, project administration, and funding acquisition.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This is contribution no. 25-059 J from the Kansas Agricultural Experimental Station. This work is part of Zhenjiao Du's PhD dissertation entitled "Machine Learning Empowered Discovery of Bioactive Peptides from Food Proteins and Beyond" at Kansas State University. This work was supported in part by the National Science Foundation (2419880) and the Plant Protein Innovation Center.

REFERENCES

(1) Kim, G. B.; Kim, J. Y.; Lee, J. A.; Norsigian, C. J.; Palsson, B. O.; Lee, S. Y. Functional Annotation of Enzyme-Encoding Genes Using Deep Learning with Transformer Layers. *Nat. Commun.* **2023**, *14* (1), 7370.

(2) Kroll, A.; Ranjan, S.; Engqvist, M. K. M.; Lercher, M. J. A General Model to Predict Small Molecule Substrates of Enzymes Based on Machine and Deep Learning. *Nat. Commun.* **2023**, *14* (1), 2787.

(3) Copley, S. D. Shining a Light on Enzyme Promiscuity. *Curr. Opin. Struct. Biol.* **2017**, *47*, 167−175.

(4) Nobeli, I.; Favia, A. D.; Thornton, J. M. Protein Promiscuity and Its Implications for Biotechnology. *Nat. Biotechnol.* **2009**, *27* (2), 157−167.

(5) Kroll, A.; Ranjan, S.; Lercher, M. J. Drug-Target Interaction Prediction Using a Multi-Modal Transformer Network Demonstrates High Generalizability to Unseen Proteins. *bioRxiv* **2023**, *12*, 2023.

(6) Zhang, D.; Jia, C.; Sun, D.; Gao, C.; Fu, D.; Cai, P.; Hu, Q.-N. Data-Driven Prediction of Molecular Biotransformations in Food Fermentation. *J. Agric. Food Chem.* **2023**, *71* (22), 8488−8496.

(7) 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. Discovery of Toxin-Degrading Enzymes with Positive Unlabeled Deep Learning. *ACS Catal.* **2024**, *14*, 3336−3348.

(8) Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bursteinas, B.; The UniProt Consortium; et al. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D480−D489.

(9) Wang, X.; Quinn, D.; Moody, T. S.; Huang, M. ALDELE: All-Purpose Deep Learning Toolkits for Predicting the Biocatalytic Activities of Enzymes. *J. Chem. Inf. Model.* **2024**, *64* (8), 3123−3139.

(10) Kalia, A.; Krishnan, D.; Hassoun, S. CSI: Contrastive data Stratification for Interaction prediction and its application to compound−protein interaction prediction. *Bioinformatics* **2023**, *39* (8), btad456.

(11) Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: Deep Drug−Target Binding Affinity Prediction. *Bioinformatics* **2018**, *34* (17), i821−i829.

(12) Yuan, W.; Chen, G.; Chen, C. Y.-C. FusionDTA: Attention-Based Feature Polymerizer and Knowledge Distillation for Drug-Target Binding Affinity Prediction. *Briefings Bioinf.* **2022**, *23* (1), bbab506.

(13) Nguyen, T.; Le, H.; Quinn, T. P.; Nguyen, T.; Le, T. D.; Venkatesh, S. GraphDTA: Predicting Drug−Target Binding Affinity with Graph Neural Networks. *Bioinformatics* **2021**, *37* (8), 1140−1147.

(14) Zhang, Y.; Li, S.; Meng, K.; Sun, S. Machine Learning for Sequence and Structure-Based Protein−Ligand Interaction Prediction. *J. Chem. Inf. Model.* **2024**, *64* (5), 1456−1472.

(15) Peng, L.; Liu, X.; Chen, M.; Liao, W.; Mao, J.; Zhou, L. MGNDTI: A Drug-Target Interaction Prediction Framework Based on Multimodal Representation Learning and the Gating Mechanism. *J. Chem. Inf. Model.* **2024**, *64* (16), 6684−6698.

(16) He, H.; Chen, G.; Chen, C. Y.-C. NHGNN-DTA: A Node-Adaptive Hybrid Graph Neural Network for Interpretable Drug−Target Binding Affinity Prediction. *Bioinformatics* **2023**, *39* (6), btad355.

(17) Nguyen, N.-Q.; Park, S.; Gim, M.; Kang, J. MulinforCPI: Enhancing Precision of Compound−Protein Interaction Prediction through Novel Perspectives on Multi-Level Information Integration. *Briefings Bioinf.* **2023**, *25* (1), bbad484.

(18) Nguyen, N.-Q.; Jang, G.; Kim, H.; Kang, J. Perceiver CPI: A Nested Cross-Attention Network for Compound−Protein Interaction Prediction. *Bioinformatics* **2023**, *39* (1), btac731.

(19) Song, N.; Dong, R.; Pu, Y.; Wang, E.; Xu, J.; Guo, F. PMF-CPI: Assessing Drug Selectivity with a Pretrained Multi-Functional Model for Compound−Protein Interactions. *J. Cheminf.* **2023**, *15* (1), 97.

(20) Wang, X.; Liu, J.; Zhang, C.; Wang, S. SSGraphCPI: A Novel Model for Predicting Compound-Protein Interactions Based on Deep Learning. *Int. J. Mol. Sci.* **2022**, *23* (7), 3780.

(21) Chen, Z.-H.; Zhao, B.-W.; Li, J.-Q.; Guo, Z.-H.; You, Z.-H. GraphCPIs: A Novel Graph-Based Computational Model for Potential Compound-Protein Interactions. *Mol. Ther.–Nucleic Acids* **2023**, *32*, 721−728.

(22) Li, Y.; Fan, Z.; Rao, J.; Chen, Z.; Chu, Q.; Zheng, M.; Li, X. An Overview of Recent Advances and Challenges in Predicting Compound-Protein Interaction (CPI). *Biomed. Rev.* **2023**, *3* (6), 465−486.

(23) Du, Z.; Comer, J.; Li, Y. Bioinformatics Approaches to Discovering Food-Derived Bioactive Peptides: Reviews and Perspectives. *TrAC, Trends Anal. Chem.* **2023**, *162*, 117051.
(24) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. **2017**, arXiv:1706.03762
(25) Ahmad, W.; Simon, E.; Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa-2: Towards Chemical Foundation Models. **2022**, arXiv:2209.01712
(26) Akdis, C. A. Does the Epithelial Barrier Hypothesis Explain the Increase in Allergy, Autoimmunity and Other Chronic Conditions? *Nat. Rev. Immunol.* **2021**, *21* (11), 739−751.
(27) Du, Z.; Ding, X.; Xu, Y.; Li, Y. UniDL4BioPep: A Universal Deep Learning Architecture for Binary Classification in Peptide Bioactivity. *Briefings Bioinf.* **2023**, *24*, 1−10.
(28) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; et al. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379* (6637), 1123−1130.
(29) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-Scale Chemical Language Representations Capture Molecular Structure and Properties. *Nat. Mach. Intell.* **2022**, *4* (12), 1256−1264.
(30) 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.; Xing, H.; Chen, H. Scientific Large Language Models: A Survey on Biological & Chemical Domains. **2024**, arXiv:2401.14656
(31) Thumuluri, V.; Martiny, H.-M.; Almagro Armenteros, J. J.; Salomon, J.; Nielsen, H.; Johansen, A. R. NetSolP: Predicting Protein Solubility in Escherichia Coli Using Language Models. *Bioinformatics* **2022**, *38* (4), 941−946.
(32) Lim, S.; Lu, Y.; Cho, C. Y.; Sung, I.; Kim, J.; Kim, Y.; Park, S.; Kim, S. A Review on Compound-Protein Interaction Prediction Methods: Data, Format, Representation and Model. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1541−1556.
(33) Mou, Z.; Eakes, J.; Cooper, C. J.; Foster, C. M.; Standaert, R. F.; Podar, M.; Doktycz, M. J.; Parks, J. M. Machine Learning-Based Prediction of Enzyme Substrate Scope: Application to Bacterial Nitrilases. *Proteins: Struct., Funct., Bioinf.* **2021**, *89* (3), 336−347.
(34) Lee, I.; Keum, J.; Nam, H. DeepConv-DTI: Prediction of Drug-Target Interactions via Deep Learning with Convolution on Protein Sequences. *PLoS Comput. Biol.* **2019**, *15* (6), No. e1007129.
(35) Qian, Y.; Li, X.; Wu, J.; Zhang, Q. MMCL-CPI: A multi-modal compound-protein interaction prediction model incorporating contrastive learning pre-training. *Comput. Biol. Chem.* **2024**, *112*, 108137.
(36) 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. Learning Transferable Visual Models From Natural Language Supervision. **2021**, arXiv:2103.00020
(37) Bhat, S.; Palepu, K.; Hong, L.; Mao, J.; Ye, T.; Iyer, R.; Zhao, L.; Chen, T.; Vincoff, S.; Watson, R.; Wang, T. Z.; Srijay, D.; Kavirayuni, V. S.; Kholina, K.; Goel, S.; Vure, P.; Deshpande, A. J.; Soderling, S. H.; DeLisa, M. P.; Chatterjee, P. De Novo Design of Peptide Binders to Conformationally Diverse Targets with Contrastive Language Modeling. *Sci. Adv.* **2025**, *11* (4), No. eadr8638.
(38) Palepu, K.; Ponnapati, M.; Bhat, S.; Tysinger, E.; Stan, T.; Brixi, G.; Koseki, S. R. T.; Chatterjee, P. Design of Peptide-Based Protein Degraders via Contrastive Deep Learning. **2022**, biorxiv 2022.05.23.493169.
(39) Gao, B.; Qiang, B.; Tan, H.; Jia, Y.; Ren, M.; Lu, M.; Liu, J.; Ma, W.-Y.; Lan, Y. DrugCLIP: Contrastive Protein-Molecule Representation Learning for Virtual Screening. *Advances in Neural Information Processing Systems*, 2023; Vol. *36*, pp 44595−44614.
(40) Tao, W.; Lin, X.; Liu, Y.; Zeng, L.; Ma, T.; Cheng, N.; Jiang, J.; Zeng, X.; Yuan, S. Bridging Chemical Structure and Conceptual Knowledge Enables Accurate Prediction of Compound-Protein Interaction. *BMC Biol.* **2024**, *22* (1), 248.
(41) Singh, R.; Sledzieski, S.; Bryson, B.; Cowen, L.; Berger, B. Contrastive Learning in Protein Language Space Predicts Interactions between Drugs and Protein Targets. *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120* (24), No. e2220778120.
(42) Wang, Z.; Wang, Z.; Yang, M.; Pang, L.; Nie, F.; Liu, S.; Gao, Z.; Zhao, G.; Ji, X.; Huang, D.; Zhu, Z.; Li, D.; Yuan, Y.; Zheng, H.; Zhang, L.; Ke, G.; Wang, D.; Yu, F. Enhancing Challenging Target Screening via Multimodal Protein-Ligand Contrastive Learning. **2024**, bioRxiv 2024.08.22.609123..
(43) Yu, T.; Cui, H.; Li, J. C.; Luo, Y.; Jiang, G.; Zhao, H. Enzyme Function Prediction Using Contrastive Learning. *Science* **2023**, *379* (6639), 1358−1363.
(44) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118* (15), No. e2016239118.
(45) Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A. Simple Framework for Contrastive Learning of Visual Representations. **2020**, arXiv:2002.05709.
(46) 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. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. **2020**, arXiv:2006.07733.
(47) Richemond, P. H.; Grill, J.-B.; Altché, F.; Tallec, C.; Strub, F.; Brock, A.; Smith, S.; De, S.; Pascanu, R.; Piot, B.; Valko, M. BYOL Works Even without Batch Statistics. **2020**, arXiv:2010.10241.
(48) Du, Z.; Xu, Y.; Liu, C.; Li, Y. pLM4Alg: Protein Language Model-Based Predictors for Allergenic Proteins and Peptides. *J. Agric. Food Chem.* **2024**, *72*, 752.
(49) Li, W.; Godzik, A. Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics* **2006**, *22* (13), 1658−1659.

https://doi.org/10.1021/acs.jcim.4c02357
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX