DeepRES: Deep learning enables reaction-based comprehensive enzyme screening

Keisuke Hirota[1], Takuji Yamada[1,2,3,4*]

1. School of Life Science and Technology, Institute of Science Tokyo
2. Metagen Inc.
3. Metagen Therapeutics Inc.
4. digzyme Inc.

*Corresponding author. School of Life Science and Technology, Institute of Science Tokyo, Tokyo, Japan; E-mail: takuji@bio.titech.ac.jp

Abstract

Background

Enzymes accelerate biochemical reactions in living organisms, thus playing an important role in metabolism. Although metabolic pathway databases are growing, many metabolic reactions, termed orphan enzymes, have not been annotated to gene sequences, which hinders functional annotation in genomic analysis. Moreover, protein databases contain many proteins of unknown function. Owing to this gap between known proteins and enzymatic reactions, various proteins of unknown function may be orphan enzymes; however, available tools cannot adequately predict these links.

Results

In this study, we developed DeepRES, an AI-based framework for comprehensive enzyme screening, to explore novel enzyme candidates from proteins of unknown function for reactions of interest. DeepRES implements enzyme screening via two steps: classification of enzymes and non-enzymes and prediction of catalytic capabilities for enzyme–reaction pairs. The two deep learning models comprising DeepRES showed comparable or superior performance to that of existing software. We performed screening of 1,255 orphan enzymes involved in the microbiome using DeepRES and successfully identified candidate proteins for 897 orphan enzymes. We then used those candidates as references for genomic analysis and explored novel biosynthetic gene clusters from microbial genomes to obtain promising candidate gene clusters, including those related to anthocyanin degradation.

Conclusions

Comprehensive enzyme screening via DeepRES, which is the first computational tool designed to associate orphan enzymes with proteins of unknown function, is expected to facilitate high-throughput identification of orphan enzyme-encoding genes. Furthermore, DeepRES can be easily integrated into the current genomic analysis pipeline to extend the functional annotation.

Background

Enzymes are proteins that exhibit catalytic activity and accelerate biochemical reactions in living organisms. A series of enzymatic reactions is referred to as a metabolic pathway, and metabolic pathway databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [1], are frequently used to infer metabolic function from genome sequences. Despite the accumulation of vast amounts of gene sequence data due to advances in sequencing technologies and analytical software, from approximately 20% to 50% of the reactions in pathway databases remain unassigned to a corresponding gene sequence [1–3]. Enzymes for which gene sequence information related to confirmed activities is lacking are defined as orphan enzymes [4,5]; this lack of sequence information hinders functional analysis, such as functional annotation and metabolic modeling. Therefore, the identification of genes that encode orphan enzymes is essential for the advancement of genomic analysis.

Moreover, protein databases such as UniProt [6] contain many proteins of unknown function. Owing to the gap between known proteins and enzymatic reactions, various proteins of unknown function are expected to be orphan enzymes. Therefore, enzyme screening for proteins of unknown function is necessary for the identification of orphan enzymes. As experimentally verifying protein function is time-consuming and labor-intensive, a computational approach is especially needed [7].

Numerous computational tools for enzyme screening have been developed, and they are categorized into two primary types: enzyme classification methods and enzyme prediction methods. Enzyme classification methods take protein sequences as inputs and output functional annotations. These include homology search tools, such as BLAST [8], and machine learning (ML) models that predict functional labels for input protein sequences [9–

14]. Following the recent success of highly accurate protein 3D structure prediction [15,16], structure-based computational methods have flourished, with the development of Foldseek [17], which is a fast protein structure search tool, and ML models for predicting enzymatic functions on the basis of protein structures [18,19]. ML models that are independent of reference databases are helpful in the exploration of orphan enzyme-encoding genes. However, existing models are designed to predict functional classifications such as EC numbers and thus have the limitation that they can handle only enzymatic reactions within the range covered by the classification schema. In our previous work, we developed an ML-based framework that covers a broader range of enzymatic reactions by employing reaction classifications on the basis of the chemical structural transformation patterns of substrates and products; however, it still cannot handle all reactions owing to its dependence on the reaction classification schema [3]. Furthermore, most existing approaches have limitations in the prediction of enzyme classification for non-enzymes as they do not consider the input of non-enzymes. In contrast, enzyme prediction methods such as E-zyme 2.0 [20] take any reaction as input and output enzymes that catalyze similar reactions. While these approaches can handle all reactions, including orphan enzyme reactions, their search space is restricted to known enzymes; they cannot search for orphan enzyme candidates from proteins of unknown function. Most existing computational tools do not adequately predict whether a protein of unknown function is likely to be an orphan enzyme.

Implementing a computational framework that enables direct and transverse searches for proteins and enzyme reactions allows for the exploration of enzyme candidates from all proteins, including those with unknown functions, for any given reaction. To this end, multi-modal AI that processes proteins and reactions has the

potential to achieve this goal, and cross-modal retrieval via multi-modal AI is expected to be a compelling avenue. Originally introduced within the computer-vision community, contrastive language–image pre-training (CLIP) [21] aligns natural-language captions with their corresponding images by projecting both modalities into a shared latent space; embedding proteins and enzymatic reactions into a shared latent space, as in CLIP, would enable comprehensive enzyme screening as a cross-modal retrieval.

In this study, we developed DeepRES, which is an AI-based framework for comprehensive enzyme screening, to explore novel enzyme candidates from proteins of unknown function for reactions of interest. DeepRES consists of a protein classification model (EnzymeCNN) and an enzyme screening model (EnzymeCLIP) (Figure 1A). We show that EnzymeCNN and EnzymeCLIP achieve comprehensive enzyme screening and outperform existing computational tools. We subsequently focused on human gut bacteria because of the abundance of orphan enzymes and genes of unknown function [2,22], and the availability of extensive metabolic pathway and protein 3D structure resources [2,23], therefore we explored candidate proteins from the ESM Metagenomic Atlas [23] for 1,255 orphan enzymes.

Results

DeepRES framework specification

DeepRES is an AI-based framework for comprehensive enzyme screening that comprises two deep learning models: EnzymeCNN and EnzymeCLIP (Figure 1A). EnzymeCNN is a convolutional neural network (CNN)-based model used to classify enzymes and non-enzymes according to protein sequence and structure (Figure 1B). EnzymeCLIP is a CLIP-like model used to predict enzymatic activities for protein–reaction pairs by projecting proteins and reactions into the same embedding space and calculating cosine similarities as scores (Figure 1C). DeepRES performs enzyme screening by taking proteins and reactions as inputs. DeepRES first classifies proteins as enzymes or non-enzymes via EnzymeCNN. For those proteins predicted to be enzymes, EnzymeCLIP then predicts their correspondence to given reactions. These processes enable DeepRES to consider that input proteins might be non-enzymes and to be independent of the reaction classification schema.

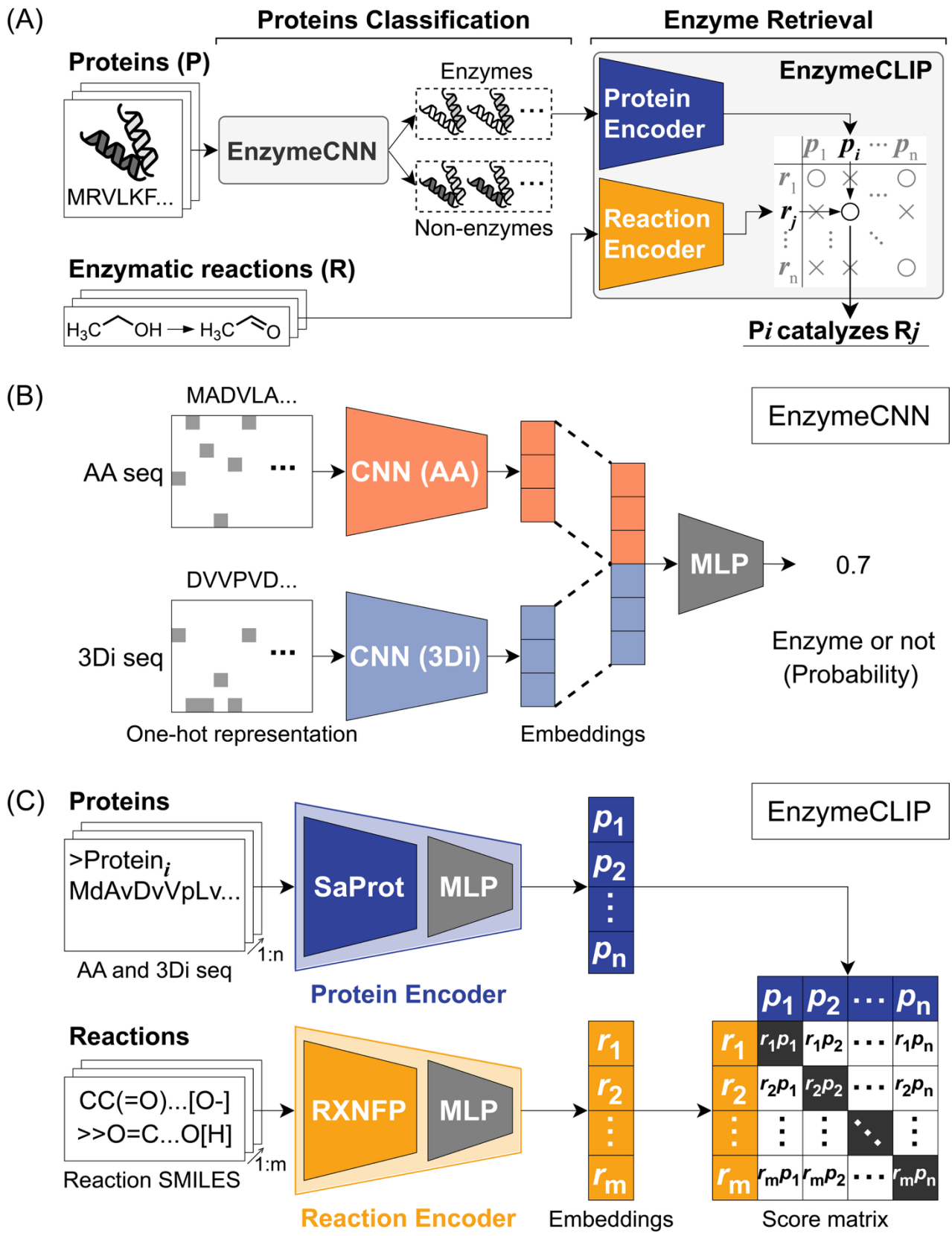

Figure 1: Overview of DeepRES. (A) DeepRES consists of two deep learning models: EnzymeCNN and EnzymeCLIP. First, DeepRES takes proteins as inputs and classifies them into enzymes and non-enzymes via EnzymeCNN. Next,

DeepRES associates the predicted enzymes with input reactions by EnzymeCLIP. (B) EnzymeCNN is a convolutional neural network (CNN)-based model for protein classification. EnzymeCNN classifies input proteins into enzymes and non-enzymes by processing amino acid sequences and 3Di sequences separately using CNNs and integrating the sequence and structural information. (C) EnzymeCLIP is a contrastive language–image pre-training (CLIP)-like model for enzyme retrieval. EnzymeCLIP is based on two pretrained language models, SaProt and RXNFP, and embeds input proteins and reactions into a shared latent space by the protein encoder and the reaction encoder. Thus, EnzymeCLIP can comprehensively predict the enzymatic activities of enzyme-reaction pairs of interest. Proteins in structure-aware sequences and enzymatic reactions in Reaction SMILES notation are passed to EnzymeCLIP.

EnzymeCNN development

EnzymeCNN effectively integrates protein sequence and structural information

We performed hyperparameter optimization of EnzymeCNN on a development dataset constructed from the Swiss-Prot [6] and the AlphaFold Protein Structure Database (AlphaFold DB) [15]. The results demonstrated that EnzymeCNN is robust to its hyperparameters and can classify enzymes and non-enzymes accurately (Table 1, Supplementary Figure S3). To isolate the contribution of each protein modality, we constructed two ablation models: EnzymeCNN-AA, which receives only amino acid sequences, and EnzymeCNN-3Di, which receives only 3Di sequences that represent protein 3D structures. After both ablation models were optimized via the same procedure as that for EnzymeCNN, the best EnzymeCNN model outperformed both ablations, which suggests the effectiveness of integrating protein sequences and structural information in the protein classification task

(Figure 2, Supplementary Figure S3). A comparison of EnzymeCNN-AA and EnzymeCNN-3Di indicated that protein 3D structures mainly help in classifying enzymes and non-enzymes.

Table 1: Results of EnzymeCNN hyperparameter tuning.

Number of ResNet layers	Embedding dim	Accuracy	F1-score	MCC	AUC
3	256	0.9671	0.9664	0.9342	0.9920
3	512	0.9671	0.9664	0.9343	0.9918
4	256	0.9696	0.9690	0.9393	0.9926
4	512	0.9689	0.9683	0.9378	0.9926
5	256	0.9722	0.9717	0.9444	0.9933
5	512	0.9692	0.9685	0.9383	0.9927

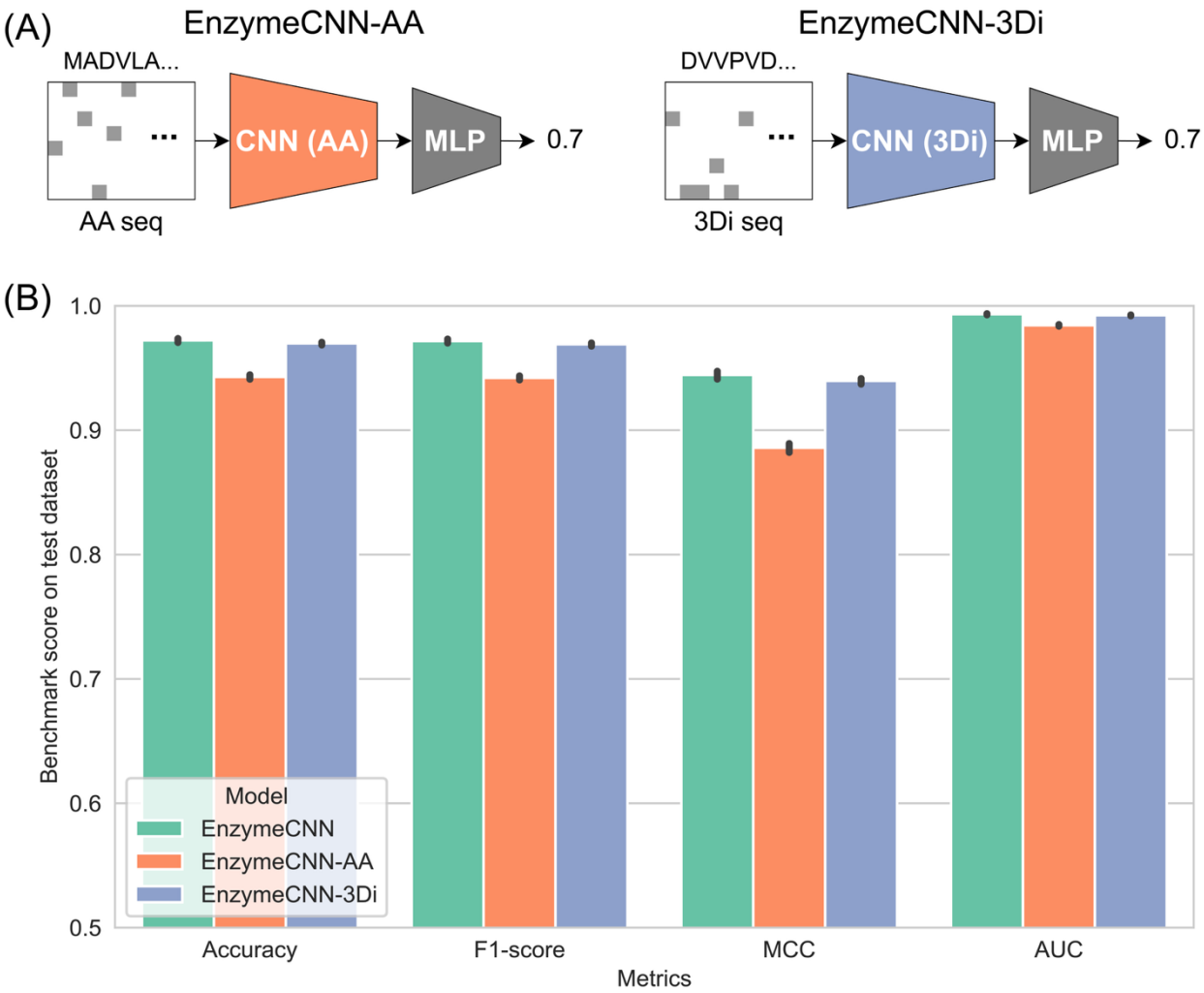

Figure 2: Performance comparison of EnzymeCNN with ablation models. (A) The two ablation models are EnzymeCNN-AA, which receives only amino acid sequences, and EnzymeCNN-3Di, which receives only 3Di sequences. (B) EnzymeCNN was compared with EnzymeCNN-AA and EnzymeCNN-3Di in terms of performance on the test dataset after hyperparameter tuning.

EnzymeCNN outperforms existing algorithms in the protein classification task

ML models are generally more sensitive to the functional similarity of proteins than are homology search algorithms [9,10]. Thus, we evaluated the performance of EnzymeCNN on proteins whose sequences or

structures are far from the training data to assess whether EnzymeCNN is better at detecting functional similarity than existing search methods are. Specifically, we removed validation and test data similar to training data in protein sequences or structures using MMseqs2 [24] and Foldseek [17], re-optimized the EnzymeCNN hyperparameters and compared EnzymeCNN with MMseqs2 and Foldseek in the protein classification task (Supplementary Tables S1 and S2). Across all similarity thresholds, EnzymeCNN outperformed the existing tools (Figure 3). Notably, when the similarity threshold was tightened, the performance of MMseqs2 and Foldseek collapsed, whereas EnzymeCNN retained its performance.

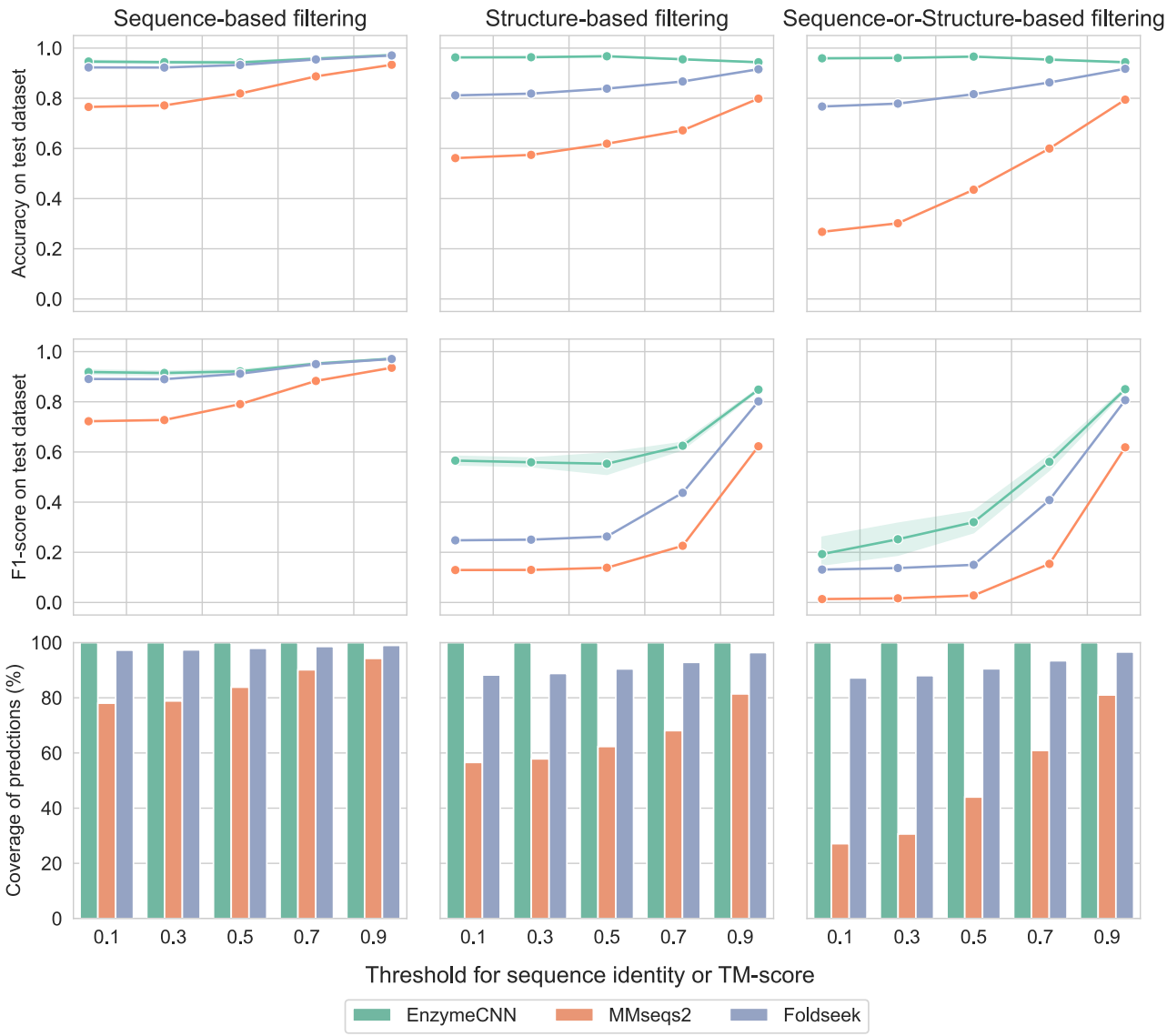

Figure 3: Performance comparison of EnzymeCNN with existing tools under protein similarity-limited situations. After removing validation and test data similar to the training data in terms of sequence identity and/or TM-score via MMseqs2 and/or Foldseek, EnzymeCNN and the two previous approaches were evaluated on the protein similarity-limited dataset. The bottom row shows the percentage of the test data that each tool was able to annotate as enzymes or non-enzymes. When MMseqs2 and Foldseek were evaluated, no hits were treated as incorrect.

The advantage of EnzymeCNN persisted when all computational approaches were benchmarked on new Swiss-Prot data. Compared with both MMseqs2 and Foldseek, EnzymeCNN consistently achieved higher overall accuracy and F1-scores (Figure 4). In addition, unlike the homology search, EnzymeCNN annotated all input proteins as enzymes or non-enzymes.

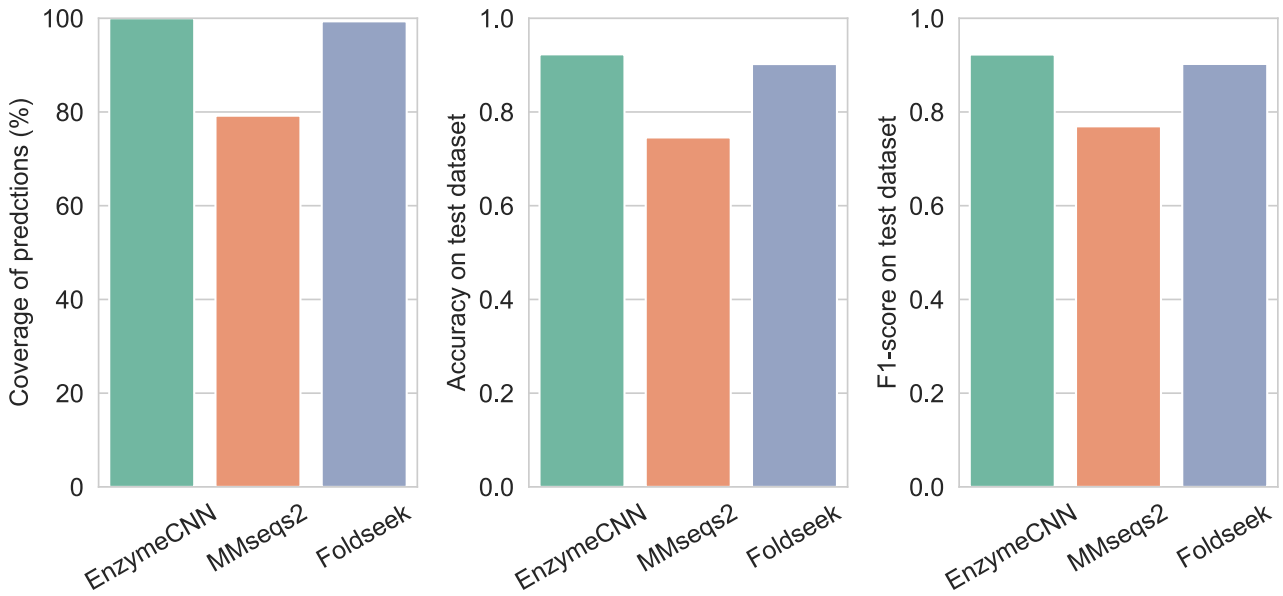

Figure 4: Performance comparison of EnzymeCNN with existing tools on the new protein dataset. EnzymeCNN, MMseqs2 and Foldseek were evaluated on the new protein dataset. The left graph shows the percentage of new data that each tool was able to annotate as enzymes or non-enzymes. In evaluating MMseqs2 and Foldseek, no-hits were treated as incorrect.

EnzymeCLIP development

EnzymeCLIP performed well in the enzyme retrieval task

Using a development dataset assembled from Swiss-Prot [6], AlphaFold DB [15] and Rhea [25], we systematically tuned the architecture of EnzymeCLIP and its training settings, such as loss functions, batch sizes

and fine-tuning of the two pretrained models (Supplementary Table S3). The experimental results demonstrated that fine-tuning the underlying language models in EnzymeCLIP is effective for enhancing performance. SaProt-based models consistently outperformed ESM-2-based models, thus underscoring the importance of integrating protein sequence and structural information in the enzyme retrieval task. In contrast, the performance of EnzymeCLIP was largely insensitive to the choice of batch size across the range we tested. The use of cyclic loss improved the performance of EnzymeCLIP, whereas replacing contrastive loss with soft contrastive loss led to a decline in performance.

EnzymeCLIP outperformed the state-of-the-art method in the EC number prediction task

Since another computational approach similar to EnzymeCLIP has not yet been established, we compared EnzymeCLIP with CLEAN [9], which is the state-of-the-art enzyme function annotation tool, in the EC number prediction task on new Swiss-Prot and Rhea data (Supplementary Figure S4). Although CLEAN is a multi-label classifier and EnzymeCLIP is designed for a screening task, EnzymeCLIP performed competitively with CLEAN in terms of top-1 accuracy and outperformed CLEAN in terms of top-10 accuracy (Figure 5).

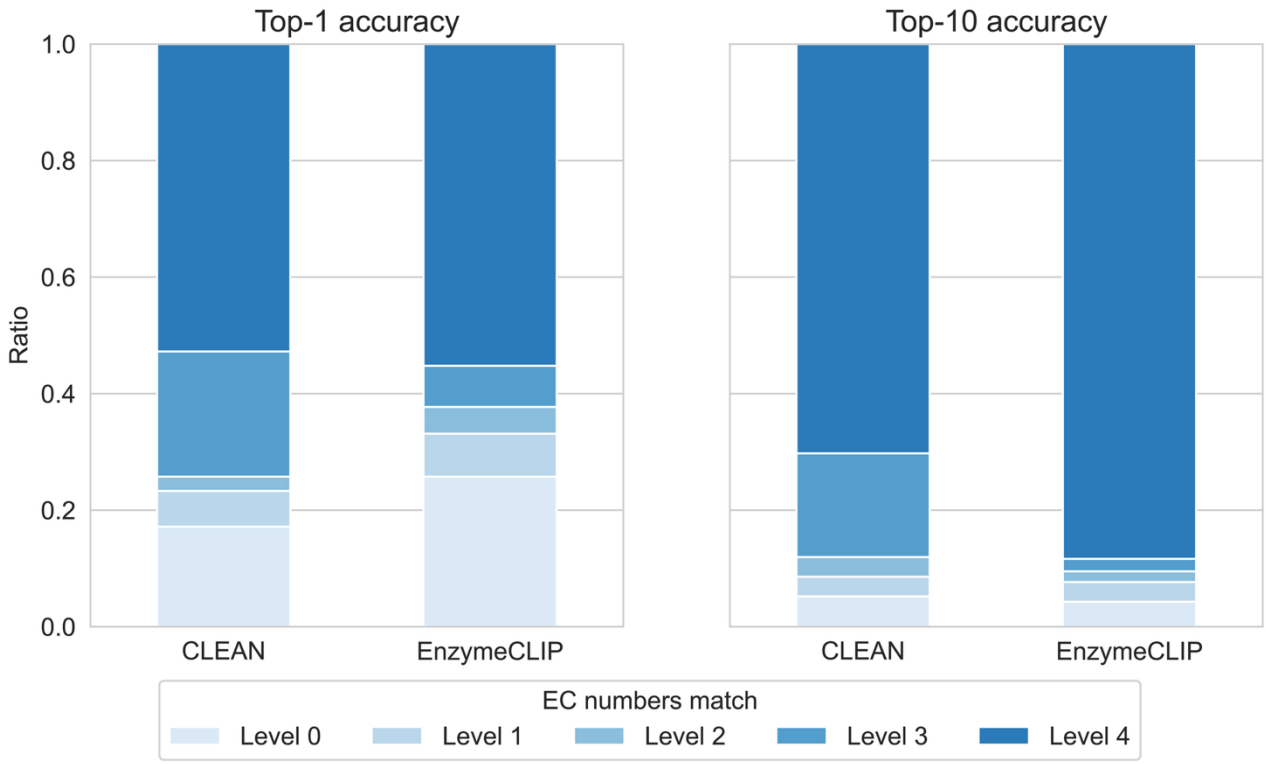

Figure 5: Performance comparison of EnzymeCNN with CLEAN in the EC number prediction task. Performance comparison of EnzymeCNN with CLEAN on the new enzyme–reaction dataset in terms of the top-1 accuracy and the top-10 accuracy. If the actual EC number is 1.1.1.1 and the predicted EC number is 1.1.2.3, the EC numbers match at level 2.

Application of DeepRES to metagenome data to explore orphan enzyme candidates

We focused on metagenome data and proteins of unknown function associated with orphan enzymes using DeepRES. Specifically, we applied DeepRES to 13,902,058 metagenomic proteins of unknown function obtained from the ESM Metagenomic Atlas [23] and 1,255 orphan enzymes involved in human gut bacteria recorded in EnteroPathway [2] (Figure 6). First, the 13,902,058 proteins were classified as enzymes and non-enzymes by EnzymeCNN, which resulted in 595,340 predicted enzymes. Next, we explored candidate proteins

220 for the 1,255 orphan enzymes identified by EnzymeCLIP, and associated 97,074 proteins of unknown function
221 with 897 orphan enzymes. Furthermore, to annotate the remaining predicted enzymes with enzymatic reactions,
222 we explored candidate proteins for well-known enzymatic reactions in Rhea (release 138) using EnzymeCLIP
223 (Supplementary Figure S5).

224 We obtained 4,744 representative metagenome-assembled genomes (MAGs) of human gut bacteria from
225 MGnify (Unified Human Gastrointestinal Genome v2.0.1)[26] and queried all genes in the MAGs against the
226 orphan enzyme candidates identified by DeepRES as reference databases using BLAST. The homology search
227 against putative orphan enzyme proteins enabled us to assign the functional annotations of the orphan enzymes
228 to the MAGs. Inspired by previous studies [3,27], we attempted to identify novel biosynthetic gene clusters
229 (BGCs) that correspond to metabolically related orphan enzymes by evaluating orphan enzyme candidates in
230 microbial genomes, with consideration of genomic and metabolomic neighboring information (Table 2). For
231 example, BGC screening for EPM0520 and EPM0522 related to anthocyanin degradation (Figure 7) identified
232 novel BGC candidates from 50 and 98 microbial genomes, respectively (Supplementary Table S4).

233

234

235

236

237

238

Table 2: List of modules subjected to a novel biosynthetic gene cluster search.

ID	Description
EPM0027	Gentiopicrin degradation
EPM0069	Sennoside metabolism
EPM0087	Arctiin degradation
EPM0191	Kalopanax saponin K degradation
EPM0304	EGCG degradation
EPM0336	Verbascoside degradation
EPM0406	Capilliposide C degradation
EPM0414	Clematichinenoside AR degradation
EPM0520	Cyanidin-3-rutinoside degradation
EPM0522	Delphinidin 3-O-rutinoside degradation
EPM0550	Procyanidin dimer B2 degradation
EPM0575	Astragaloside II degradation
EPM0576	Soyasaponin Ab degradation
EPM0603	Ellagic acid degradation
EPM0609	Catechin degradation
EPM0665	3,5-Dicaffeoylquinic acid degradation
EPM0666	3,5-Diferuloylquinic acid degradation
EPM0750	Procyanidin dimer B2 degradation
EPM0751	Procyanidin dimer B2 degradation
EPM1203	Mogroside V degradation

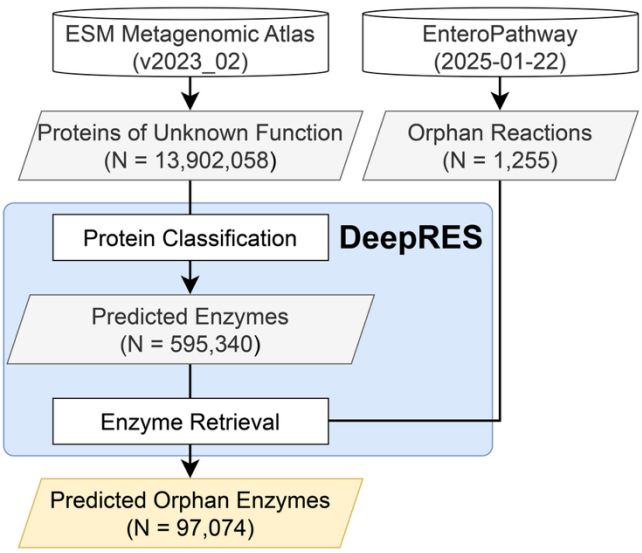

Figure 6: Flow and results of the application of DeepRES to metagenome data.

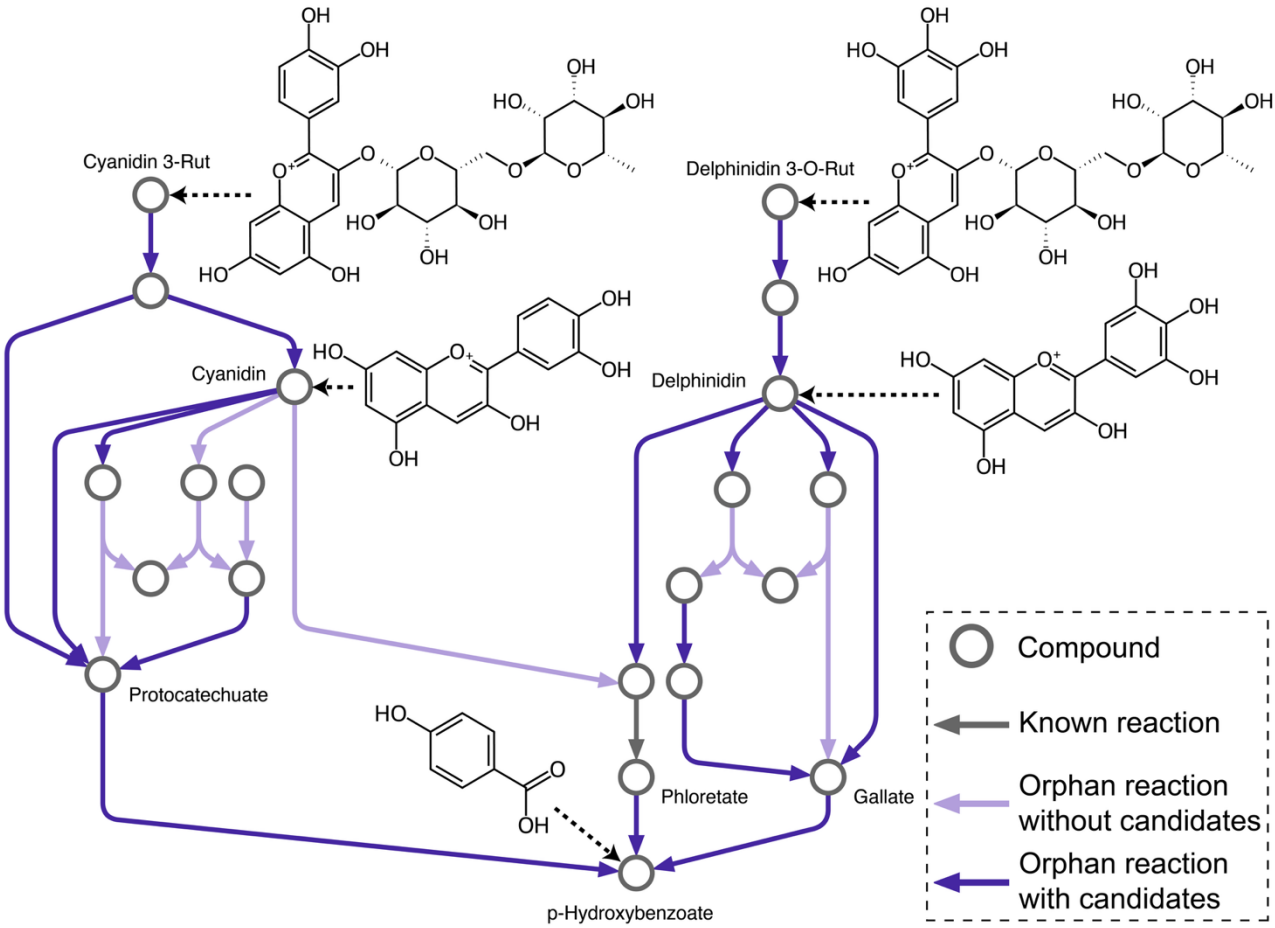

Figure 7: Examples of BGC screening with putative orphan enzyme proteins obtained by DeepRES. The two anthocyanin degradation modules with the most orphan enzymes are EPM0520 (cyanidin-3-rutinoside degradation) and EPM0522 (delphinidin 3-O-rutinoside degradation).

Discussions

Metabolic pathway databases are continually growing due to the accumulation of results from previous studies, and many enzymatic reactions in the databases remain unassigned to corresponding gene sequences; the associated enzymes are referred to as orphan enzymes. In this study, which focused on the possibility that some proteins of unknown function are orphan enzymes, we developed DeepRES, which is an AI-based

255 framework for comprehensive enzyme screening, to explore novel enzyme candidates from proteins of
256 unknown function for reactions of interest. To our knowledge, DeepRES is the first method designed to associate
257 any reaction, including those that involve orphan enzymes, with any protein, including functionally unknown
258 proteins. The two models that comprise DeepRES outperformed existing computational tools. We applied
259 DeepRES to metagenome data, which consisted of 13,902,058 metagenomic proteins of unknown function and
260 1,255 orphan enzymes involved in the metabolism of human gut bacteria. We classified the 13,902,058 proteins
261 into enzymes and non-enzymes via EnzymeCNN and successfully mapped the 97,074 predicted enzymes to
262 897 orphan enzymes using EnzymeCLIP.

263 In addition to its primary role in comprehensive enzyme screening, DeepRES provides a catalog of putative
264 enzymes for homology search. The use of the orphan enzyme candidates identified by DeepRES as a reference
265 database is also beneficial because this approach does not require a GPU and can be easily integrated into the
266 current genomic analysis pipeline, particularly for functional annotation. We demonstrated that this approach
267 enables us to identify orphan enzyme candidate genes that are linked to a taxonomy and a genomic location
268 within microbial genomes. Furthermore, following the findings of previous studies that integrating genomic and
269 metabolomic neighboring information facilitates the identification of orphan enzyme-encoding genes [3,27],
270 we successfully identified BGC candidates from microbial genomes for 15 EnteroPathway modules, which were
271 composed mostly of orphan enzymes (Supplementary Table S4). These modules included anthocyanin
272 degradation (EPM0520 and EPM0522). Anthocyanins, such as cyanidin and delphinidin have been suggested
273 to act on obesity, type 2 diabetes, inflammatory bowel disease and colorectal cancer through their conversion

274 into various metabolites by the human gut microbiota, thus they have attracted attention for dietary therapy
275 [28,29]. Despite its high importance, the mechanisms of anthocyanin degradation in the human gut microbiota
276 remain largely unclear [1,2]. Novel BGC candidates were detected for EPM0520 and EPM0522 in 50 and 98
277 microbial genomes, respectively, primarily in *Bacteroides* (Supplementary Table S4). These results are
278 promising because *Bacteroides* is capable of metabolizing anthocyanin [28,29]. In addition, more than 80% of
279 the bacteria with BGC candidates involved in anthocyanin degradation belong to *Bacteroidota*, which suggests
280 a potential association between the anthocyanin-induced decrease in the *Firmicutes*/*Bacteroidetes* ratio in the
281 gut microbiota [28,30] and the novel BGC candidates identified in this study. Moreover, although DeepRES
282 was used for microbiome research in this study, DeepRES is applicable to eukaryotes and viruses.

283 While DeepRES successfully mapped many previously uncharacterized proteins to specific enzymatic
284 reactions, including orphan enzymes, approximately 40% of the proteins predicted by EnzymeCNN to be
285 enzymes were not linked to enzymatic reactions (Supplementary Figure S5). This result arose from the nature
286 of EnzymeCLIP which uses a reaction dataset as a reference database, and indicates the limitations of
287 DeepRES's applicability, as well as the possibility that the remaining predicted enzymes might catalyze
288 unknown enzymatic reactions.

289 Unlike existing approaches, DeepRES is independent of reaction classification, and its prediction can be
290 converted into any preferred classification schema, as demonstrated in its comparison with CLEAN in the EC
291 number prediction task. However, there is room for improvement in the numerical representation of enzymatic
292 reactions in DeepRES. DeepRES is based on RXNFP, which embeds Reaction SMILES representation via

natural-language processing (NLP) techniques; however, the NLP models pose difficulties in recognizing chirality [31]. Given that enzymatic reactions are characterized by high stereoselectivity, improving the underlying reaction encoding is expected to further increase the effectiveness of DeepRES.

Although increasing the batch size is generally effective in contrastive learning [21,32,33], this trend was not observed during the development of EnzymeCLIP. The multimodal data of proteins and enzymatic reactions include many-to-many relationships due to homologous proteins, protein multifunctionality and convergent evolution. These complex associations may offset the benefit of larger batches by introducing many-to-many relationships within mini-batches that CLIP cannot account for. That is, the results suggest the limitations of a CLIP-like approach for biological multi-modal data. Inspired by SoftCLIP [34], we introduced a soft contrastive loss to relax the assumptions regarding one-to-one relationships in the CLIP strategy; however, this approach decreased the performance of EnzymeCLIP. This could be explained by the possibility that the global similarity of proteins obscures the similarities of their enzymatic functions. Taken together, CLIP is applicable to any multi-modal data, thus enabling comprehensive enzyme screening in this study; however, its application to biological multi-modal data still leaves room for improvement.

Conclusions

We developed DeepRES, which is a deep learning-based computational tool for comprehensive enzyme screening. Key features that distinguish DeepRES from previous approaches are the explicit consideration of proteins of unknown function and its independence of reaction classification schema, such as EC numbers.

312 DeepRES identified candidate proteins for 897 orphan enzymes associated with human gut bacteria.
313 Furthermore, we explored novel gene clusters involved in human-gut interactions by employing those
314 candidates as reference datasets in a homology search. The results demonstrate the potential of DeepRES to
315 bridge the gap between our understanding of gene sequences and functions and to expand the current genome
316 analysis. DeepRES is expected to contribute to the high-throughput and cost-effective identification of genes
317 that encode orphan enzymes, thereby promoting a deeper understanding of metabolism and advancing enzyme
318 engineering.

319

320

321 Methods

322 Dataset creation

323 To develop DeepRES, we constructed a protein dataset for protein classification and an enzyme–reaction
324 dataset for enzyme screening. All the data were retrieved from UniProt [6], AlphaFold DB [15] and Rhea [25].

325 First, all Swiss-Prot protein sequences were obtained from UniProt (release 2024_03) and subjected to
326 homology clustering using MMseqs2 (v15-6f452) [24] with an identity threshold of 0.9, which yielded 335,684
327 cluster representatives. Predicted protein 3D structure data from Swiss-Prot were retrieved from AlphaFold DB
328 (v4), and those data were converted to 3Di sequences using Foldseek (v9-427df8a) [17]. The 21,520 cluster
329 representatives that lacked 3Di sequences were discarded. The 19,642 cluster representatives longer than 850
330 residues were also discarded because of computational constraints. Amino acid and 3Di sequences were then

merged for each protein, which resulted in 294,522 protein sequences.

Next, the protein dataset was split 8:1:1 into training, validation, and test datasets, with a consistent ratio of enzymes to non-enzymes ensured for each split. This resulted in 235,617 training data points, 29,452 validation data points, and 29,453 test data points. The of enzyme and non-enzyme labels were assigned according to the presence and absence, respectively, of EC number annotations in Swiss-Prot; enzymes accounted for approximately 49% of each split.

Finally, the protein dataset was divided into enzymes and non-enzymes, and the enzymes were mapped to enzymatic reactions obtained from Rhea (release 133) to construct an enzyme–reaction dataset that contained 139,156 enzyme–reaction pairs. Since proteins and reactions have many-to-many relationships due to protein multifunctionality and convergent evolution, the enzyme–reaction dataset was first split at a ratio of 8:1:1 into training, validation and test datasets based on the associated reactions. Then, various training and validation data were removed to avoid overlap in proteins between data splits. This resulted in 106,174 training data points, 14,113 validation data points, and 18,869 test data points.

In addition, to compare the performance of DeepRES with those of previous computational tools, we constructed a new dataset that was not used in their development. Specifically, we utilized Swiss-Prot data updated from version 2024_03 to 2024_06 and Rhea data updated from release 133 to 136 to prepare new protein and enzyme–reaction datasets. As the 3D structures of most of the new proteins are not yet included in the AlphaFold DB, we performed protein 3D structure prediction via AlphaFold2 [15].

EnzymeCNN development

EnzymeCNN is a CNN-based model for classifying enzymes and non-enzymes based on protein sequences and structures, which are represented as amino acid and 3Di sequences. As CNNs have proven versatile for sequential data, including text, speech and protein sequences [10,35], we applied CNNs for protein classification especially by utilizing dilated convolution techniques [36], to broaden the receptive field exponentially without losing resolution. Hyperparameter tuning and performance tests were conducted on the development protein dataset five times with different seeds. Each model was trained for 50 epochs, and the checkpoint that yielded the highest validation accuracy was subsequently evaluated on the hold-out test dataset. To validate the benefits of EnzymeCNN approach, which integrates sequence and structural information, we built two ablation variants, EnzymeCNN-AA (sequence only) and EnzymeCNN-3Di (structure only), and tested them like we tested EnzymeCNN.

We next tested the performance of EnzymeCNN under limited sequence or structural homology. Specifically, we removed training data that were similar to validation or test data via homology searches with MMseqs2 and Foldseek, then re-optimized the hyperparameters and evaluated the performance of EnzymeCNN. The searches were performed with the sequence identity or TM-score thresholds set to 0.1, 0.3, …, 0.9 and the other parameters set to the MMseqs2 default values. For benchmarking, we additionally classified enzymes and non-enzymes by querying the test data against the training dataset using MMseqs2 and Foldseek with default settings.

Finally, we evaluated the performance of EnzymeCNN on the new protein dataset. The development dataset was split 9:1 into training and validation datasets, which were used to build EnzymeCNN, MMseqs2 and

Foldseek were again used as benchmark methods.

EnzymeCLIP development

EnzymeCLIP is a CLIP-like model for comprehensive enzyme screening built on two pretrained, modality-specific language models—SaProt [37], which is a structure-aware protein language model, and RXNFP [38], which is a BERT-based encoder for chemical reactions. EnzymeCLIP takes protein sequence and structural information, as well as reaction information, as inputs to predict catalytic activity for all combinations of input proteins and reactions. Protein information is supplied to EnzymeCLIP in the "structure-aware vocabulary" format introduced with SaProt, which integrates primary- and tertiary-structure features into a unified text representation. Reaction information is provided in Reaction SMILES format, which is a textual representation of chemical reactions. EnzymeCLIP embeds proteins and reactions into a latent space using the two separate encoders and is trained to pull vector representations of an enzymatic reaction and its catalytic protein close and to push those of non-catalytic pairs apart. Consequently, catalytic activity can be inferred by computing the cosine similarity between the resulting vectors. Thus, the catalytic activities of protein–reaction pairs can be predicted by calculating cosine similarities between the embedding vectors of proteins and reactions.

To improve the performance of EnzymeCLIP, both SaProt and RXNFP are fine-tuned within EnzymeCLIP using Low-Rank Adaptation (LoRA) [39]; the hyperparameters of LoRA are the same as those used in previous studies [39,40]. To assess the contribution of each element in EnzymeCLIP, we constructed ablation variants that (i) replace SaProt with ESM-2 [23], (ii) omit the fine-tuning of pretrained language models and (iii) vary

the batch size during training. We further examined loss functions inspired by previous studies [34,41]: contrastive loss, soft contrastive loss and cyclic loss. These loss functions are computed on the basis of a batch of N protein–reaction pairs $\{P_i, R_i\}_{i=1}^{N}$ where P_i represents a protein and R_i represents a reaction. EnzymeCLIP converts P_i and R_i into a protein embedding $P_i^e \in \mathbb{R}^{256}$ and a reaction embedding $R_i^e \in \mathbb{R}^{256}$, respectively. A protein-reaction similarity vector $\boldsymbol{p}_i(P,R)$ and a reaction-protein similarity vector $\boldsymbol{p}_i(R,P)$ are calculated as follows:

$$\boldsymbol{p}_i(P,R) = \frac{\exp(\langle P_i^e, R_i^e \rangle/\tau)}{\sum_{j=1}^{N} \exp\left(\langle P_i^e, R_j^e \rangle/\tau\right)}$$

$$\boldsymbol{p}_i(R,P) = \frac{\exp(\langle R_i^e, P_i^e \rangle/\tau)}{\sum_{j=1}^{N} \exp\left(\langle R_i^e, P_j^e \rangle/\tau\right)}$$

where $\langle \cdot,\cdot \rangle$ represents the inner product and where τ represents a trainable temperature parameter. The CLIP strategy uses the one-hot vector that corresponds to the input pairs $\boldsymbol{y}_i = \{y_{ij}\}_{j=1}^{N}$ as labels; therefore, the contrastive loss $\mathcal{L}_{\text{CLIP}}$ is computed as follows:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2N}\sum_{i=1}^{N} H\left(\boldsymbol{y}_i, \boldsymbol{p}_i(P,R)\right) + \frac{1}{2N}\sum_{j=1}^{N} H\left(\boldsymbol{y}_i, \boldsymbol{p}_i(R,P)\right)$$

$$= -\frac{1}{2N}\sum_{i=1}^{N} \log\left[\frac{\exp(\langle P_i^e, R_i^e \rangle/\tau)}{\sum_{j=1}^{N} \exp\left(\langle P_i^e, R_j^e \rangle/\tau\right)}\right] - \frac{1}{2N}\sum_{j=1}^{N} \log\left[\frac{\exp\left(\langle P_j^e, R_j^e \rangle/\tau\right)}{\sum_{i=1}^{N} \exp\left(\langle P_i^e, R_j^e \rangle/\tau\right)}\right]$$

We utilized the similarity between proteins $\boldsymbol{p}_i(P,P)$ to soften the one-hot label vector $\boldsymbol{y}_i = \{y_{ij}\}_{j=1}^{N}$ and calculate the soft contrastive loss $\mathcal{L}_{\text{soft}}$ as follows:

$$\tilde{\boldsymbol{y}}_i = (1-\alpha)\boldsymbol{y}_i + \alpha \boldsymbol{p}_i(P,P)$$

$$\mathcal{L}_{\text{soft}} = \frac{1}{2N}\sum_{i=1}^{N} H(\tilde{\boldsymbol{y}}_i, \boldsymbol{p}_i(P,R)) + \frac{1}{2N}\sum_{j=1}^{N} H(\tilde{\boldsymbol{y}}_i, \boldsymbol{p}_i(R,P))$$

where α is a hyperparameter, which is set to 0.3. Moreover, to align the shared embedding space for each

modality in EnzymeCLIP, we introduce the cyclic loss $\mathcal{L}_{\text{cyclic}}$, via following regularization term:

$$\mathcal{L}_{\text{Cyclic}} = \frac{\lambda_1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\langle P_i^e, R_j^e\rangle - \langle P_j^e, R_i^e\rangle\right)^2 + \frac{\lambda_2}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\langle P_i^e, P_j^e\rangle - \langle R_j^e, R_i^e\rangle\right)^2$$

where λ_1 and λ_2 are the hyperparameters, which are set to 0.25.

In total, 48 enzyme screening models (the original EnzymeCLIP and 47 variants) were trained for 10 epochs on the development enzyme–reaction dataset 5 times with different seeds. Their performance was evaluated in terms of the enrichment factor (EF) [42], which is a metric commonly used in screening tasks.

$$\text{EF} = \frac{\text{Hits}_{\text{sampled}}/N_{\text{sampled}}}{\text{Hits}_{\text{total}}/N_{\text{total}}}$$

EF measures how concentrated positives are within highly ranked predictions. For example, $\text{EF}_{0.10}$ evaluates the top 10% of predictions and takes values from 1 to 10, with a higher EF indicating better screening. For each model, the checkpoint that achieves the highest validation $\text{EF}_{0.10}$ was subsequently evaluated on the hold-out test dataset.

Finally, we evaluated the performance of EnzymeCLIP on the new enzyme–reaction dataset. To build EnzymeCLIP, the development dataset was split 9:1 into training and validation datasets on the basis of reactions. Then, some of the training data were removed to avoid overlap in proteins between the data splits. Since another computational method similar to EnzymeCLIP has not yet been established, we compared EnzymeCLIP with CLEAN [9], which is the state-of-the-art enzyme function annotation tool, in the EC number prediction task. Specifically, while CLEAN predicts EC numbers directly via multi-label classification, EnzymeCLIP annotates EC numbers indirectly for the input protein by comparing the input protein embedding with those of enzymatic reactions and retrieving the EC numbers associated with the enzymatic reactions mapped to the input protein.

Given the conceptual difference between EnzymeCLIP, which is a screening model, and CLEAN, which is a classification model, we evaluated their performance in terms of the top-k accuracy.

Orphan enzyme exploration

To associate proteins of unknown function with orphan enzymes and explore novel enzyme candidates, we applied DeepRES to metagenome data, including the ESM Metagenomic Atlas [23], which is a resource of metagenomic protein structures predicted by ESMFold, and EnteroPathway [2], which is a curated metabolic pathway database of the human gut microbiota. We retrieved 42,236,858 proteins from the ESM Metagenomic Atlas (v2023_02) via Foldseek and performed a homology search of the metagenomic proteins against UniRef90 (release 2025_01) [6] using Diamond BLASTp (v2.0.9) [43] with default parameters. Afterward, the proteins were treated as functional unknowns when the homology search (i) yielded no hits, (ii) aligned to "uncharacterized" or "hypothetical" proteins, or (iii) aligned to proteins without any description, which resulted in 13,902,058 proteins of unknown function. We also identified 1,255 orphan enzymes that correspond to enzymatic reactions that are not associated with sequence information on the basis of the correspondence of each entry with KEGG or UniProt information provided by EnteroPathway. For each reaction, the molecular structures of all the substrates and products were retrieved and converted to reaction SMILES representations; 1,150 orphan enzymes for which complete reaction SMILES representations could be generated were retained for downstream analysis.

We applied DeepRES to the 13.9M proteins of unknown function and 1,150 orphan enzymes. In addition, for

444 proteins predicted to be enzymes by EnzymeCNN yet not linked to any orphan enzyme, we assigned functional
445 annotations by mapping them to the known enzymatic reactions in Rhea (release 138) [25] using EnzymeCLIP.
446 Here, the top 10% border of scores for positive examples in the test dataset, as determined during EnzymeCLIP
447 development, was employed as the threshold for the EnzymeCLIP score to consider an enzyme–reaction pair to
448 be linked.
449

450 Exploring novel biosynthetic gene clusters

451 Novel enzyme candidates identified by DeepRES could be utilized as a reference dataset for functional
452 annotation in genome analysis. Therefore, for each orphan enzyme for which we identified candidate proteins
453 by DeepRES, we mapped all genes in 4,744 representative MAGs of human gut bacteria obtained from MGnify
454 (Unified Human Gastrointestinal Genome v2.0.1) [26] to the reference dataset using Diamond BLASTp (v2.0.9)
455 [43]. Previous studies have demonstrated that leveraging a combination of metabolic pathway adjacency and
456 genomic neighborhood information is effective for identifying orphan enzyme-encoding genes [3,27].
457 Following this finding, we explored novel BGCs by assessing whether the candidate genes are densely
458 populated in microbial genomes for EnteroPathway modules, which are composed mostly of orphan enzymes.
459 Here, we specifically investigated novel BGC candidates for 20 modules (Table 2) that consisted of more than
460 10 reactions, more than 90% of which were orphan, and DeepRES identified candidate proteins for more than
461 50% of the orphan enzymes. The criterion for the novel BGCs was that at least five homologs of the orphan
462 enzymes in the module were included in the range of 10 continuous genes in the genome.

References

1. Kanehisa M, Furumichi M, Sato Y, Matsuura Y, Ishiguro-Watanabe M. KEGG: biological systems database as a model of the real world. Nucleic Acids Res. 2025;53:D672–7.

2. Shiroma H, Darzi Y, Terajima E, Nakagawa Z, Tsuchikura H, Tsukuda N, et al. Enteropathway: the metabolic pathway database for the human gut microbiota. Brief Bioinform. 2024;25:bbae419.

3. Hirota K, Salim F, Yamada T. DeepES: deep learning-based enzyme screening to identify orphan enzyme genes. Cowen L, editor. Bioinformatics. 2025;41:btaf053.

4. Lespinet O, Labedan B. Orphan Enzymes? Science. 2005;307:42–42.

5. Pouliot Y, Karp PD. A survey of orphan enzyme activities. BMC Bioinformatics. 2007;8:244.

6. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res. 2025;53:D609–17.

7. Zhang Z, Li C. Enzyme annotation for orphan reactions and its applications in biomanufacturing. Green Chem Eng. 2023;4:137–45.

8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

9. Yu T, Cui H, Li JC, Luo Y, Jiang G, Zhao H. Enzyme function prediction using contrastive learning. Science. 2023;379:1358–63.

10. Sanderson T, Bileschi ML, Belanger D, Colwell LJ. ProteInfer, deep neural networks for protein functional inference. Dötsch V, Staller MV, editors. eLife. 2023;12:e80942.

11. Kim GB, Kim JY, Lee JA, Norsigian CJ, Palsson BO, Lee SY. Functional annotation of enzyme-encoding genes using deep learning with transformer layers. Nat Commun. 2023;14:7370.

12. Buton N, Coste F, Le Cunff Y. Predicting enzymatic function of protein sequences with attention. Bioinformatics. 2023;39:btad620.

13. Tan Q, Xiao J, Chen J, Wang Y, Zhang Z, Zhao T, et al. ifDEEPre: large protein language-based deep learning enables interpretable and fast predictions of enzyme commission numbers. Brief Bioinform. 2024;25:bbae225.

14. Ayres G, Munsamy G, Heinzinger M, Ferruz N, Yang K, Bergman B, et al. Annotating the microbial dark matter with HiFi-NN. iScience. 2025;28:112480.

15. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.

16. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630:493–500.

17. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024;42:243–6.

18. Jiao P, Wang B, Wang X, Liu B, Wang Y, Li J. Struct2GO: protein function prediction based on graph pooling algorithm and AlphaFold2 structure information. Bioinformatics. 2023;39:btad637.

19. Song Y, Yuan Q, Chen S, Zeng Y, Zhao H, Yang Y. Accurately predicting enzyme functions through

geometric graph learning on ESMFold-predicted structures. Nat Commun. 2024;15:8180.
20. Moriya Y, Yamada T, Okuda S, Nakagawa Z, Kotera M, Tokimatsu T, et al. Identification of Enzyme Genes Using Chemical Structure Alignments of Substrate–Product Pairs. J Chem Inf Model. 2016;56:510–6.
21. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning Transferable Visual Models From Natural Language Supervision. Proc 38th Int Conf Mach Learn. PMLR; 2021. p. 8748–63.
22. Rodríguez del Río Á, Giner-Lamia J, Cantalapiedra CP, Botas J, Deng Z, Hernández-Plaza A, et al. Functional and evolutionary significance of unknown genes from uncultivated taxa. Nature. 2024;626:377–84.
23. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379:1123–30.
24. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026–8.
25. Bansal P, Morgat A, Axelsen KB, Muthukrishnan V, Coudert E, Aimo L, et al. Rhea, the reaction knowledgebase in 2022. Nucleic Acids Res. 2022;50:D693–700.
26. Richardson L, Allen B, Baldi G, Beracochea M, Bileschi ML, Burdett T, et al. MGnify: the microbiome sequence data analysis resource in 2023. Nucleic Acids Res. 2023;51:D753–9.
27. Yamada T, Waller AS, Raes J, Zelezniak A, Perchat N, Perret A, et al. Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours. Mol Syst Biol. 2012;8:581.
28. Du L, Ding X, Tian Y, Chen J, Li W. Effect of anthocyanins on metabolic syndrome through interacting with gut microbiota. Pharmacol Res. 2024;210:107511.
29. Liang A, Leonard ,William, Beasley ,Jesse T., Fang ,Zhongxiang, Zhang ,Pangzhen, and Ranadheera CS. Anthocyanins-gut microbiota-health axis: A review. Crit Rev Food Sci Nutr. 2024;64:7563–88.
30. Kapoor P, Tiwari A, Sharma S, Tiwari V, Sheoran B, Ali U, et al. Effect of anthocyanins on gut health markers, Firmicutes-Bacteroidetes ratio and short-chain fatty acids: a systematic review via meta-analysis. Sci Rep. 2023;13:1729.
31. Yoshikai Y, Mizuno T, Nemoto S, Kusuhara H. Difficulty in chirality recognition for Transformer architectures learning chemical structures from string representations. Nat Commun. 2024;15:1197.
32. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. Proc 37th Int Conf Mach Learn. PMLR; 2020. p. 1597–607.
33. Pham H, Dai Z, Ghiasi G, Kawaguchi K, Liu H, Yu AW, et al. Combined scaling for zero-shot transfer learning. Neurocomputing. 2023;555:126658.
34. Gao Y, Liu J, Xu Z, Wu T, Zhang E, Li K, et al. SoftCLIP: Softer Cross-Modal Alignment Makes CLIP Stronger. Proc AAAI Conf Artif Intell. 2024;38:1860–8.
35. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data. 2021;8:53.
36. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. Proc 4th Int Conf Learn Rep. ICLR 2016.
37. Su J, Han C, Zhou Y, Shan J, Zhou X, Yuan F. SaProt: Protein Language Modeling with Structure-aware Vocabulary. Proc 12th Int Conf Learn Rep. ICLR 2024.

38. Schwaller P, Probst D, Vaucher AC, Nair VH, Kreutter D, Laino T, et al. Mapping the space of chemical reactions using attention-based neural networks. Nat Mach Intell. 2021;3:144–52.
39. Hu EJ, shen yelong, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-rank adaptation of large language models. Proc 10th Int Conf Learn Rep. ICLR 2022
40. Schmirler R, Heinzinger M, Rost B. Fine-tuning protein language models boosts predictions across diverse tasks. Nat Commun. 2024;15:7407.
41. Goel S, Bansal H, Bhatia S, Rossi RA, Vinay V, Grover A. CyCLIP: Cyclic Contrastive Language-Image Pretraining. Preprint at http://arxiv.org/abs/2205.14459 (2022)
42. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, et al. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem. 2004;47:1750–9.
43. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18:366–8.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Source codes for the DeepRES are available on GitHub under the MIT license (https://github.com/yamadalab/deepres). Model weights (EnzymeCNN and EnzymeCLIP), EnteroPathway reaction information, the homology search results for ESM Metagenomic Atlas against UniRef90, and the results of DeepRES application to ESM Metagenomic Atlas, EnteroPathway and Rhea are accessible through Zenodo (https://doi.org/10.5281/zenodo.16347933).

Competing interests

T.Y. is a founder of Metagen Inc., Metagen Therapeutics Inc., and digzyme Inc. Metagen Inc. focuses on the design and control of the gut environment for human health. Metagen Therapeutics Inc. focuses on drug discovery and development which utilizes microbiome science. All of the companies had no control over the interpretation, writing, or publication of this work. The terms of these arrangements are being managed by the Institute of Science Tokyo in accordance with its conflict of interest policies.

Funding

This work has been supported by the grants from the Japan Society for the Promotion of Science (KAKENHI JP16H06279 (PAGS) to T.Y. and from JST SPRING, Japan Grant Number JPMJSP2180 to K.H.

Authors' contributions

K.H. and T.Y. contributed to the conception and design of the study. K.H. carried out the development of the deep learning models, the genomic analyses, and creating the python library. K.H. wrote the manuscript with contributions from T.Y. T.Y. supervised the study. The authors read and approved the final manuscript.

Acknowledgements

We thank Felix Salim for fruitful discussions. This study was carried out using the TSUBAME4.0 supercomputer at Institute of Science Tokyo.