
★ Supplemental Document

Enzymatic Reaction Recommendation with Interpretable Positive-unlabeled Learning

Zhenkun Shi^{1(✉)}, Jun Zhu², Dehang Wang², Qianqian Yuan¹, Zhitao Mao¹, Ruoyu Wang¹, Haoran Li¹, Weining Wu^{2(✉)}, Xiaoping Liao^{1(✉)}, Hongwu Ma^{1(✉)}

¹Biodesign Center, Key Laboratory of Engineering Biology for Low-carbon Manufacturing, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, 300308, Tianjin, China.

²College of Biotechnology, Tianjin University of Science & Technology, Tianjin, China.

✉zhenkun.shi@tib.cas.cn

Abstract

This is a supplementary document for the paper "RXNRECCer: Active Learning with Protein Language Models for Fine-Grained Enzyme Reaction Prediction". It provides details on preparing the data, selecting models, fine-tuning parameters, performance evaluation, as well as supplementary figures and tables that provide experimental details and support our conclusions. It also includes information on how to use the web service and APIs for high-throughput enzymatic reaction mining.

SI Related Work

Protein function annotation remains a central challenge in computational biology. Among various approaches, Enzyme Commission (**EC**) number prediction has become a key proxy for inferring enzymatic function and catalytic activity directly from amino acid sequences. Over the years, a wide range of computational methods have been developed to tackle this task, ranging from rule-based systems and profile-based search engines to deep learning models.

In addition to supervised EC number prediction models, traditional homology-based methods such as multiple sequence alignment (MSA) remain widely used due to their simplicity and interpretability. In our study, we employ **DIAMOND**, a fast alignment tool, to retrieve functional annotations based on sequence similarity against curated enzyme databases.

More recently, **protein language models** have emerged as powerful tools for capturing contextual information in amino acid sequences. Models such as **UniRep**, **ESM**, and **T5** generate fixed-length embeddings that encode evolutionary and structural features. These embeddings can be used to assess similarity in a learned latent space, enabling alignment-free annotation strategies. In our work, we compare proteins by computing embedding distances to annotated entries in the database, thereby predicting enzymatic function in a nearest-neighbor fashion.

In the following subsections, we summarize five representative EC prediction methods used as baselines in this study. Subsequent sections will also detail the additional

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

homology- and embedding-based approaches (DIAMOND, UniRep, ESM, T5) integrated
22 into our framework for comprehensive benchmarking.
23

CatFam (2009)

CatFam [1] is a profile-controlled, sequence-based method designed for large-scale catalytic function annotation. It constructs enzyme-specific profiles using sequence alignments and applies profile-specific thresholds to control the false positive rate. This allows users to choose between high-precision annotation or exploratory discovery with better recall. Unlike global thresholding methods, CatFam ensures that each profile achieves consistent performance, which is crucial for genome-wide enzyme annotation. Benchmarking on three test sets and 13 bacterial genomes showed that CatFam outperforms earlier tools such as PRIAM in both precision and coverage. While it has lower precision than more recent deep learning methods, its recall remains competitive, and its source code is still actively maintained. We included CatFam as a baseline to represent early, profile-based approaches with tunable confidence levels.

PRIAM-V2 (2018)

PRIAM-V2 [2] is a rule-based method for automated EC number annotation, originally proposed in 2003 and later updated in version 2.0. It uses position-specific scoring matrices (PSSMs) and regular expression patterns tailored to each enzyme family, trained on the full Swiss-Prot database. PRIAM supports both single-sequence and genome-wide annotation, making it suitable for microbial genome studies. The method achieves high recall and is known for its interpretability and robustness in detecting remote homologs. The updated version includes improvements in scoring and coverage. Although it may have lower precision than modern learning-based methods, its code availability, established methodology, and reproducibility make it a strong classical baseline.

DeepEC (2019)

DeepEC [3] is a deep learning-based framework for fast and accurate EC number prediction. It employs a three-stage convolutional neural network (CNN) architecture. The first CNN layer performs enzyme/non-enzyme classification, the second predicts the third-level EC class, and the third layer completes the full four-level EC assignment. To enhance performance, DeepEC also incorporates homology analysis as an optional refinement step. The method is optimized for minimal computational cost and memory usage, making it suitable for high-throughput applications. DeepEC has demonstrated sensitivity to mutations in catalytic domains or binding residues, enabling fine-grained discrimination. Although the source code for training is not available, pretrained models are provided for local use. Due to its strong performance and practicality, we include DeepEC as a deep learning benchmark.

ECREcer (2022)

ECREcer [4], developed by our group, is a multitask deep learning framework designed for accurate and robust enzyme function prediction. It integrates state-of-the-art protein language models (e.g., ESM) to generate biologically meaningful embeddings from raw amino acid sequences. These embeddings are passed to a hierarchical bidirectional GRU network with attention, enabling the model to jointly perform enzyme/non-enzyme classification, multifunction enzyme detection, and EC number assignment. The multitask design allows the model to exploit shared biological signals across tasks. In

addition, ECRECer employs a penalty-based regularization strategy to mitigate negative transfer between tasks, and uses a greedy decoding mechanism optimized for the F1 score. Extensive benchmarks show that ECRECer consistently outperforms existing methods in both accuracy and robustness, particularly for novel or low-homology proteins. It is available as a high-throughput annotation tool via both a web server and a standalone package.

CLEAN (2023)

CLEAN [5] is a contrastive learning-based framework for EC number prediction, developed to improve annotation accuracy and generalization. It constructs a function-aware embedding space by minimizing the distance between protein sequences sharing the same EC number, while maximizing the distance between those with different functions. The model uses ESM-1b for initial sequence encoding, followed by a feedforward projection head for embedding refinement. During inference, the Euclidean distance between a query embedding and cluster centroids of known EC numbers is computed, and the closest EC numbers are returned as predictions. CLEAN further applies a statistical filtering strategy based on maximum separation margin and p-value thresholds to retain only high-confidence annotations. This method is particularly effective in predicting rare, multifunctional, or low-similarity enzymes. Due to its performance and interpretability, we selected CLEAN as a recent state-of-the-art baseline in our evaluation.

DIAMOND (2015)

DIAMOND [6] is a high-speed sequence alignment tool designed for large-scale protein searches. It adopts a seed-and-extend algorithm similar to BLAST but achieves several orders of magnitude faster performance, making it well-suited for high-throughput annotation tasks. In our study, DIAMOND is employed as a homology-based baseline: each query protein is aligned against curated reference databases annotated with either known EC numbers (MSA-via-EC) or known reactions (MSA-via-RXN), and functional annotations are transferred from the top-scoring hit. While alignment-based methods are inherently limited by sequence divergence and database coverage, they remain highly effective for conserved proteins and serve as a robust reference in comparative evaluations.

UniRep (2019)

UniRep [7] is one of the earliest protein language models trained on millions of sequences to learn informative fixed-length representations. Based on a multiplicative LSTM architecture, UniRep encodes sequence-level information without requiring explicit alignment or structural features. In our implementation, UniRep embeddings are computed for both query and reference proteins. Pairwise distances are then used to identify the most functionally similar entries, and EC numbers are transferred from the nearest neighbors. This approach provides an embedding-based alternative to traditional alignment, capturing subtle functional signals even in low-homology regions.

ESM (2021)

ESM (Evolutionary Scale Modeling) [8] is a family of transformer-based protein language models pretrained on large-scale sequence databases. ESM models, such as ESM-1b and ESM-2, achieve strong performance across various structural and functional prediction tasks. In this study, we extract residue-averaged ESM embeddings for all proteins and compute cosine distances between query proteins and entries in a reference EC-annotated dataset. Predictions are made based on the closest matches in embedding space. ESM

serves as a representative of modern transformer-based encoders and demonstrates
improved sensitivity to remote homologs compared to earlier models.

ProstT5 (2024)

ProstT5 [9] is a protein language model based on the T5 encoder–decoder architecture,
trained to translate between amino acid sequences and 3Di structure representations
derived from AlphaFold2-predicted structures. The model was fine-tuned using a large
non-redundant dataset of over 17 million proteins with high-quality 3Di annotations,
enabling it to learn joint representations of sequence and structure through bilingual
translation tasks. In our study, we use ProstT5 as a pretrained encoder to extract fixed-
length protein embeddings, which are then compared against a reference database for
annotation transfer via nearest-neighbor search. While ProstT5 has demonstrated strong
performance in structure-aware tasks such as remote homology detection and CATH
classification, we observed that it underperforms compared to our own fine-tuned ESM2
model when applied to reaction-level enzyme function prediction. We therefore include
ProstT5 primarily as a representative general-purpose PLM baseline to benchmark
against structure-augmented models.

S1 SI Appendix Materials and Methods

S1.1 Software and code availability

All software was coded in Python. The framework is implemented using the PyTorch
library, and parts of comparison baselines are implemented using scikit-learn. The
datasets, the results, and source codes used to reimplement our methods are publicly
available from <https://github.com/kingstdio/RXNRECCer>.

S1.2 Preprocessing

To collect high-confidence Reaction–Protein (R–P) pairs, we utilized two expert-curated
data sources with comprehensive coverage: SwissProt [10] and Rhea [11]. The data
preprocessing consists of six steps (s1–s6), as detailed in Algorithm 1:

Algorithm 1 Data Preprocessing

- | | |
|---|--------|
| 1: Download UniProt snapshot data and extract protein records | ▷ [s1] |
| 2: Retrieve protein–reaction relation pairs $\langle R, P \rangle$ from UniProt | ▷ [s2] |
| 3: Obtain curated reaction data from RHEA | ▷ [s3] |
| 4: Filter out protein records with sequence changes | ▷ [s4] |
| 5: Filter out protein records with sequence less than 10 amino acids | ▷ [s5] |
| 6: Remove reaction entries lacking valid $\langle R, P \rangle$ associations | ▷ [s6] |
| 7: Construct the benchmark dataset, denoted as ds_b | ▷ [s7] |

S1.3 Dataset

To comprehensively evaluate the performance and generalization ability of enzyme–reaction
prediction models, we constructed two datasets tailored for complementary evaluation
settings: a 10-fold cross-validation set (ds_{rcv}) for in-distribution assessment, and a
temporally independent test set (ds_{rcp}) for real-world generalization benchmarking.

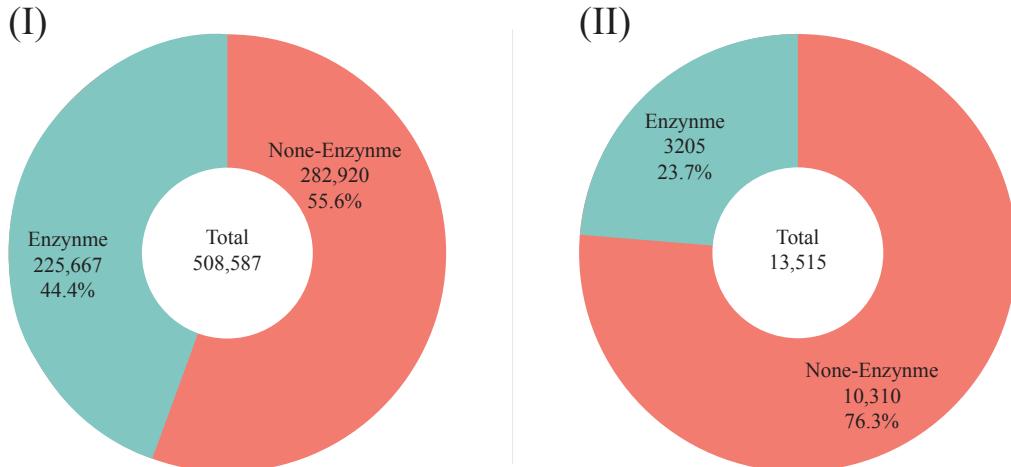
S1.3.1 Construction of *ds_rcv* Dataset

143

The *ds_rcv* dataset was derived from the UniProtKB/Swiss-Prot release of January 2018 and comprises 508,587 protein records, with 44.4% labeled as enzymes and 55.6% as non-enzymes (Fig. S1(I)). To ensure long-term consistency and avoid potential retrospective bias, we excluded entries whose sequences were later removed or modified in the January 2024 UniProt snapshot. This filtering step guarantees that the training data reflect only historical information available at the time of model development.

Additionally, protein sequences shorter than 10 amino acids were removed prior to training. These ultra-short sequences accounted for only 0.16% of the data and included a negligible proportion (0.76%) of enzymatic annotations (Fig. S2). Due to their lack of structural and catalytic features, such sequences are generally regarded as biologically irrelevant for enzyme function prediction tasks [12].

After preprocessing, the cleaned dataset was randomly partitioned into 10 equal folds for cross-validation. In each fold, 90% of the data were used for training and 10% for testing. This setup enables reproducible and balanced evaluation across a wide range of enzyme function prediction methods, including EC-based tools and end-to-end PLM-based models.



Supplementary Figure S1. Proportions of enzyme and non-enzyme sequences in the *ds_rcv* dataset (I) and the *ds_rcp* test set (II).

S1.3.2 Construction of *ds_rcp* Dataset

160

To assess real-world generalization, we constructed an independent test set, *ds_rcp* (Recently Curated Proteins), consisting of 13,515 proteins newly added to UniProtKB/Swiss-Prot between January 2018 and January 2024. Proteins overlapping with *ds_rcv* were excluded to ensure strict temporal independence, and any reaction annotations previously seen in *ds_rcv* were also removed to maintain reaction-level novelty.

Compared to the cross-validation set, *ds_rcp* represents a more challenging and diverse benchmark. As shown in Fig. S1(II), enzyme annotations are substantially less prevalent in this dataset (23.7% vs. 44.4%), better reflecting the enzyme/non-enzyme imbalance encountered in real biological discovery scenarios. Furthermore, the sequence similarity between training and testing proteins is significantly reduced. Only 19.98% of sequences in *ds_rcp* share more than 80% identity with any training sequence from *ds_rcv*, in contrast to 46.21% in *ds_rcv* itself (Fig. S3). The coverage of known enzyme classes is also lower: only 16.44% of enzyme-labeled proteins in *ds_rcp* belong to EC categories

161

162

163

164

165

166

167

168

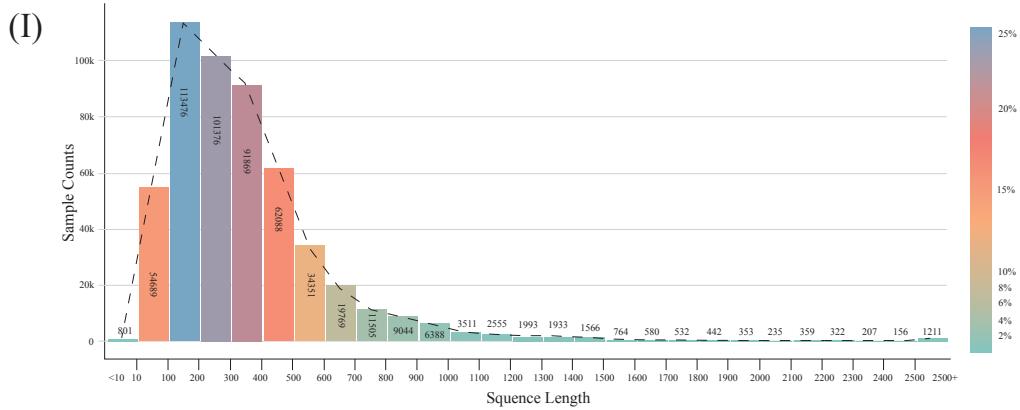
169

170

171

172

173



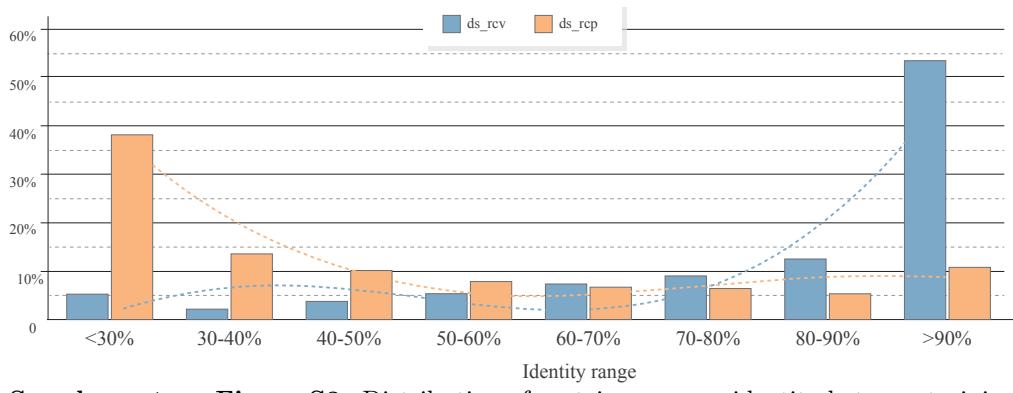
Supplementary Figure S2. Distribution of protein sequence lengths in the *ds_rcv* dataset. Sequences shorter than 10 amino acids were excluded during preprocessing.

present in *ds_rcv*, compared to 66.14% within *ds_rcv*. These differences highlight the increased novelty and distribution shift of *ds_rcp*.

To ensure a fair and consistent comparison with state-of-the-art baselines such as DeepEC, CLEAN, and ECRECer—which were all trained on protein databases released prior to February 2018—we applied the same temporal cutoff when building *ds_rcv*. This ensures that evaluation on *ds_rcp* is free from information leakage and serves as a realistic proxy for future deployment scenarios, where models must generalize to novel, low-homology enzymes. Full statistics for both datasets are presented in Table S1.

Supplementary Table S1. Summary of training and test set statistics across 10 folds of *ds_rcv* and the independent test set *ds_rcp*.

Dataset	Total	Non-Enzyme	Enzyme	#Reactions	Unique P–R Pairs
ds_rcv_fold1_train	457,729	254,621	203,108	10,208	273,904
ds_rcv_fold1_test	50,858	28,299	22,559	4,398	30,109
ds_rcv_fold2_train	457,729	254,584	203,145	10,186	273,396
ds_rcv_fold2_test	50,858	28,336	22,522	4,690	30,617
ds_rcv_fold3_train	457,729	254,843	202,886	10,205	273,386
ds_rcv_fold3_test	50,858	28,077	22,781	4,544	30,627
ds_rcv_fold4_train	457,729	254,616	203,113	10,200	273,495
ds_rcv_fold4_test	50,858	28,304	22,554	4,559	30,518
ds_rcv_fold5_train	457,729	254,703	203,026	10,233	273,516
ds_rcv_fold5_test	50,858	28,217	22,641	4,560	30,497
ds_rcv_fold6_train	457,729	254,610	203,119	10,205	273,909
ds_rcv_fold6_test	50,858	28,310	22,548	4,448	30,104
ds_rcv_fold7_train	457,729	254,484	203,245	10,191	273,925
ds_rcv_fold7_test	50,858	28,436	22,422	4,537	30,088
ds_rcv_fold8_train	457,729	254,516	203,213	10,210	273,804
ds_rcv_fold8_test	50,858	28,404	22,454	4,514	30,209
ds_rcv_fold9_train	457,729	254,618	203,111	10,236	273,816
ds_rcv_fold9_test	50,858	28,302	22,556	4,398	30,197
ds_rcv_fold10_train	457,729	254,507	203,222	10,200	273,677
ds_rcv_fold10_test	50,858	28,413	22,445	4,581	30,336
ds_rcp	13,515	10,310	3,205	2,185	5,795



Supplementary Figure S3. Distribution of protein sequence identity between training and test sets in the two datasets. The *ds_rcv* dataset (blue) was constructed via random partitioning, while the *ds_rcp* dataset (orange) uses a temporal split. Identity scores reflect the highest sequence identity between each test protein and any training sequence.

S2 Models

182

S2.1 Optimized Parameters for RXNRECCer-S1

183

The RXNRECCer-S1 model combines a pretrained protein language model with a lightweight neural classifier head for multi-label reaction prediction. It is specifically designed to balance representation power and computational efficiency. The architecture leverages deep sequence embeddings from ESM2 and integrates a GRU–attention mechanism to capture both global dependencies and localized motifs relevant to catalytic activity. The full parameter breakdown is shown below:

# RXNRECCer-S1		
Layer (type)	Output Shape	Param #
SeqEmbedding	(L, 1280)	197M
bi-GRU	(L, 1024)	10,526,720
Transformer	(1024)	1,050,625
Dense	(64)	65,600
FC	(num_of_labels)	64 x N
<hr/>		
Total trainable parameters:		~208.6M
Frozen ESM2 layers (layers 0-22):		~453M
<hr/>		

Protein sequences are first encoded using the pretrained ESM2-650M model, which outputs residue-level embeddings of 1280 dimensions. To reduce computational overhead and prevent overfitting, we freeze the first 23 Transformer layers and fine-tune only the last 10 layers (layers 23–32). This partial fine-tuning retains general sequence knowledge while enabling task-specific adaptation.

The GRU layer processes the sequence embeddings bidirectionally to model long-range dependencies, outputting a 1024-dimensional vector per sequence. A self-attention mechanism then aggregates this variable-length sequence into a fixed-size embedding by learning position-aware attention weights. The aggregated vector passes through a two-layer classifier: a fully connected layer ($1024 \rightarrow 64$) with nonlinearity, followed by a final output layer mapping to the reaction label space ($64 \rightarrow 10479$). Dropout with $p = 0.2$ is applied after the GRU to regularize training.

This streamlined yet expressive architecture enables RXNRECCer-S1 to achieve high annotation accuracy with reduced model complexity, making it suitable for large-scale proteome-level reaction prediction tasks.

S2.1.1 Effect of ESM2 Layer Freezing Depth

221

To evaluate the optimal fine-tuning strategy, we conducted ablation experiments by freezing different numbers of layers in the ESM2-650M model. Specifically, we compared the performance of models with 10, 5, and 1 unfrozen Transformer layers. All models were trained for 20 epochs with an 8:2 train-validation split. The table below summarizes performance on the validation set using mean accuracy (mAccuracy), mean precision (mPrecision), mean recall (mRecall), and mean F1 score (mF1).

Supplementary Table S2. Validation performance under different ESM2 freezing strategies. Best results highlighted in bold.

Frozen Layers	mAccuracy	mPrecision	mRecall	mF1
Layers 0–22 (10 unfrozen)	0.9719	0.9747	0.7749	0.7660
Layers 0–27 (5 unfrozen)	0.9272	0.9829	0.6420	0.6504
Layers 0–31 (1 unfrozen)	0.9629	0.9753	0.6864	0.6783

These results demonstrate that freezing the first 23 layers and fine-tuning the remaining 10 achieves the best overall performance. Unfreezing fewer layers leads to weaker recall and F1, while unfreezing only the final layer significantly limits representation learning capacity.

S2.2 Prompts

S2.2.1 Prompt 1: Obtaining Evidence for Predictions

Prompt 1: Evidence Seeking

You are a biochemical expert. Given a protein sequence and its predicted catalyzed-reaction results (in JSON format)—where “-” denotes a non-catalytic protein and a list denotes predicted reactions—provide concise evidence and interpretation for each prediction. You may consult UniProt, BRENDA, KEGG, ExPASy, and other authoritative sources.

Key considerations:

1. Sequence-Level Annotation: (Performed upstream; do not reiterate these steps in your output.)
2. Non-Catalytic Predictions: “-” indicates no catalytic activity. Provide direct evidence (e.g., absence of catalytic motifs).
3. Multiple Reactions: If multiple reactions are predicted, furnish evidence for each.
4. Ranking: Order predicted reactions by likelihood (1 = most likely).
5. Justification: Cite conserved domains, active-site residues, structural motifs, homology, or known mechanisms. Do NOT describe the analysis pipeline (BLAST, Pfam, motif scans); assume it has already been done upstream.
6. Confidence Score: Assign a score [0.0-1.0] reflecting how strongly sequence features support each reaction. If “-” is selected, give it the highest confidence.
7. Literature References: If you invoke published studies or databases as evidence, cite real, verifiable references (authors, journal, year, DOI). Do not fabricate citations.

Output (strict JSON using only the key “results”):

```
[  
  {  
    "reaction_id": "RHEA:XXXXX",  
    "confidence": 0.00-1.00,  
    "reason": "Concise evidence summary (no pipeline details)." }  
]
```

228
229
230
231

232
233

234

S2.2.2 Prompt 2: Fine-grained Results Ranking and Justification for Proteins Integrated TrEMBL

235

236

Prompt 2: Results Ranking and Justification

You are a biochemical expert. Given a protein sequence and a set of candidate catalyzed reactions (in JSON format), analyze and determine which reaction(s) are most likely catalyzed by the protein. You may consult relevant biochemical databases including *UniProt*, *BRENDA*, and *ExPASy* to retrieve supporting information.

Key considerations:

1. Non-Catalytic Cases:

- A reaction represented by a single dash (“-”) indicates a predicted lack of catalytic activity.
- If all reactions are marked as “-”, or none of the valid reactions are plausible, select the dash (“-”) reaction to indicate non-catalytic function.

2. Multiple Reactions:

- Proteins may have multiple active sites. Select more than one reaction if supported by sequence features.

3. Ranking:

- Assign a likelihood-based ranking to selected reactions (1 = most likely).

4. Justification:

- Provide a brief explanation for each decision, referencing conserved domains, active site residues, structural motifs, or known catalytic mechanisms.

5. Confidence Score:

- Assign a confidence score between 0 and 1 based on how well the protein sequence aligns with known catalytic features for each reaction.
- If no valid reaction is likely, assign the highest confidence to the dash (“-”) reaction.

Output format (strictly JSON):

```
[  
  {  
    "reaction_id": "xxx",  
    "selected": "yes" or "no",  
    "rank": <integer>,      // Only for selected = "yes"  
    "confidence": <float>, // Between 0 and 1  
    "reason": "Explanation based on sequence and reaction data."  
  },  
  ...  
]
```

▲ Only use the key *“results”* in the output. Do not include any extra explanation or metadata.

237

S2.2.3 Prompt 3: Fine-grained Results Ranking and Justification for Proteins Not Integrated UniProt

238

239

Prompt 3: Sequence-Based Reaction Prediction with Fallback Analysis

You are a biochemical expert. Given a protein sequence (with or without a UniProt ID) and a set of candidate catalyzed reactions (in JSON format), analyze and determine which reaction(s) are most likely catalyzed by the protein. You may consult relevant biochemical databases including *UniProt*, *BRENDA*, and *ExPASy* to retrieve supporting information.

Key considerations:

1. ID-Based Annotation:

- If a UniProt ID is provided, first retrieve functional annotations and known EC numbers from UniProt.

2. Fallback Sequence Analysis:

- If no UniProt ID is available, automatically perform:
 - (a) A BLAST (or DIAMOND) search against UniProt/NCBI NR, extracting the top 5 homologs and their EC/functional annotations.
 - (b) A Pfam/InterProScan domain scan to identify characteristic catalytic domains (e.g., PLP-dependent aminotransferase family).
 - (c) A motif scan for key active-site residues (e.g., PLP-binding lysine motifs, signature sequence patterns).

3. Non-Catalytic Cases:

- A reaction represented by a single dash ("–") indicates a predicted lack of catalytic activity.
- If all valid reactions are implausible after ID-based or fallback analysis, select the dash ("–") reaction.

4. Multiple Reactions:

- Proteins may have multiple active sites. Select more than one reaction if supported by sequence features.

5. Ranking:

- Assign a likelihood-based ranking to selected reactions (1 = most likely).

6. Justification:

- Provide a brief explanation for each decision, referencing database annotations, conserved domains, active-site residues, structural motifs, or known catalytic mechanisms.

7. Confidence Score:

- Assign a confidence score between 0 and 1 based on how well the sequence and analysis support each reaction.
- If no valid reaction is likely, assign the highest confidence to the dash ("–") reaction.

Output format (strictly JSON):

```
[  
  {"reaction_id": "xxx",  
   "selected": "yes" or "no",  
   "rank": <integer>,      // Only if selected = "yes"  
   "confidence": <float>, // Between 0 and 1  
   "reason": "Explanation based on sequence and reaction data."},  
  ...  
]
```

▲ Only output the key “results” in the JSON array. Do not include any extra explanation or metadata.

240

S2.3 Input Examples

241

S2.3.1 Sequence can be provided with a UniProt ID (TrEMBL)

242

Input Example 1: Protein sequence with UniProt ID

INPUT:

```
[  
 {  
   "protein information": {  
     "uniprot id": "Q9VK50",  
     "protein amino acid sequence": "MPRSLGNQCLNYSKERMWEPGYLKAKCAELRLESE  
FRLYRIRLWKSYLLTFFMLHIFVTSVHCALLLATIERRSIIYFDVALSIGCALVLILVLSVNDEF  
IAKHTWYMYASSIFASLTFLVFAADTESIYHTYAHSWILGTFYDTYIIYMIYMFLPIHFISGAVLLA  
LLVSGLYIILYFVIFIAQGFAQFASALFSVGGMSPDIVHYIICLNIVGIFYRVMNDTVVRSSFLDRHQ  
YIKEKIWLNRNARLQEKKQLLDSILPPQISLPLQKDIQGRIVMAKQGIHSWTAMERTMAIQIHDPVSI  
LYADVVNYTHLTTLTVEMLVKVLHDLYGRFDLAAYRYKVQRIKFLGDCYYCAGLSDPDPDHANN  
CVILGLSMINHIMEVRDIHGLDINMRIGVHSGNLFAGVIGEAKLQFDIWGLDVTIANVLESTGVPG  
CVHISGATLNNLDVNRFDIEDGPEEARHPPLLKYRIRSYIIRQDLHMDDEDSDEFLGDLHSISLC  
NMGAQPRISDSANQSLRALFHEELREFRKMPVSAFSPKRLLGICRFNTGKEVPAHQNLNICLTFT  
DPLIERAYLKQTDYMYKYSIILSASVGCSLVYIELMDTQMICSSCFVLPASVATIQCILALIAWYK  
ILNNILPSHIVDVYLNLSAKHELYFENYRMVSVMFAMLINFEMDLRSLRVLNNEIIAEFDTLLFYK  
RSTVISLQSTYTHAENNNSIASTSRTLQSL"},  
   "reaction information": {  
     "candidate reaction 1": {  
       "reaction id": "RHEA:15389",  
       "reaction equation": "ATP = 3',5'-cyclic AMP + diphosphate",  
     },  
     "candidate reaction 2": {  
       "reaction id": "RHEA:13665",  
       "reaction equation": "GTP = 3',5'-cyclic GMP + diphosphate"}  
   }  
 }]
```

243

S2.3.2 Sequence can be provided with a UniProt ID (in Lab)

244

Input Example 2: Protein sequence without UniProt ID

INPUT:

```
{  
   "protein information": {  
     "uniprot id": "-",  
     "protein amino acid sequence": "MGSQETNLPPHVLIIFPLPIQGLDITFIVS..."},  
   "reaction information": {  
     "candidate reaction 1": {  
       "reaction id": "RHEA:39895",  
       "reaction equation": "7-deoxyloganetate + UDP-alpha-D-glucose  
= 7-deoxyloganate + H(+) + UDP",  
       "reaction equation in ChEBI format": "CHEBI:76846 +  
CHEBI:58885 = CHEBI:76844 + CHEBI:15378 + CHEBI:58223",  
       "reaction equation in SMILES format": "[C][C@H]1CC...",  
       "reaction associated Enzyme Commission Number": "  
EC:2.4.1.323"},  
     "candidate reaction 2": {  
       "reaction id": "-",  
       "reaction equation": "-"}  
   }  
 }
```

245

S2.4 Cases for GPT

246

S2.4.1 GPT Case 1: Provide Evidence for None-enzyme Prediction

247

Example: without UniProt ID

► Query: Prompt 1 +

```
{  
    "protein information": {  
        "uniprot id": "-",  
        "protein amino acid sequence": "MAKPILLSIYLCLIIIVALFNGCLAQSGGRQQHKFGQ  
CQLNRLDALERTNRIEAEAGVIESWDPNHQQLQCAGVAVVRRTIEPNGLLLPHYSNAPQLVYIARGRGIT  
GVLFPGCPETFEESQRQSQQGQRREFQQDRHQKIRHFREGDIIIAFPAGVAHWCYNDGSSPVVAIFLLDTH  
NNANQLDQNPRNFYLAGNPDEFRPQGQQYEQHRRQQHQQRGEHGEQQRDLGNNVSGFDAEFLADA  
FNVDTETARRLQSENDHRGSIVRVEGRQLQVIRPRWSREEQEHEERKERERERESESERRQSRGGRRDDN  
GLEETICTLSRENIGDPSRADIYTEEAGRISTVNSHNLPILRWLQLSAERGALYSDALYVPHWNLNAHS  
VVYALRGRAEVQVVDFGQTVFDELREGQLLTIPQNFAVVKRARDEGF EWVSFKTNENAMVSPLAGRTS  
AIRALPEEVLVNAFQIPREDARRLKFN RQESTLVR SRSSRSERRA EV"  
    },  
    "reaction information": {  
        "predicted reaction 1": {  
            "reaction id": "-",  
            "reaction equation": "-"  
        }  
    }  
}
```

► Answer ✓

```
{  
    "results": [  
        {  
            "reaction_id": "-",  
            "confidence": 1.0,  
            "reason": "The protein sequence lacks conserved catalytic motifs,  
active-site residues, or domains characteristic of known  
enzymes. No evidence of catalytic function is present based on  
sequence features (e.g., absence of serine protease, kinase,  
or oxidoreductase signatures)."  
        }  
    ]  
}
```

248

S2.4.2 GPT Case 2: Provide Evidence for Enzyme Prediction From TrEMBL. 249

Example: with UniProt ID

► Query: Prompt 1 +

```
{
    "protein information": {
        "uniprot id": "A0A1V1IGJ0",
        "protein amino acid sequence": "MKFPGKRKS KHYFPVNARDPLLQQFQPE..."
    },
    "reaction information": {
        "candidate reaction 1": {
            "reaction id": "RHEA:27710",
            "reaction equation": "ATP + guanosine = ADP + GMP + H(+)",
            "reaction equation in ChEBI format": "CHEBI:30616 + CHEBI:
                16750 = CHEBI:456216 + CHEBI:58115 + CHEBI:15378",
            "reaction equation in SMILES format": "Nc1ncnc2n(cnc12) [C@@H]
                ]1O[C@H] (COP([O-]) (=O) OP([O-]) (=O) OP([O-]) ([O-]) =O) [C@@H
                ](O) [C@H]1O.Nc1nc2n(cnc2c(=O) [nH]1) [C@@H]1O[C@H] (CO) [
                C@@H] (O) [C@H]1O>>Nc1ncnc2n(cnc12)...",
            "reaction associated EC": "EC:2.7.1.73"
        },
        "candidate reaction 2": {
            "reaction id": "RHEA:21140",
            "reaction equation": "ATP + inosine = ADP + H(+) + IMP",
            "reaction equation in ChEBI format": "CHEBI:30616 + CHEBI:
                17596 = CHEBI:456216 + CHEBI:15378 + CHEBI:58053",
            "reaction equation in SMILES format": "Nc1ncnc2n(cnc12) [C@@H]
                ]1O[C@H] (COP([O-]) (=O) OP([O-]) (=O) OP([O-]) ([O-]) =O) [C@@H
                ](O) [C@H]1O.OC[C@H]1O[C@H] ([C@H] (O) [C@@H]1O)n1ncnc2c(O)
                ncnc12>>Nc1ncnc2n(cnc12) [C@@H]1O[C@H] (COP([O-]) (=O) OP([O-
                -]) ([O-]) =O) [C@H] (O) [C@H]1O.[H+].O[C@@H]1[C@@H] (COP([O-
                -]) ([O-]) =O) O[C@H] ([C@@H]1O)n1ncnc2c1nc[nH]c2=O",
            "reaction associated EC": "EC:2.7.1.73"
        }
    }
}
```

► Answer ✓

```
{
    'results': [
        {'reaction_id': 'RHEA:27710',
            'confidence': 0.95,
            'reason': 'The protein sequence contains conserved motifs
                characteristic of nucleoside kinases, including the GXGXXG ATP-
                binding motif and residues associated with guanosine kinase
                activity. Homology to characterized guanosine kinases (EC
                2.7.1.73) is high (see UniProt Q9HOU4, PDB: 2Q0U), and the
                substrate specificity loop matches those known to prefer
                guanosine. Literature supports this function in similar bacterial
                proteins (Bzowska et al., Acta Biochim Pol, 2000, PMID:
                11079780).'},
        {'reaction_id': 'RHEA:21140',
            'confidence': 0.8,
            'reason': 'The sequence also shares homology with nucleoside kinases
                capable of phosphorylating inosine, but key substrate recognition
                residues more closely match guanosine-specific kinases. Some
                guanosine kinases show minor activity toward inosine (Bzowska et
                al., 2000), but this is less favored due to differences in the
                binding pocket.'}
    ]
}
```

S2.4.3 GPT Case 3: Provide Evidence for Enzyme Prediction without UniProt ID.

251
252

Example: without UniProt ID

► Query: Prompt 1 +

```
{  
    "protein information": {  
        "uniprot id": "-",  
        "protein amino acid sequence": "MLHILCQGTPFEIGYEHGSAAKAVIARSIDFAVDLIR  
GKTKKTDEELKQVLSQLGRVIEERWPKYEEIRGIAKGAERDVSEIVMLNTRTEFAYGLKAARDGCTTAY  
CQLPNGALQGQNWDFFSATKENLIRLTIRQAGLPTIKFITEAGIIGKVGFNSAGVAVVNLYNALHLQGLRPT  
GVPSHIALRIAESTSPSQAYDRIVEQGGMAASAFIMVGNNGHEAFGLEFSPTSIRKQVLDANGRMVHTNH  
CLLQHGKNEKEILDPLPDSWRHQRMEFLDGFDGTKQAFQLWADEDNYPPSICRAYEEGKSRGATLFNI  
IYDHARREATVRLGRPTNPDEMFVMRFDEEDERSALNARL"  
    },  
    "reaction information": {  
        "predicted reaction 1": {  
            "reaction id": "RHEA:20720",  
            "reaction equation": "H2O + isopenicillin N + phenylacetyl-CoA =  
CoA + H(+) + L-2-amino adipate + penicillin G",  
            "reaction equation in ChEBI format": "CHEBI:15377 + CHEBI:58399 +  
CHEBI:57390 = CHEBI:57287 + CHEBI:15378 + CHEBI:58672 + CHEBI:  
51354",  
            "reaction equation in SMILES format": "[H]O[H].[H][C@]12SC(C)(C)[  
C@@H](N1C(=O)[C@H]2NC(=O)CCC[C@H]([NH3+])C([O-])=O)C([O-])=O.CC(C)(  
COP([O-])(=O)OP([O-])(=O)OC[C@H]1O[C@H]([C@H](O)[C@@H]1OP([O-])([O-]  
)=O)n1cnc2c(N)ncnc12)[C@H](O)C(=O)NCCC(=O)NCCSC(=O)Cc1cccc1>>CC(  
C)(COP([O-])(=O)OP([O-])(=O)OC[C@H]1O[C@H]([C@H](O)[C@@H]1OP([O-])([O-]  
)=O)n1cnc2c(N)ncnc12)[C@H](O)C(=O)NCCC(=O)NCCS.[H+].[NH3+][C@H]  
[H](CCCC([O-])=O)C([O-])=O.[H][C@]12SC(C)(C)[C@@H](N1C(=O)[C@H]2NC(=  
O)Cc1cccc1)C([O-])=O",  
            "reaction associated Enzyme Commission Number": "EC:2.3.1.164"  
        }  
    }  
}
```

► Answer ✓

```
{  
    "results": [  
        {  
            "reaction_id": "RHEA:20720",  
            "confidence": 0.95,  
            "reason": "The protein sequence contains conserved motifs and  
active-site residues characteristic of acyltransferases  
involved in penicillin biosynthesis, specifically  
isopenicillin N acyltransferase (EC 2.3.1.164). Key residues (e.g., Ser, His, Asp triad) and the overall domain architecture  
match those found in characterized isopenicillin N  
acyltransferases (UniProt: P00811, PDB: 1QME). This supports  
the predicted reaction: conversion of isopenicillin N and  
phenylacetyl-CoA to penicillin G. See: Coque et al., J. Biol.  
Chem. 267: 6262-6270 (1992), PMID: 1556132; UniProt P00811."  
        }  
    ]  
}
```

253

S2.4.4 Case 4: Reaction-Aware Distillation for Fine-Grained Catalytic and Non-Catalytic Predictions

254
255

Example: Ground Truth | Non-Enzyme (No UniProt ID Available)

256

► Query Prompt 3 +

```
{  
    "protein information": {  
        "uniprot id": "-",  
        "protein amino acid sequence": "MTHREELLPPMKWDAWGDPAAKPLSD..."}  
        ,  
        "reaction information": {  
            "candidate reaction 1": {  
                "reaction id": "-",  
                "reaction equation": "-"  
            },  
            "candidate reaction 2": {  
                "reaction id": "RHEA:20049",  
                "reaction equation": "1-(5-phospho-beta-D-ribosyl)-5'-AMP +  
H2O = 1-(5-phospho-beta-D-ribosyl)-5-[(5-phospho-beta-D-  
ribosylamino)methylideneamino]imidazole-4-carboxamide",  
                "reaction equation in ChEBI format": "CHEBI:59457 + CHEBI:  
15377 = CHEBI:58435",  
                "reaction equation in SMILES format": "O[C@H]1[C@@H](O)[C@H](O[C@H]1COP([O-])([O-])=O)n1cnc2c1ncn([C@@H]1O[C@H](  
COP([O-])([O-])=O)[C@H](O)[C@H]1O)c2=N.[H]O[H]>>NC(=O)  
c1ncn([C@H]2O[C@H](COP([O-])([O-])=O)[C@H](O)[C@H]2O)  
c1\\N=C\\N[C@@H]1O[C@H](COP([O-])([O-])=O)[C@H](O)[C@H]1O",  
                "reaction associated EC Number": "EC:3.5.4.19"}  
            }  
        }  
}
```

► Answer ✓

```
{  
    'results': [  
        {  
            'reaction_id': '-',  
            'selected': 'yes',  
            'rank': 1,  
            'confidence': 0.95,  
            'reason': 'At ~120 aa the sequence is too small and lacks the  
            conserved amidase signature and catalytic Glu/His-Gly motifs  
            characteristic of EC 3.5.4.19 enzymes.'  
        },  
        {  
            'reaction_id': 'RHEA:20049',  
            'selected': 'no',  
            'confidence': 0.1,  
            'reason': 'No evidence of conserved domains or sequence motifs  
            associated with EC 3.5.4.19 (phosphoribosylformylglycinamidine  
            cyclo-ligase) is found in the provided sequence. The protein  
            does not align with known catalytic features for this reaction  
            .'  
        }  
    ]  
}
```

257

S2.4.5 Case 5: Ambiguous Justification with Mixed Catalytic and Non-Catalytic Predictions

258
259

Example Query: Ground Truth | Enzyme (No UniProt Annotation Available)

► Query Prompt 2 +

```
{  
    "protein information": {  
        "uniprot id": "-",  
        "protein amino acid sequence": "MSVESFERFLSRRVLSMQESQTMKITGLAKKMQA  
        ..."},  
    "reaction information": {  
        "candidate reaction 1": {  
            "reaction id": "RHEA:21824",  
            "reaction equation": "2-oxoglutarate + L-aspartate = L-glutamate  
            + oxaloacetate",  
            "reaction equation in ChEBI format": "CHEBI:16810 + CHEBI:29991  
            = CHEBI:29985 + CHEBI:16452",  
            "reaction equation in SMILES format": "[O-]C(=O)CCC(=O)C([O-])=O  
            . [NH3+] [C@@H] (CC([O-])=O)C([O-])=O>>[NH3+] [C@@H] (CCC([O-])=O  
            )C([O-])=O.[O-]C(=O)CC(=O)C([O-])=O",  
            "reaction associated Enzyme Commission Number": "EC:2.6.1.1"},  
        "candidate reaction 2": {  
            "reaction id": "RHEA:22880",  
            "reaction equation": "2-oxoglutarate + L-arogenate = L-glutamate  
            + prephenate",  
            "reaction equation in ChEBI format": "CHEBI:16810 + CHEBI:58180  
            = CHEBI:29985 + CHEBI:29934",  
            "reaction equation in SMILES format": "[O-]C(=O)CCC(=O)C([O-])=O  
            . [NH3+] [C@@H] (CC1(C=CC(O)C=C1)C([O-])=O)C([O-])=O>>[NH3+] [  
            C@@H] (CCC([O-])=O)C([O-])=O.O[C@H]1C=C[C@] (CC(=O)C([O-])=O)  
            (C=C1)C([O-])=O",  
            "reaction associated Enzyme Commission Number": "EC:2.6.1.79"},  
        "candidate reaction 3": {  
            "reaction id": "RHEA:12601",  
            "reaction equation": "2-oxoglutarate + L-2-amino adipate = 2-  
            oxoadipate + L-glutamate",  
            "reaction equation in ChEBI format": "CHEBI:16810 + CHEBI:58672  
            = CHEBI:57499 + CHEBI:29985",  
            "reaction equation in SMILES format": "[O-]C(=O)CCC(=O)C([O-])=O  
            . [NH3+] [C@@H] (CCCC([O-])=O)C([O-])=O>>[O-]C(=O)CCCC(=O)C([O-]  
            )=O.[NH3+] [C@@H] (CCC([O-])=O)C([O-])=O",  
            "reaction associated Enzyme Commission Number": "EC:2.6.1.39"},  
        "candidate reaction 4": {  
            "reaction id": "RHEA:20445",  
            "reaction equation": "L-arogenate + oxaloacetate = L-aspartate +  
            prephenate",  
            "reaction equation in ChEBI format": "CHEBI:58180 + CHEBI:16452  
            = CHEBI:29991 + CHEBI:29934",  
            "reaction equation in SMILES format": "[NH3+] [C@@H] (CC1(C=CC(O)C  
            =C1)C([O-])=O)C([O-])=O.[O-]C(=O)CC(=O)C([O-])=O>>[NH3+] [  
            C@@H] (CC([O-])=O)C([O-])=O.O[C@H]1C=C[C@] (CC(=O)C([O-])=O)  
            (C=C1)C([O-])=O",  
            "reaction associated Enzyme Commission Number": "EC:2.6.1.78"},  
        "candidate reaction 5": {  
            "reaction id": "-",  
            "reaction equation": "-"}  
    }  
}
```

260

261

Example Response: Ground Truth | Enzyme (No UniProt Annotation Available)

• Answer X

```
{  
  "results": [  
    {  
      "reaction_id": "RHEA:21824",  
      "selected": "no",  
      "confidence": 0.05,  
      "reason": "Sequence lacks the conserved PLP-binding Lys and class  
                 I aminotransferase signature motifs characteristic of EC  
                 2.6.1.1 enzymes."  
    },  
    {  
      "reaction_id": "RHEA:22880",  
      "selected": "no",  
      "confidence": 0.05,  
      "reason": "No homology to known prephenate aminotransferases (EC  
                 2.6.1.79) and absence of necessary active-site residues."  
    },  
    {  
      "reaction_id": "RHEA:12601",  
      "selected": "no",  
      "confidence": 0.05,  
      "reason": "Missing the hallmark sequence motifs of amino adipate  
                 aminotransferases (EC 2.6.1.39)."  
    },  
    {  
      "reaction_id": "RHEA:20445",  
      "selected": "no",  
      "confidence": 0.05,  
      "reason": "Does not display the conserved regions required for  
                 arogenate/oxaloacetate transamination (EC 2.6.1.78)."  
    },  
    {  
      "reaction_id": "-",  
      "selected": "yes",  
      "rank": 1,  
      "confidence": 0.95,  
      "reason": "No recognizable aminotransferase or other catalytic  
                 domains are present, indicating a non-catalytic function."  
    }  
  ]  
}
```

262

S2.4.6 Case 6: Accurate Interpretation Leveraging Both Reaction and Non-Reaction Predictions

263

264

Example Query: Ground Truth | Enzyme (No UniProt Annotation Available)

► Query Prompt 3 +

```
{  
    "protein information": {  
        "uniprot id": "-",  
        "protein amino acid sequence": "MSVESFERFLSRRVLSMQESQTMKITGLAKKMQA  
        ..."},  
    "reaction information": {  
        "candidate reaction 1": {  
            "reaction id": "RHEA:21824",  
            "reaction equation": "2-oxoglutarate + L-aspartate = L-glutamate  
            + oxaloacetate",  
            "reaction equation in ChEBI format": "CHEBI:16810 + CHEBI:29991  
            = CHEBI:29985 + CHEBI:16452",  
            "reaction equation in SMILES format": "[O-]C(=O)CCC(=O)C([O-])=O  
            . [NH3+] [C@@H] (CC([O-])=O)C([O-])=O>>[NH3+] [C@@H] (CCC([O-])=O  
            )C([O-])=O.[O-]C(=O)CC(=O)C([O-])=O",  
            "reaction associated Enzyme Commission Number": "EC:2.6.1.1"},  
        "candidate reaction 2": {  
            "reaction id": "RHEA:22880",  
            "reaction equation": "2-oxoglutarate + L-arogenate = L-glutamate  
            + prephenate",  
            "reaction equation in ChEBI format": "CHEBI:16810 + CHEBI:58180  
            = CHEBI:29985 + CHEBI:29934",  
            "reaction equation in SMILES format": "[O-]C(=O)CCC(=O)C([O-])=O  
            . [NH3+] [C@@H] (CC1(C=CC(O)C=C1)C([O-])=O)C([O-])=O>>[NH3+] [  
            C@@H] (CCC([O-])=O)C([O-])=O.O[C@H]1C=C[C@] (CC(=O)C([O-])=O) (  
            C=C1)C([O-])=O",  
            "reaction associated Enzyme Commission Number": "EC:2.6.1.79"},  
        "candidate reaction 3": {  
            "reaction id": "RHEA:12601",  
            "reaction equation": "2-oxoglutarate + L-2-aminoadipate = 2-  
            oxoadipate + L-glutamate",  
            "reaction equation in ChEBI format": "CHEBI:16810 + CHEBI:58672  
            = CHEBI:57499 + CHEBI:29985",  
            "reaction equation in SMILES format": "[O-]C(=O)CCC(=O)C([O-])=O  
            . [NH3+] [C@@H] (CCCC([O-])=O)C([O-])=O>>[O-]C(=O)CCCC(=O)C([O-]  
            )=O.[NH3+] [C@@H] (CCC([O-])=O)C([O-])=O",  
            "reaction associated Enzyme Commission Number": "EC:2.6.1.39"},  
        "candidate reaction 4": {  
            "reaction id": "RHEA:20445",  
            "reaction equation": "L-arogenate + oxaloacetate = L-aspartate +  
            prephenate",  
            "reaction equation in ChEBI format": "CHEBI:58180 + CHEBI:16452  
            = CHEBI:29991 + CHEBI:29934",  
            "reaction equation in SMILES format": "[NH3+] [C@@H] (CC1(C=CC(O)C  
            =C1)C([O-])=O)C([O-])=O.[O-]C(=O)CC(=O)C([O-])=O>>[NH3+] [  
            C@@H] (CC([O-])=O)C([O-])=O.O.O[C@H]1C=C[C@] (CC(=O)C([O-])=O) (C  
            =C1)C([O-])=O",  
            "reaction associated Enzyme Commission Number": "EC:2.6.1.78"},  
        "candidate reaction 5": {  
            "reaction id": "-",  
            "reaction equation": "-"}  
    }  
}
```

265

Example Response: Ground Truth | Enzyme (No UniProt Annotation Available)

• Answer ✓

```
{  
  "results": [  
    {  
      "reaction_id": "RHEA:21824",  
      "selected": "yes",  
      "rank": 1,  
      "confidence": 0.94,  
      "reason": "Sequence contains a conserved PLP-dependent  
                 aminotransferase class I domain, including the key Lys residue  
                 for Schiff base formation and the GlyGly motif, matching  
                 AspAT (EC 2.6.1.1) signatures."  
    },  
    {  
      "reaction_id": "RHEA:20445",  
      "selected": "yes",  
      "rank": 2,  
      "confidence": 0.88,  
      "reason": "Domain architecture and active-site residues also  
                 correspond to prephenate aminotransferase activity (EC  
                 2.6.1.78), consistent with dual specificity observed in  
                 related enzymes."  
    },  
    {  
      "reaction_id": "RHEA:22880",  
      "selected": "no",  
      "confidence": 0.04,  
      "reason": "Lacks the substrate-binding pocket features required  
                 for arogenate-specific aminotransferase (EC 2.6.1.79)."  
    },  
    {  
      "reaction_id": "RHEA:12601",  
      "selected": "no",  
      "confidence": 0.04,  
      "reason": "Absent the extended aminoadipate-binding loops and  
                 catalytic motifs of aminoadipate aminotransferase (EC  
                 2.6.1.39)."  
    },  
    {  
      "reaction_id": "-",  
      "selected": "no",  
      "confidence": 0.04,  
      "reason": "Clear catalytic domains are present, ruling out a non-  
                 catalytic function."  
    }  
  ]  
}
```

266

S3 Algorithm

267

Algorithm 2 : Active Learning via Attention Discrepancy

Require: Untrained dataset \mathcal{D}_u , training dataset \mathcal{D}_l , attention module $A(\cdot)$, encoder GRU , classifier FC , rounds R , samples per round K

- 1: **for** round $r = 1$ to R **do**
- 2: **for** each enzyme $e_i \in \mathcal{D}_u$ **do**
- 3: $z_i \leftarrow GRU(ESM(e_i))$
- 4: $p_i^{w/o} \leftarrow FC(z_i)$ ▷ Without attention
- 5: $p_i^w \leftarrow FC(A(z_i))$ ▷ With attention
- 6: $\delta e_i \leftarrow |p_i^w - p_i^{w/o}|$
- 7: Select top- K enzymes S with highest δe_i
- 8: $\mathcal{D}_l \leftarrow \mathcal{D}_l \cup S; \mathcal{D}_u \leftarrow \mathcal{D}_u \setminus S$
- 9: Retrain model M on updated \mathcal{D}_l
- 10: **return** Final model M

Algorithm 3 : Clustering-based Active Learning Strategy

Require: Untrained dataset \mathcal{D}_u , training dataset \mathcal{D}_l , model M , clusters C_1, \dots, C_m , rounds R , samples per round K

- 1: **for** round $r = 1$ to R **do**
- 2: Train model M on \mathcal{D}_l
- 3: **for** each cluster C_j **do**
- 4: Compute error: $err_j = \frac{1}{|C_j|} \sum e_i \in C_j \mathbb{I}(M(e_i) \neq \hat{y}_i)$
- 5: Compute sampling weight: $\alpha_j = \frac{err_j}{\sum_k err_k + \epsilon}$
- 6: Determine sample count: $k_j = \lfloor K \cdot \alpha_j \rfloor$
- 7: Randomly select k_j samples from C_j and add to S
- 8: $\mathcal{D}_l \leftarrow \mathcal{D}_l \cup S; \mathcal{D}_u \leftarrow \mathcal{D}_u \setminus S$
- 9: **return** Final model M

S4 SI Appendix Tables

268

S4.1 Performance Evaluation under 10-Fold Cross-Validation

269

We conducted 10-fold cross-validation on a randomly split subset of the SwissProt dataset (*ds_rcv*) to systematically benchmark the performance of various enzyme function prediction methods. As shown in the tables below, EC-based methods such as ECRECer and CatFam achieve moderate accuracy, with ECRECer performing best among them. In contrast, protein language model-based methods (e.g., T5, ESM, UniRep) exhibit significantly higher accuracy and recall, demonstrating the effectiveness of embedding-based function transfer. Notably, our proposed RXNRECer-S1 method outperforms all baselines across all evaluation metrics, achieving an mF1 score of 0.9847, highlighting its strong predictive capacity even under randomized data splits.

270

271

272

273

274

275

276

277

278

Supplementary Table S3. Performance of EC-based annotation methods on reaction prediction (10-fold cross-validation).

Method	mAccuracy	mPrecision	mRecall	mF1	noPrediction
ECRECer	0.8330 ± 0.0271	0.9129 ± 0.0200	0.7842 ± 0.0376	0.8193 ± 0.0265	0
MSA-via-EC	0.8210 ± 0.0325	0.8722 ± 0.0271	0.7739 ± 0.0309	0.7828 ± 0.0271	2392.90 ± 40.71
CatFam	0.7500 ± 0.0271	0.7923 ± 0.0277	0.6956 ± 0.0418	0.7406 ± 0.0344	0
CLEAN	0.3190 ± 0.0470	0.3489 ± 0.0566	0.3538 ± 0.0496	0.3511 ± 0.0525	0
DeepEC	0.2910 ± 0.0412	0.3544 ± 0.0454	0.3120 ± 0.0393	0.3318 ± 0.0418	30920.50 ± 88.40
PRIAM	0.0810 ± 0.0296	0.1735 ± 0.0310	0.3795 ± 0.0544	0.2371 ± 0.0354	14366.70 ± 106.82

Note: noPrediction = number of samples for which the method made no prediction.

Supplementary Table S4. Performance of protein language model- and reaction similarity-based methods on reaction prediction (10-fold cross-validation).

Method	mAccuracy	mPrecision	mRecall	mF1	noPrediction
T5-cosine	0.9546 ± 0.0010	0.9173 ± 0.0036	0.9836 ± 0.0005	0.9493 ± 0.0020	0
T5-euclidean	0.9541 ± 0.0010	0.9165 ± 0.0035	0.9836 ± 0.0005	0.9488 ± 0.0019	0
ESM-cosine	0.9369 ± 0.0016	0.8941 ± 0.0039	0.9803 ± 0.0009	0.9352 ± 0.0023	0
ESM-euclidean	0.9351 ± 0.0016	0.8894 ± 0.0036	0.9799 ± 0.0009	0.9325 ± 0.0022	0
UniRep-cosine	0.8908 ± 0.0016	0.8276 ± 0.0039	0.9622 ± 0.0010	0.8898 ± 0.0023	0
UniRep-euclidean	0.8897 ± 0.0015	0.8260 ± 0.0036	0.9618 ± 0.0010	0.8887 ± 0.0021	0
MSA-via-RXN	0.8380 ± 0.0374	0.8771 ± 0.1106	0.9589 ± 0.0142	0.8292 ± 0.0809	2392.90 ± 40.71

Note: noPrediction = number of samples for which the method made no prediction.

Supplementary Table S5. Performance of proposed RXNRECer method on reaction prediction (10-fold cross-validation).

Method	mAccuracy	mPrecision	mRecall	mF1
RXNRECer-S1	0.9873 ± 0.0006	0.9890 ± 0.0009	0.9804 ± 0.0013	0.9847 ± 0.0010

279

S4.2 Performance Evaluation under Recently Curated Proteins

To assess the real-world generalization capability of all evaluated methods, we further tested them on an independent dataset (Recently Curated Proteins, *ds_rcp*) comprising proteins that were recently curated in UniProtKB/Swiss-Prot (2018–2024). This dataset contains no overlap with any training data and reflects realistic deployment scenarios,

280

281

282

283

particularly involving novel or low-homology enzymes. The evaluation results are summarized below, grouped by EC-based annotation methods, embedding similarity-based methods, and PLM-based end-to-end models.

284

285

286

Supplementary Table S6. Performance of EC-based annotation methods on reaction prediction (*ds_rcp*).

Method	mAccuracy	mPrecision	mRecall	mF1	noPrediction
ECRECCer	0.6899	0.9680	0.6517	0.7148	0
MSA-via-EC	0.4334	0.9079	0.4092	0.5044	3974
CatFam	0.7705	0.8700	0.6704	0.6203	0
CLEAN	0.0789	0.8790	0.1455	0.0977	0
DeepEC	0.0349	0.9784	0.0524	0.0597	9840
PRIAM	0.0178	0.8534	0.1601	0.0754	4232

Note: noPrediction = number of samples for which the method made no prediction.

Supplementary Table S7. Performance of protein language model- and reaction similarity-based methods on reaction prediction (*ds_rcp*).

Method	mAccuracy	mPrecision	mRecall	mF1	noPrediction
T5-cosine	0.7469	0.8939	0.7448	0.7782	0
T5-euclidean	0.7461	0.8926	0.7436	0.7763	0
ESM-cosine	0.7274	0.8849	0.7238	0.7608	0
ESM-euclidean	0.7213	0.8826	0.7190	0.7577	0
UniRep-cosine	0.6898	0.8631	0.6700	0.7126	0
UniRep-euclidean	0.6885	0.8558	0.6683	0.7090	0
MSA-via-RXN	0.5013	0.9196	0.5457	0.6423	4280

Note: noPrediction = number of samples for which the method made no prediction.

Supplementary Table S8. Performance of proposed RXNRECer method on reaction prediction (*ds_rcp*).

Method	mAccuracy	mPrecision	mRecall	mF1
RXNRECer-S1	0.8347	0.8940	0.8184	0.8060
RXNRECer-S2	0.8571	0.8344	0.9306	0.8545

Supplementary Table S9. Protein sequence identity distributions for *ds_rcv* and *ds_rcp* datasets.

pident range	MSA Identity for <i>ds_rcv</i>		MSA Identity for <i>ds_rcp</i>	
	Count	Proportion	Count	Proportion
<30%	2767	5.44%	5167	38.23%
30–40%	1176	2.31%	1842	13.63%
40–50%	1996	3.92%	1386	10.26%
50–60%	2804	5.51%	1082	8.01%
60–70%	3810	7.49%	926	6.85%
70–80%	4666	9.17%	890	6.59%
80–90%	6429	12.64%	741	5.48%
>90%	27210	53.50%	1481	10.96%

Supplementary Table S10. Comparison of annotation coverage and prediction completeness across different methods for the *Fusarium venenatum* proteome (FS12832).

Method	Enzyme	Non-Enzyme	No-Prediction	Partial-EC	Full-EC	Reaction	EC-Orphan
RXNRECer	3297	9535	0	0	3297	3297	0
ECRECer	4536	8296	0	1262	3274	2585	1951
MSA	2638	3749	6445	403	2235	2133	549
DeepEC	984	0	11848	0	984	736	253
CatFam	1100	11732	0	0	1100	794	318
PRIAM	1100	0	11732	0	1100	794	318
CLEAN	12832	0	0	0	12832	9127	3705

Supplementary Table S11. Pairwise structural alignment results for RXNRECer vs CLEAN case

PDB1	PDB2	TM-to-PDB1	TM-to-PDB2	TM-Avg	TM-Max	RMSD
2473_4wo8	2473_4wod	0.9984	0.9984	0.9984	0.9984	0.3700
2473_4wod	2473_4woe	0.9743	0.9610	0.9676	0.9743	1.6100
2473_4wo8	2473_4woe	0.9714	0.9581	0.9647	0.9714	1.7100
5128_2ixo	XP_025583253.1	0.9401	0.6612	0.8006	0.9401	1.7700
5128_2ixn	XP_025583253.1	0.9371	0.6289	0.7830	0.9371	1.6700
5128_2ixn	XP_025588571.1	0.9341	0.6407	0.7874	0.9341	1.6900
5128_2ixo	5128_2ixn	0.8693	0.9121	0.8907	0.9121	1.7500
5128_2ixo	XP_025588571.1	0.9087	0.6554	0.7820	0.9087	1.9800
XP_025583253.1	XP_025588571.1	0.6824	0.6977	0.6901	0.6977	2.2800
2473_4woe	5128_2ixn	0.1535	0.3046	0.2290	0.3046	6.1900
2473_4wod	5128_2ixn	0.1521	0.2961	0.2241	0.2961	6.3400
2473_4wo8	5128_2ixn	0.1513	0.2939	0.2226	0.2939	6.1800
2473_4wo8	5128_2ixo	0.1502	0.2784	0.2143	0.2784	6.7900
2473_4woe	XP_025588571.1	0.1894	0.2784	0.2339	0.2784	7.1600
2473_4wod	XP_025588571.1	0.1921	0.2762	0.2341	0.2762	7.5900
2473_4wod	5128_2ixo	0.1458	0.2719	0.2088	0.2719	6.6700
2473_4wo8	XP_025588571.1	0.1861	0.2682	0.2271	0.2682	7.4900
2473_4woe	5128_2ixo	0.1421	0.2660	0.2041	0.2660	6.8000
2473_4woe	XP_025583253.1	0.1677	0.2407	0.2042	0.2407	7.6200
2473_4wod	XP_025583253.1	0.1682	0.2385	0.2033	0.2385	7.5700
2473_4wo8	XP_025583253.1	0.1581	0.2251	0.1916	0.2251	7.2300

Supplementary Table S12. Pairwise structural alignment results for protein glyco-transferase case (EC:2.4.99.19)

PDB1	PDB2	TM-to-PDB1	TM-to-PDB2	TM-Avg	TM-Max	RMSD
6gxc	5ogl	0.9988	0.9974	0.9981	0.9988	0.3200
3rce	6gxc	0.9921	0.9599	0.9760	0.9921	0.7400
3rce	5ogl	0.9915	0.9581	0.9748	0.9915	0.8500
AF-Q0P9C8	3aag	0.3284	0.9805	0.6544	0.9805	0.8400
3rce	AF-Q0P9C8	0.9763	0.9357	0.9560	0.9763	1.3600
6gxc	AF-Q0P9C8	0.9748	0.9653	0.9700	0.9748	1.4100
AF-Q0P9C8	5ogl	0.9661	0.9742	0.9701	0.9742	1.4300
3rce	3aag	0.3341	0.9438	0.6389	0.9438	1.4100
5ogl	3aag	0.3217	0.9358	0.6287	0.9358	1.5400
6gxc	3aag	0.3210	0.9331	0.6271	0.9331	1.5100
XP_025588202.2	XP_065464856.1	0.8976	0.9273	0.9124	0.9273	1.6300
XP_025580849.2	XP_065464856.1	0.8891	0.9270	0.9080	0.9270	2.3400
XP_025587400.1	XP_065464856.1	0.8870	0.9139	0.9005	0.9139	1.8900
XP_025580849.2	XP_025588202.2	0.8689	0.8775	0.8732	0.8775	1.7700
XP_025580849.2	XP_025587400.1	0.8672	0.8773	0.8722	0.8773	2.7000
XP_025587400.1	XP_025588202.2	0.8598	0.8582	0.8590	0.8598	2.5300
XP_025582350.1	XP_065464856.1	0.7318	0.7933	0.7625	0.7933	3.5700
XP_025582350.1	XP_025588202.2	0.7519	0.7908	0.7714	0.7908	3.7000
XP_025580849.2	XP_025582350.1	0.7768	0.7453	0.7610	0.7768	3.8000
XP_025582350.1	XP_025587400.1	0.7326	0.7717	0.7521	0.7717	3.8700
XP_025589614.1	XP_065464856.1	0.7515	0.7602	0.7558	0.7602	3.3200
XP_025580849.2	XP_025589614.1	0.7268	0.7478	0.7373	0.7478	3.1300
XP_025588202.2	XP_025589614.1	0.7330	0.7470	0.7400	0.7470	3.3200
XP_025587400.1	XP_025589614.1	0.7334	0.7461	0.7397	0.7461	3.3200
XP_025582350.1	XP_025589614.1	0.6787	0.7275	0.7031	0.7275	3.4600
3aag	XP_025589614.1	0.3432	0.2010	0.2721	0.3432	5.7900
3aag	XP_025580849.2	0.3196	0.1885	0.2540	0.3196	6.5300
3aag	XP_025588202.2	0.3090	0.1860	0.2475	0.3090	6.8400
5ogl	XP_025582350.1	0.2366	0.3015	0.2690	0.3015	6.6200
6gxc	XP_025582350.1	0.2369	0.3010	0.2689	0.3010	6.7200
3aag	XP_025587400.1	0.2987	0.1793	0.2390	0.2987	6.5100
3aag	XP_065464856.1	0.2967	0.1757	0.2362	0.2967	6.1300
3aag	XP_025582350.1	0.2892	0.1616	0.2254	0.2892	6.4000
3rce	XP_025582350.1	0.2327	0.2859	0.2593	0.2859	7.5900
AF-Q0P9C8	XP_025587400.1	0.2156	0.2857	0.2507	0.2857	7.1500
AF-Q0P9C8	XP_025588202.2	0.2116	0.2836	0.2476	0.2836	6.9200
AF-Q0P9C8	XP_025582350.1	0.2204	0.2803	0.2504	0.2803	7.3600
3rce	XP_065464856.1	0.2116	0.2783	0.2450	0.2783	7.1100
AF-Q0P9C8	XP_065464856.1	0.2011	0.2753	0.2382	0.2753	7.0100
AF-Q0P9C8	XP_025589614.1	0.2032	0.2728	0.2380	0.2728	7.0700
AF-Q0P9C8	XP_025580849.2	0.2055	0.2716	0.2386	0.2716	7.0700
6gxc	XP_025587400.1	0.2031	0.2716	0.2374	0.2716	6.7000
5ogl	XP_025587400.1	0.2016	0.2689	0.2353	0.2689	6.6800
6gxc	XP_025580849.2	0.2033	0.2656	0.2344	0.2656	7.0500
3rce	XP_025587400.1	0.2051	0.2647	0.2349	0.2647	6.8100
3rce	XP_025580849.2	0.2064	0.2637	0.2351	0.2637	6.6900
5ogl	XP_025588202.2	0.2016	0.2632	0.2324	0.2632	7.6700
5ogl	XP_025580849.2	0.2006	0.2621	0.2313	0.2621	7.3200
5ogl	XP_025589614.1	0.1970	0.2616	0.2293	0.2616	7.1800
6gxc	XP_025588202.2	0.1968	0.2594	0.2281	0.2594	6.8200
3rce	XP_025588202.2	0.2010	0.2578	0.2294	0.2578	6.8200
6gxc	XP_065464856.1	0.1882	0.2558	0.2220	0.2558	6.7800
6gxc	XP_025589614.1	0.1898	0.2525	0.2211	0.2525	7.5900
5ogl	XP_065464856.1	0.1882	0.2525	0.2203	0.2525	7.5400
3rce	XP_025589614.1	0.1844	0.2426	0.2135	0.2426	6.7600

Supplementary Table S13. Runtime of individual methods on ds_{rcp} grouped by speed categories. Methods exceeding 5000s were excluded from large-scale ensemble evaluation due to prohibitive runtime.

Category	Method	Time (s)
Ultra-fast (<60s)	MSAvia-RXN	13.77
	MSAvia-EC	31.50
Relatively fast (60–900s)	RXNRECer-S1	815.08
Fast (900–2000s)	T5_cos	1026.41
	PRIAM	1095.75
	ESM_cos	1327.04
	CatFam	1401.73
Moderate (2000–6000s)	DeepEC	4497.51
	ECRECer	4596.76
	UniRep.cos	4888.82
Slow (>6000s)	CLEAN	7744.88
Very slow (>10 ⁵ s)	ECPred	150386.52

Supplementary Table S14. Performance and runtime of individual models and ensembles on ds_{rcp} . Runtime is reported in seconds.

Ensemble ID / Methods	mAcc.	mPrec.	mRec.	mF1	Type	#Pred.	Runtime (s)
DeepEC	0.0349	0.9784	0.0524	0.0597	—	1	4497.51
ECRECer	0.6899	0.9680	0.6517	0.7148	—	1	4596.76
MSA-via-RXN	0.5013	0.9196	0.5457	0.6423	—	1	13.77
MSA-via-EC	0.4334	0.9079	0.4092	0.5044	—	1	31.50
RXNRECer-S1	0.8347	0.8940	0.8184	0.8060	—	1	815.08
T5	0.7469	0.8939	0.7448	0.7782	—	1	1026.41
ESM	0.7274	0.8849	0.7238	0.7608	—	1	1327.04
CLEAN	0.0789	0.8790	0.1455	0.0977	—	1	7744.88
CatFam	0.7705	0.8700	0.6704	0.6203	—	1	1401.73
UniRep	0.6898	0.8631	0.6700	0.7126	—	1	4888.82
PRIAM	0.0178	0.8534	0.1601	0.0754	—	1	1095.75
E1	0.5985	0.7531	0.9462	0.8132	M	2	5411.84
E1	0.5985	0.7531	0.9462	0.8132	R	2	5411.84
E2	0.5989	0.7536	0.9475	0.8141	M	3	5425.61
E3	0.3552	0.6868	0.9510	0.7538	R	3	6507.59
E4	0.5989	0.7530	0.9477	0.8137	R	4	5457.11
E5	0.3330	0.6739	0.9525	0.7451	R	4	11396.41
E6	0.5984	0.7524	0.9477	0.8133	R	5	6858.84
E7	0.0868	0.6580	0.9532	0.7277	R	5	19141.29
E8	0.5982	0.7514	0.9477	0.8127	R	6	7885.25
E9	0.0868	0.6581	0.9534	0.7278	R	6	20167.70
E10	0.5989	0.7507	0.9344	0.8055	M	7	16619.69
E11	0.0868	0.6581	0.9535	0.7278	R	7	24665.21

Note: M = Majority Vote; R = Recall Boosted; “—” denotes individual (non-ensemble) methods.

E1 = {ECRECer, RXNRECer-S1};

E2 = {ECRECer, MSA-via-RXN, RXNRECer-S1};

E3 = {ECRECer, PRIAM, RXNRECer-S1};

E4 = {MSA-via-EC, ECRECer, MSA-via-RXN, RXNRECer-S1};

E5 = {ECRECer, PRIAM, UniRep, RXNRECer-S1};

E6 = {MSA-via-EC, ECRECer, CatFam, MSA-via-RXN, RXNRECer-S1};

E7 = {CLEAN, ECRECer, PRIAM, UniRep, RXNRECer-S1};

E8 = {MSA-via-EC, ECRECer, CatFam, MSA-via-RXN, T5, RXNRECer-S1};

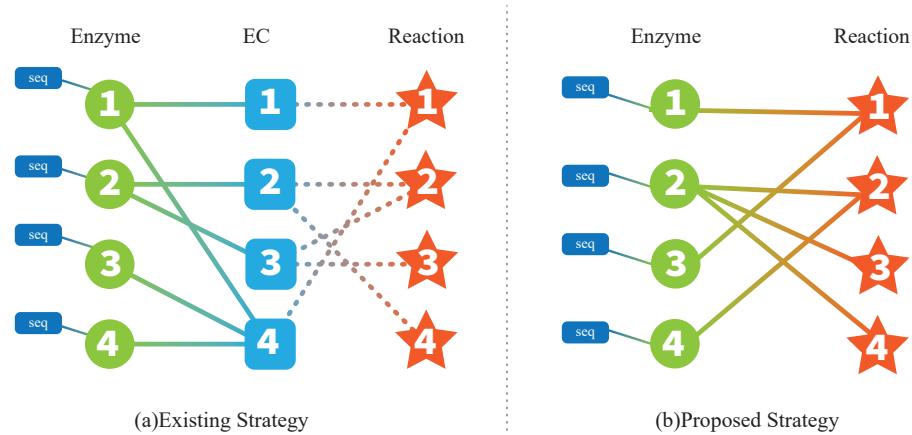
E9 = {CLEAN, ECRECer, PRIAM, UniRep, T5, RXNRECer-S1};

E10 = {CLEAN, ECRECer, PRIAM, MSA-via-RXN, ESM, T5, RXNRECer-S1};

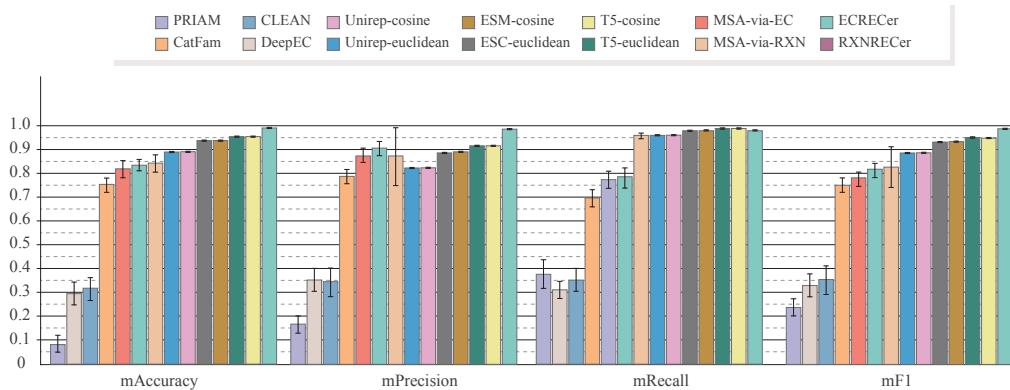
E11 = {DeepEC, CLEAN, ECRECer, PRIAM, UniRep, T5, RXNRECer-S1}.

S5 SI Appendix Figures

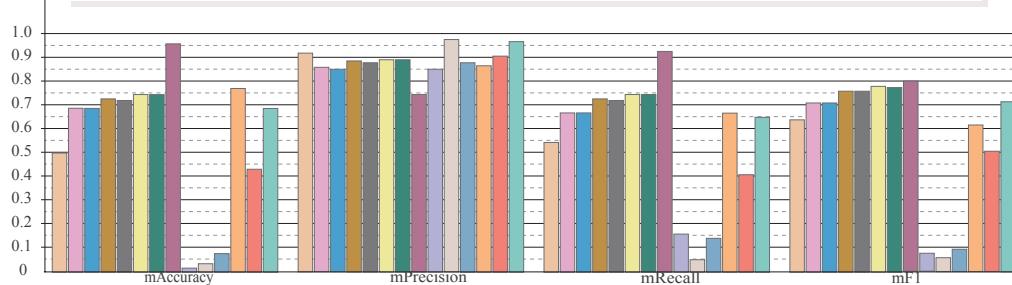
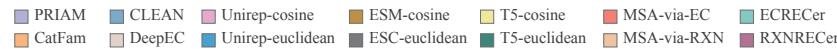
287



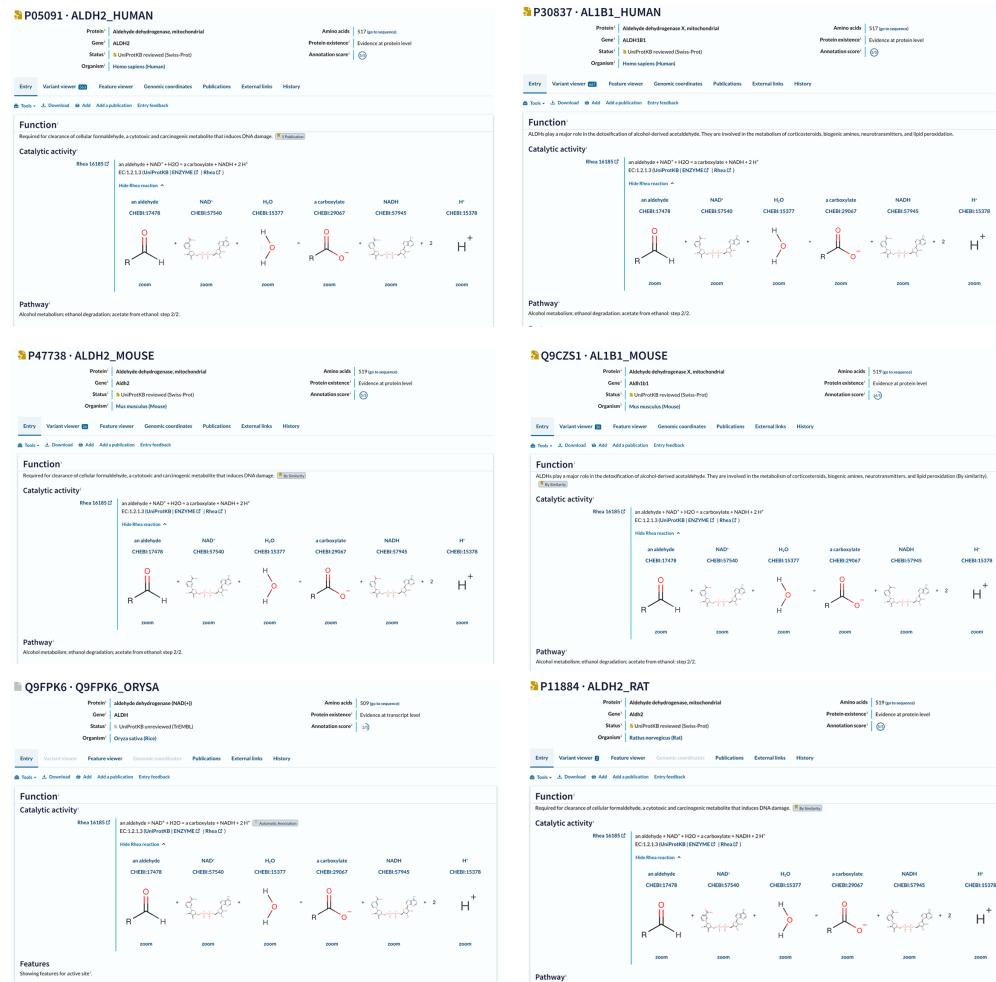
Supplementary Figure S4. A schematic illustration of: a) existing enzyme-reaction relation mutual prediction strategy; b) our proposed enzyme-reaction relation mutual prediction strategy.



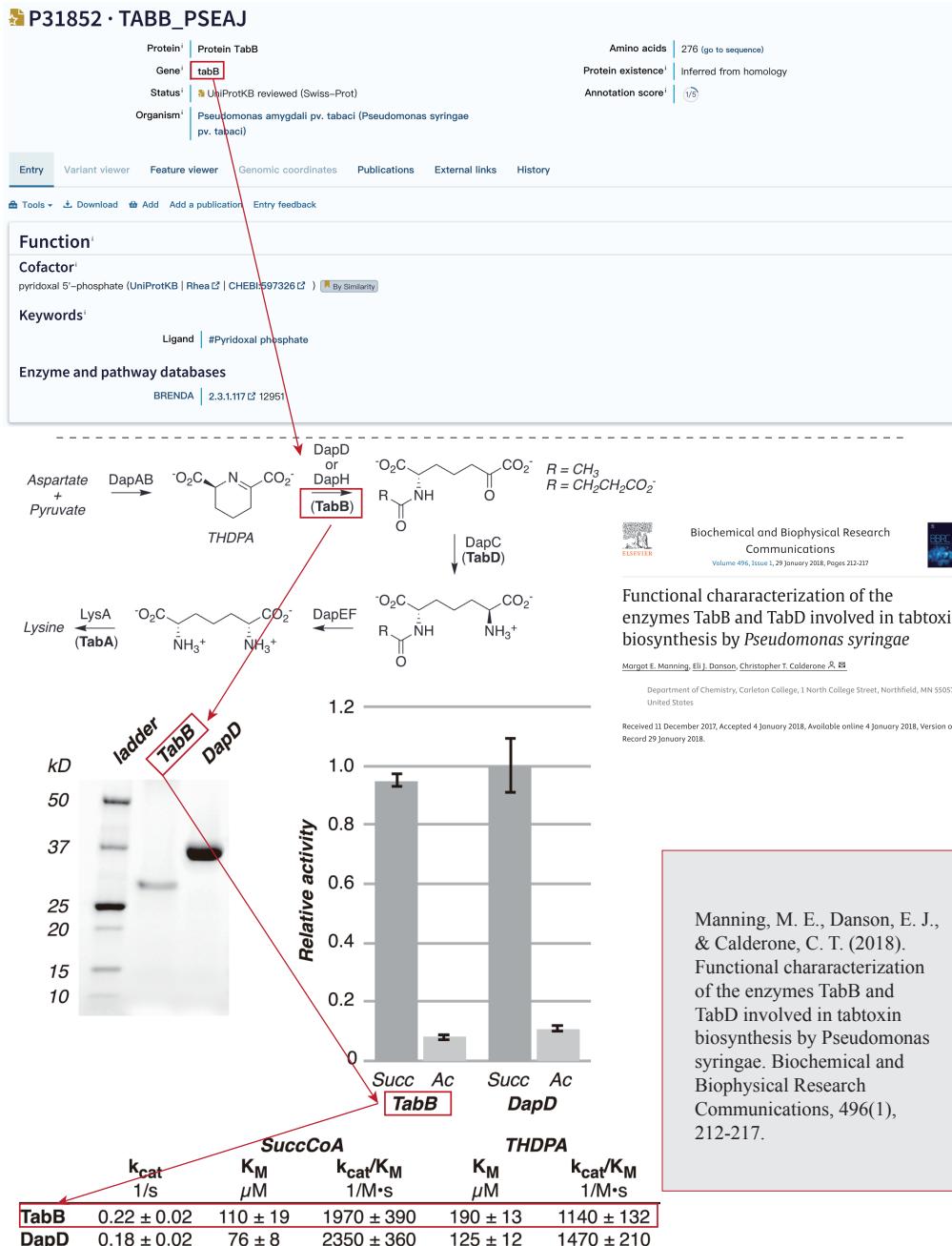
Supplementary Figure S5. Performance comparison for enzyme reaction prediction on ds_rcv datasets.



Supplementary Figure S6. Performance comparison for enzyme reaction prediction on ds_rcp datasets.



Supplementary Figure S7. UniProt-listed functional annotations for representative proteins analyzed in Case II.



Supplementary Figure S8. Experimental evidence supporting the catalytic function of TabB (UniProtKB ID: P31852) from *Pseudomonas syringae*. Manning *et al.* [13] demonstrated that TabB catalyzes the transfer of a succinyl group from succinyl-CoA to tetrahydrodipicolinate (THDPA), forming N-succinyl-2-amino-6-oxoheptanedioate and CoA (RHEA:17325). This biochemical validation provides direct support for RXNREcer's prediction in Case III.

P73408 · Y1840_SYNY3

Proteinⁱ Uncharacterized protein **slr1840**
 Statusⁱ UniProtKB reviewed (Swiss-Prot)
 Organismⁱ *Synechocystis* sp. (strain ATCC 27184 / PCC 6803 / Kazusa)

Amino acids 371 ([go to sequence](#))
 Protein existenceⁱ Inferred from homology
 Annotation scoreⁱ 1/8

Entry Variant viewer Feature viewer Genomic coordinates Publications External links History

Tools Download Add Add a publication Entry feedback

Functionⁱ

Gene Ontologyⁱ

GO annotations GO-CAM models New

Gene Ontology (GO) annotations organized by slimming set.

Slimming set: Bacteria and Archaea (Prokaryotes) ▾

Table 1
 Oligonucleotides used for PCR amplification (restriction sites underlined) of GK genes from *Arabidopsis thaliana* (*AtGLYK*, *Atg80380*), *Oryza sativa* (*OsGLYK*, *Os01g0682500*), *Nostoc* sp. PCC 7120 (*NosGLYK*, CYORF *alr2873*), *Saccharomyces cerevisiae* (*ScGLYK*, *YGR205W*), *Synechocystis* sp. PCC 6803 (*SynGK*, **slr1840**), and *E. coli* K12 (*EcoGK1*, *BAE77171*; *EcoGK2*, *BAE76292*)

Gene	Direction	Sequence
<i>AtGLYK</i>	Forward Reverse	5'-CTCGAGACAGTGGATGTCTTCGGTGTCAG-3' 5'- <u>CTCGA</u> GTTAGTTGGAGTATCGGGTCCCT-3'
<i>OsGLYK</i>	Forward Reverse	5'-CTCGAGATCTCCTCCGTCAGGAC-3' 5'- <u>GGTAC</u> CTCATCTACCCCCACATAGGATTCTC-3'
<i>ScGLYK</i>	Forward Reverse	5'-CTCGAGCCCTCCCTCTCTTATTATCTCCAAG-3' 5'- <u>GGTAC</u> CTTACCAATTCCGGTTGATAAACTTGAC-3'
<i>NosGLYK</i>	Forward Reverse	5'-CTCGAGATGCAAGTTGGGCAGAAAAGAACG-3' 5'- <u>GGTAC</u> CTTACCAATTCCGGTTGATAAACTTGAC-3'
<i>SynGK</i>	Forward Reverse	5'-GCTCGAGATGCGTCAAATTGGAT-3' 5'- <u>CGAAT</u> TCTCAGTTTGACTTTGGA-3'
<i>EcoGK1</i>	Forward Reverse	5'-CTCGAGATGAAAATCGTAATCG-3' 5'- <u>GAATT</u> TCTACCCCGCGTTGCGC-3'
<i>EcoGK2</i>	Forward Reverse	5'-CTCGAGATGAAGATTGTCAATTGC-3' 5'- <u>GAATT</u> TCTTAGTTTAATTCCC-3'

Table 2
 Kinetic parameters and product formation

Enzyme	K_m (D-glycerate) [mM]	K_m (ATP) [mM]	V_{max} [$\mu\text{mol min}^{-1} \text{mg}^{-1}$]	Product
<i>Class I</i>				
<i>EcoGK1</i>	0.086 ± 0.001	0.121 ± 0.016	465 ± 10	2PGA
<i>EcoGK2</i>	0.056 ± 0.006	0.200 ± 0.014	70.5 ± 9.5	2PGA
<i>SynGK</i>	0.043 ± 0.011	0.212 ± 0.033	14.5 ± 2.9	2PGA
<i>Class III</i>				
<i>NosGLYK</i>	0.104 ± 0.011	0.206 ± 0.005	42.6 ± 10	3PGA
<i>ScGLYK</i>	0.208 ± 0.010	0.890 ± 0.043	268 ± 16	3PGA
<i>OsGLYK</i>	0.450 ± 0.054	0.812 ± 0.054	295 ± 65	3PGA
<i>AtGLYK</i>	0.266 ± 0.003	0.834 ± 0.138	364 ± 54	3PGA
<i>AtGLYK^{nat}</i>	0.264 ± 0.036	0.783 ± 0.059	Not determined	3PGA

Bartsch, O., Hagemann, M., & Bauwe, H. (2008). Only plant-type (GLYK) glycerate kinases produce D-glycerate 3-phosphate. *Febs Letters*, 582(20), 3025-3028.

Supplementary Figure S9. Experimental evidence supporting the catalytic function of Slr1840 (UniProtKB ID: P73408) from *Synechocystis* sp. PCC 6803. Bartsch *et al.* [14] demonstrated that Slr1840 catalyzes the phosphorylation of (R)-glycerate using ATP to produce ADP and 2-phospho-(R)-glycerate (RHEA:23516). This biochemical validation provides direct support for RXNRECEr's prediction in Case III.

References

1. Chenggang Yu, Nela Zavaljevski, Valmik Desai, and Jaques Reifman. Genome-wide enzyme annotation with precision control: Catalytic families (catfam) databases. *Proteins: Structure, Function, and Bioinformatics*, 74(2):449–460, 2009.
 2. Clotilde Claudel-Renard, Claude Chevalet, Thomas Faraut, and Daniel Kahn. Enzyme-specific profiles for genome annotation: Priam. *Nucleic acids research*, 31(22):6633–6639, 2003.
 3. Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proceedings of the National Academy of Sciences*, 116(28):13996–14001, 2019.
 4. Zhenkun Shi, Qianqian Yuan, Ruoyu Wang, Hoaran Li, Xiaoping Liao, and Hongwu Ma. Ecrecer: Enzyme commission number recommendation and benchmarking based on multiagent dual-core learning. *arXiv preprint arXiv:2202.03632*, 2022.
 5. Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023.
 6. Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60, 2015.
 7. Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.
 8. Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.
 9. Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 2024.
 10. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025.
 11. Parit Bansal, Anne Morgat, Kristian B Axelsen, Venkatesh Muthukrishnan, Elisabeth Coudert, Lucila Aimo, Nevila Hyka-Nouspikel, Elisabeth Gasteiger, Arnaud Kerhennou, Teresa Batista Neto, et al. Rhea, the reaction knowledgebase in 2022. *Nucleic acids research*, 50(D1):D693–D700, 2022.
 12. M Su, Y Ling, J Yu, J Wu, and J Xiao. Small proteins: untapped area of potential biological importance. *front. genet.* 4: 286, 2013.
 13. Margot E Manning, Eli J Danson, and Christopher T Calderone. Functional characterization of the enzymes tabb and tabd involved in tabtoxin biosynthesis by pseudomonas syringae. *Biochemical and Biophysical Research Communications*, 496(1):212–217, 2018.
 14. Oliver Bartsch, Martin Hagemann, and Hermann Bauwe. Only plant-type (glyk) glycerate kinases produce d-glycerate 3-phosphate. *Febs Letters*, 582(20):3025–3028, 2008.
-