Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimodal AI unveils a billion-scale microbial enzyme atlas

Xiaochen Bo

boxiaoc@163.com

Institute of Health Service and Transfusion Medicine https://orcid.org/0000-0003-3490-5812

Ming Ni

Academy of Military Medical Sciences

Jianting Gong

Institute of Health Service and Transfusion Medicine

Pengjia Ma

4School of Information Science and Technology, Northeast Normal University

Wanyi Chen

Institute of Health Service and Transfusion Medicine

Yibo Zhang

Institute of Health Service and Transfusion Medicine

Lingchun Qiao

Institute of Health Service and Transfusion Medicine

Xiaomin Zhang

Institute of Health Service and Transfusion Medicine

Haoyue Gu

Institute of Health Service and Transfusion Medicine

Junling Yang

Institute of Health Service and Transfusion Medicine

Songyang Diao

School of Information Science and Technology, Northeast Normal University

Wanlin Liu

State Key Laboratory of Medical Proteomics, National Center for Protein Sciences (Beijing)

Zilin Ren

Changchun Veterinary Research Institute, Chinese Academy of Agricultural Sciences
https://orcid.org/0000-0002-3621-1024

Yunming Fang

College of Chemical Engineering and Technology, Beijing University of Chemical Technology, Beijing 100029, PR China https://orcid.org/0000-0001-9586-5581

Tianwei Tan

State Key Laboratory of Green Biomanufacturing, Beijing University of Chemical Technology

Biological Sciences - Article

Keywords:

Posted Date: June 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9629622/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Multimodal AI unveils a billion-scale microbial enzyme atlas

Ming Ni[1,2†], Jianting Gong[1,4,†], Pengjia Ma[1,4,†], Wanyi Chen[1,3,5], Yibo Zhang[1,3], Lingchun Qiao[1,4], Xiaomin Zhang[1], Haoyue Gu[1,5], Junling Yang[1,5], Songyang Diao[4], Wanlin Liu[7], Zilin Ren[4], Yunming Fang[3], Xiaochen Bo[1,7,*], Tianwei Tan[3]

[1]Institute of Health Service and Transfusion Medicine, Beijing 100850, People's Republic of China

[2]Beijing Institute of Radiation Medicine, Beijing 100850, People's Republic of China

[3]State Key Laboratory of Green Biomanufacturing, Beijing University of Chemical Technology, Beijing 100029, People's Republic of China

[4]School of Information Science and Technology, Northeast Normal University, Changchun 130117, People's Republic of China

[5]Key Laboratory of Geriatric Nutrition and Health, School of Light Industry Science and Engineering, Beijing Technology and Business University, 100048, Beijing, People's Republic of China

[6]Institute of Bioengineering, College of Chemical and Biological Engineering, Zhejiang University, Hangzhou, Zhejiang 310058, People's Republic of China

[7]State Key Laboratory of Medical Proteomics, National Center for Protein Sciences (Beijing), Beijing 102206, People's Republic of China.

[†]Ming Ni, Jianting Gong and Pengjia Ma contributed equally to this work.

[*]Corresponding author: Xiaochen Bo, E-mail: boxc@bmi.ac.cn.

Author contributions

M.N., Y.F. and X.B. conceived the study, wrote and revised the paper. M.N., J.G., and P.M. performed the computational investigation and interpreted the data, and prepared the figures and tables. J.G., and P.M. built and trained the models. W.C. and Y.Z. performed the experiments and interpreted the data. L.Q., X.Z. and H.G. performed data collection and analysis. J.Y., S.D., W.L. and Z.R. assisted with the data collection and analysis. Y.F., X.B., and T.T. obtained funding, coordinated the study, and supervised the research.

Abstract

Microbial enzymes constitute a vast reservoir for biocatalyst discovery, yet the functional catalog of enzymes within the microbial proteome remains largely uncharted. Here, we develop RAMER, a pretrained multimodal artificial intelligence (AI) model for enzymes, integrating protein sequences, structures and catalytic reactions. RAMER substantially improves recall for enzyme function annotation at level-4 Enzyme Commission (EC) classes. Applying RAMER to 3.08 billion proteins from marine, freshwater, terrestrial, and extreme microbiomes, we constructed an atlas of 1.04 billion enzymes assigned to over 5600 EC classes, representing nearly 30 times the number of automatically annotated enzymes in UniProtKB/TrEMBL. For microbiomes from distinct environments, enzymes account for 40% to 51.1% of the proteomes and exhibit broadly similar catalytic function profiles. Beyond EC classes, RAMER enables fine-grained organization of specific enzymes in a manner that reflects functional structural domains and conformations, further facilitating enzyme mining and exploration of natural enzyme diversity. With RAMER, we mined and validated fluorinases from *Actinomycetota* and PETases from the Mariana Trench microbiome, some of which possessed new structural conformations. Together, the atlas provides a comprehensive resource and a perspective of microbial enzymes, and RAMER offers an AI framework for function-driven biocatalyst discovery.

Introduction

Enzymes serve as the functional molecules that underpin life's essential biological processes. The protein universe is replete with enzymes[1–3], and those of microbial origin are particularly diverse. The richness of the so-called "microbial dark matter" has been increasingly revealed by metagenomic and genomic sequencing[4,5]. Biological functions are highly efficiently realized through enzyme-catalyzed reactions that convert substrates into products, linking molecular-level catalysis to organismal physiology. Beyond their fundamental roles in life, the remarkable catalytic specificity and efficiency of enzymes have enabled widespread applications in biomanufacturing, biomedicine, life sciences, food and agriculture, as well as environmental and energy sustainability.

Continuous exploration and understanding of the natural enzymatic repertoire are therefore of great scientific and practical significance[6]. Advances in artificial intelligence (AI) for the biological and chemical sciences have provided paradigm-shifting tools for this endeavor. Tools such as AlphaFold2/3[7,8] and ESMFold[9] provide unprecedented convenience for determining enzyme structures to investigate their functions. Biological foundation models, exemplified by protein large language models (PLMs) such as Evolutionary Scale Modeling (ESM)[9–11], ProtT5[12] and Ankh[13], are widely employed for enzyme function-related downstream tasks beyond sequence homology. These tasks include, for example, functional classification (CLEAN[14] and EnzymeCAGE[15]) and catalytic activity prediction (CatPred[16] and CataPro[17]).

Microbiomes are a rich reservoir for functional enzyme mining, and AI-powered tools and techniques offer effective access to this resource. There are many recent examples, such as the discovery of a cellulose-cleaving metalloenzyme from sugarcane-bagasse soil microbiomes[18], the mining of highly active urethanases for polyurethane recycling from public metagenomic data[19], and the identification of ancestral CRISPR-Cas13 ribonucleases[20] in AlphaFold Protein Structure Database (AFDB)[21]. Therefore, it is imperative to expand the functional enzyme catalog and establish it as an accessible public resource. UniProtKB/TrEMBL[22] now contains nearly 36 million enzymes with computationally predicted annotations under the Enzyme Commission (EC) scheme (as of April 9, 2026), which is still limited relative to the vast protein universe and the unexplored microbial dark matter. Moreover, the

accuracy of these annotations, particularly for specific catalytic reactions (i.e., defining level-4 EC classes), needs further improvement.

Here, we present RAMER (Reaction-Aware Multi-modal model for Enzyme Representation), a pretrained model that integrates enzyme sequences, structures, and catalytic reactions. RAMER exhibits improved performance in predicting level-4 EC classes based on enzyme amino acid (aa) sequences. By applying RAMER to 3.08 billion microbial proteins from microbiomes across distinct habitats, we annotated 1.04 billion enzymes distributed across more than 5,600 EC classes, forming an atlas and comprehensive view for enzymes. Moreover, RAMER enables a more functionally relevant organization of enzymes within the same EC class than that achieved by PLMs or structure-based models. We experimentally validated novel fluorinases and PETases identified by RAMER, demonstrating its utility in enzyme mining.

Results

RAMER is a reaction-aware pre-trained multimodal model for enzyme representation

We hypothesized that a pretrained model with a zero-shot strategy is suitable for enzyme-related annotation of large-scale protein datasets. We therefore constructed and trained RAMER, a pretrained AI model that integrates enzymatic reactions with their corresponding enzyme sequences and protein structures. Reactions and enzyme proteins are biological entities. We did not include any other manually curated or artificial modalities in the training, such as EC numbers, Gene Ontology terms, or natural language functional descriptions of enzymes.

RAMER has a tri-modal architecture, as illustrated in **Fig. 1a** (see **SFig. 1a** for model details). For each modality, we adopted a representation based on other biological and chemical fundamental models. First, to integrate the modality of enzymatic reactions, the simplified molecular input line entry system (SMILES) specification of substrates and products of reactions were first embedded using MolT5[23], a chemical language model. Reaction directions were encoded using positional embeddings, which were added to the projection layer of the MolT5 embeddings. Then, the combined representations were fed into a Transformer encoder for the training of reaction-aware network. For embedding of protein sequences, we evaluated several PLMs. Although ESM-3B exhibited an advantage without fine-tuning, after fine-tuning

on enzyme sequences from Swiss-Prot, ProtT5[12] achieved slightly better results than ESM-3B (**SFig. 2**). Therefore, we selected ProtT5 as the embedding input for the enzyme sequence modality. Finally, given that protein structures are more related to enzyme functions than sequences and the limitations of PLMs in capturing structural information[24], we utilized GearNet[25], a geometric structure pretrained model for protein structures, to embed enzyme structures obtained from AFDB. Projection layers were appended to each encoder to unify dimensions.

For training and fine-tuning of RAMER, we constructed a matched multimodal dataset comprising 200,024 enzymes and 10,596 reactions, generating 260 thousand sequence-structure-reaction trios (**SFig. 1b**). We adopted a contrastive learning architecture[26] to multimodal fusion based on these trios. Specifically, the alignment losses for reaction-sequence and structure-sequence pairs were trained separately on these trios. The training optimized the last six layers of ProtT5, the reaction-aware network and three modalities projector layers, comprising 37 million parameters. Price-149 dataset[27], which was excluded from the training datasets, was used to select the final model of RAMER **(SFig. 1c**).

RAMER enables high-recall enzyme function prediction

We obtained over 550,000 proteins from Swiss-Prot database (till Oct 2025), half of which (48.6%) are assigned with Enzyme Commission (EC) numbers. RAMER was used to generate the representations of these protein sequences, which were visualized via dimensionality reduction techniques as uniform manifold approximation projection (UMAP)[28]. The proteins with and without EC numbers as well as the different major classes of enzymes (level-1 EC numbers) exhibit a clear separation (**Fig. 1b**). In contrast, with other PLMs, both the distinction between enzymes and non-enzymes and the separation of major classes are visually mixed (**SFig. 3**). We further calculated the clustering indices at all four levels of EC numbers, and RAMER exhibited higher performance in clustering than other PLMs and the representation module of CLEAN (**Fig. 1c-d and SFig. 4**). Similarly, we evaluated performance of RAMER for clustering 15,046 unique reactions from the RHEA database[29], which were comparable with MolT5 embeddings of reactions (UMAPs in **SFig. 5a-b**, clustering indices in **SFig. 5c-f**). It is understandable since enzyme function annotation is based on catalytic reactions.

Next, we applied RAMER for the task of enzyme function prediction—specifically,

EC numbers—from protein sequences. We found that a zero-shot approach was more favorable than a contrastive learning approach (**STab. 1**). Namely, to predict the EC number of a given enzyme, we computed the cosine similarity between the query enzyme's embedding and the embeddings of enzymes in a background dataset (i.e., knowledgebase), identified the entry with the highest similarity, and assigned its EC number to the query enzyme. Considering the versatility of enzymes, we employed two modes for the EC number prediction: the 'top-1' match mode and the 'max-sep' mode involving a set of similar enzymes by using the maximum separation technique.

The RAMER-based EC prediction was then compared with tools including CLEAN, GraphEC[30], DeepECtransformer[31], ProteInfer[32], and ProTrek[33]. We employed the same background dataset for zero-shot prediction as in CLEAN, and the New-392 dataset consisting of 392 enzymes spanning 177 EC numbers[14] were used as a validation set. We excluded the New-392 enzymes from the training set of RAMER. For the prediction of level-1 to level-3 EC numbers, RAMER and other tools, except GraphEC, were generally comparable. The top-1 mode of RAMER slightly surpassed other tools (**Fig. 1e** and **STab. 2**).

Intriguingly, RAMER, especially the max-sep mode, exhibited notable improvements in predicting challenging level-4 EC numbers (**Fig. 1f**). The RAMER max-sep mode achieved substantially higher recall (0.662) than these of CLEAN (0.479) and other tools, yielding a 0.125 increase in F1 score over CLEAN. Additionally, we curated another validation set, named Ram-255, comprising 255 enzymes with 175 EC numbers released after March 30, 2023, as a supplement to New-392 (**SFile. 1**). RAMER consistently had an advantage for level-4 EC number prediction in the Ram-255 set (F1 0.102 larger than CLEAN, **Fig. 1g**).

RAMER-based prediction is also less affected by insufficient training data and sequence dissimilarity than the other models. Enzymes in validation sets New-392 and Ram-255 were grouped by their abundance in level-4 EC numbers, and RAMER max-sep model exhibit consistently higher performance for all the groups than CLEAN and GraphEC (**Fig. 1h** and **SFig. 6a-b**). Next, we gradually excluded validation enzymes that shared varying degrees of sequence homology with those in the training set. While F1 scores decreased with the thresholds for all tools, RAMER maintained its surpass over CLEAN and GraphEC (**Fig. 2i** and **SFig. 6c-d**). Moreover, we obtained 118

enzymes with new level-4 EC numbers (released after March 30, 2023, **SFile. 2**) that did not exist in the training sets or the background set, and we found that RAMER surpassed CLEAN and GraphEC in the prediction of level-1 to level-3 EC numbers (**Fig. 1g** and **SFig. 6e-f**).

Additionally, we conducted an ablation study to assess the contribution of the three modalities to level-4 EC number prediction based on New-392. As shown in **Fig. 1k**, compared with the baseline model using only enzyme sequence information, the addition of protein structural information and catalytic reaction information individually improved precision and recall. Notably, the reaction modality led to a greater enhancement than the structure modality. When simultaneously integrating both reactions and structures (i.e., the full RAMER model), the recall showed further incremental improvement, whereas we observed a slight decrease in precision. This result shows that both reaction and structural information improve functional annotation performance, with the reaction modality contributing predominantly.

A functional atlas of 1.04 billion enzymes of microbiomes from diverse habitats

The high recall of RAMER in predicting EC numbers make it well-suited for large-scale functional annotation of enzymes in the enormous protein universe. Therefore, to obtain a comprehensive functional atlas of enzymes, as illustrated in **Fig. 2a**, we first obtained 3.08 billion microbial protein sequences from MGnify[5], Global Catalogue of Metagenomics (gcMeta)[34], global ocean microbiome genome catalogue (GOMC)[35], and recently reported microbiome dataset from Mariana Trench[36] (**STab. 3**). The RAMER embeddings of the 3.08 billion proteins were calculated. A machine-learning classifier was first constructed to extract enzymes from the total protein set (see Methods); proteins in Swiss-Prot annotated as enzymes and non-enzymes were used to train the classifier, and they were found to be distinguishable using RAMER embeddings, as well as ESM and ProtT5 embeddings (**SFig. 7**). A total of 1.04 billion enzymes were identified, and, subsequently, EC numbers of these enzymes were predicted by using the RAMER top-1 model, forming a billion-scale enzyme atlas.

We observed that the proportion of enzymes in the overall proteins varied from different databases (**Fig. 2b**). This variation is primarily attributable to differences in protein length distributions, which reflect the assembly completeness of the metagenome assemblies. Specifically, MGnify is relatively enriched with short proteins,

resulting in a low enzyme proportion of 30.9%. In contrast, proteins from gcMeta and Mariana Trench sediments exhibited size distributions similar to that of Swiss-Prot, with a median length of approximately 300 aa (**SFig. 8**). We thus only selected the proteins with a >200 aa size in GOMC for the following analyses. The microbiomes from gcMeta, together with the Mariana Trench sediments, span 25 environments including marine, freshwater, terrestrial, and extreme habitats. Their enzyme proportions in coding genes vary noticeably, ranging from 40% (in Mariana Trench sediment and groundwater) to 51.1% (in marine seawater), with the majority between 43% and 46% (**Fig. 2b**). On the other hand, the distributions of major level-1 EC number enzymes are similar across microbiomes from the 25 environments (**Fig. 2c**). Oxidoreductases (EC 1) and transferases (EC 2) together account for approximately half of the enzymes in all these microbiomes, while transferases and hydrolases (EC 3) are the two most abundant classes in each environment. Notably, groundwater microbiomes harbor the lowest proportion of oxidoreductases, while those from Mariana Trench sediments show proportions comparable to other environments. Thus, as an extreme case, we further analyzed 14.3 million "dark matter" proteins originating from bacteria in NMPFamsDB[4,37]**(STab. 3)**. Although only 2.21% (n = 320,930) of these were identified as enzymes, the major functional classes of these enzymes followed a pattern broadly similar to those in other databases and environments, albeit with substantially more hydrolases (accounting for 36.7%) and fewer ligases and translocases (**SFig. 9**).

Abundances of enzymes in the atlas varied dramatically across different functional classes (**Fig. 2d** and **SFile. 3**). The largest level-2 EC class was the phosphorus-containing group transferases (EC 2.7), representing over 0.1 billion sequences; whereas the smallest classes are EC 6.7 and EC 3.12, each of which contained two enzymes annotated in Swiss-Prot. That is, RAMER did not assign any enzymes in the billion-scale dataset to these two level-2 EC numbers (**Fig. 2e**), while three other level-2 EC classes with <10 Swiss-Prot enzymes expanded to tens to hundreds of thousands of enzymes. Overall, enzyme abundances across level-2 EC classes scaled linearly, with a median expansion factor of approximately 3600-fold relative to enzymes in Swiss-Prot (linear regression R^2 = 0.875, **Fig. 2e**).

We next compared the distributions of enzyme abundances across functional classes from diverse environments, i.e., the correlation between functional profiles of

microbial enzymes and their microbiome habitats. The enzymes from Swiss-Prot include eukaryotic origins and might reflect research bias of the scientific community, leading to a significantly different distribution; those from MGnify and GOMC had length bias in enzymes possibly induced by assembling completeness (**Fig. 2f**). In contrast, the datasets from gcMeta and Mariana Trench sediment are likely free from such biases, and the comparative analysis among environments was therefore based on them. The microbial enzyme atlas exhibited striking similarities among different environments: the majority of pairwise comparisons (88.3%) yielded a Pearson's correlation coefficient greater than 0.95 at the level-3 EC classes (**Fig. 2f**). Outliers were present, and to identify subtle differences in their functional profiles, we next performed a Z-score (distance from the mean measured in standard deviations) analysis at level-3 EC classes across environments (**SFile. 4**).

Based on the Z-score profiles, the enzyme atlas of the 25 environments could be roughly clustered into three groups (**Fig. 2g**). Two of the groups were respectively represented by several extreme habitats and terrestrial habitats, harboring enzymes overrepresented in different functional classes. The third group is a mixture of habitats except for terrestrial. Notably, the groundwater microbiome displayed the most divergent functional profile away from those of other environments (**Fig. 2g** and **SFig. 10a**). We therefore examined its functional profile at higher resolution. At the level-4 EC classes, ligases forming aminoacyl-tRNA and related compounds (EC 6.1.1.-) were consistently and exceptionally overrepresented in groundwater compared to other environments (**Fig. 2h**). This is possibly ascribed to the natural selection of microorganisms with efficient protein synthesis machinery in the stable and oligotrophic groundwater. We also observed that oxidoreductases with an iron-sulfur protein as acceptor (EC 1.2.7) were consistently overrepresented in microbiomes from extreme and sedimentary environments (**SFig. 10b**), possibly due to the dominance of *Archaea* in these microbial communities.

RAMER provides fine-grained clustering of enzymes related to functional domains beyond EC classes

RAMER is a pretrained multimodal model for enzymes, and its embeddings can be used for fine-grained clustering of enzymes at a resolution beyond EC classes. Therefore, we explored functional implications of the RAMER-derived enzyme

clusters. This question becomes critical as the enzyme atlas expands to billion-scale, turning the majority of EC classes into highly abundant ones. For example, 550 tyrosine-tRNA ligases (EC 6.1.1.1) are deposited in Swiss-Prot (March 2026), and the number increased to 815,776 in the atlas. We downsampled nearly 88.5 thousand EC 6.1.1.1 enzymes from the 25 environmental microbiomes, and they formed a complex pattern of visual RAMER-derived clusters in the UMAP projection (**Fig. 2i**). EC 6.1.1.1 enzymes were exceptionally overrepresented in groundwater microbiomes (Z-score = 3.77); however, distribution of the clusters showed no correlation with the groundwater origins (clustering metrics: ARI = –0.004, NMI = 0.013). To assess whether this pattern holds generally, we respectively calculated clustering metrics on all level-4 EC classes containing more than 100,000 enzymes (n = 645). The clustering metrics were uniformly low, indicating that the RAMER-derived clusters were independent of microbial habitats (**Fig. 2j**).

Given that these fine-grained clusters did not reflect environmental origins, we employed plastic-degrading enzymes as a case study to elucidate the biological implications of RAMER-derived clusters. We first examined how the RAMER organized one type of plastic-degrading enzymes, PETases (poly(ethylene terephthalate) hydrolases, EC 3.1.1.101). Ten enzymes in Swiss-Prot are annotated as PETases, and they formed a tight visual cluster in the RAMER-derived UMAP projection of Swiss-Prot enzymes (**Fig. 3a**, left). By comparison, in the billion-scale atlas, 71.46 thousand enzymes were assigned to EC 3.1.1.101, forming a cluster within the 4.8 million EC 3.1.1 enzymes (**Fig. 3a**, middle). After down-sampling and filtering for protein completeness and presence of the conserved Gly-X-Ser-X-Gly motif, we retained 778 predicted PETases for further analysis (**SFile. 5**).

Together with 63 known nature PETases assigned to EC 3.1.1.101 (**SFile. 6**), the 778 predicted EC 3.1.1.1 enzymes were organized into seven clearly separated RAMER-derived clusters (**Fig. 3a**, right). Except for PET27 that harbors an extra domain (A0A330MQ60, removed from UniProtKB; predicted as EC 3.1.1.3 by RAMER), all the known nature PETases fell into one cluster. We found that these clusters exhibited overall phylogenetic consistency (**Fig. 2b**, left), as enzymes of each cluster tended to be closely distributed in the phylogenetic tree. By contrast, the UMAP projection based on ESM-embeddings was distinctly different from that derived by RAMER; many phylogenetic branches could not be distinguished by ESM-derived

clustering (**Fig. 3b**, right), which might be attributed to the low sequence homology among these sequences (**SFig. 11a**).

Next, we predicted protein structures of these enzymes by using AphaFold3 and performed multiple structural alignment (MSTA) analysis. Despite the low sequence homology, these enzymes exhibit overall similarity in their structures (**SFig. 11b**), and RAMER-derived clusters exhibited coherence with the MSTA (**Fig. 3c**, left). By contrast, the GearNet embeddings, which are based on protein structures, organized these enzymes into two clusters that could not be mapped onto the MSTA tree (**Fig. 3c**, left). We examined structural conformations and selected representatives from the seven RAMER-derived clusters (**Fig. 3c**, right; representative sequences in **SFile. 5**). Interestingly, while the members from cluster 1 adopted the well-recognized conformation of known PETases, those from other clusters exhibited similar but divergent conformations around the Ser-His-Asp catalytic triad (**Fig. 3d**; except for cluster 5 whose triad was incomplete). For example, compared to cluster 4, clusters 2, 3, 5 and 6 had fewer β-strands in the sheet, and clusters 4–7 all contained additional extra domains.

Taken together, RAMER organized these predicted PETases in a manner that aligns with both phylogenetic and structural coherence, forming tighter clusters than those obtained with ESM or GearNet embeddings. Nevertheless, the clustering could not be fully recapitulated by sequence or structural similarity alone. This raises an intriguing question: do the enzymes in RAMER-derived clusters differ in catalytic activities?

To address this, we analyzed another plastic-degrading enzyme, PURases, enzymes with hydrolytic activity toward polyurethane. Recently, Chen et al. reported the discovery of several efficient PURases based on a urethanase termed Aes72[38] by using the GARSE framework of enzyme mining[19,39]. PURases currently are not assigned an EC number, and we directly calculated the RAMER-embedding cosine similarity (cos-sim) values between Aes72 and the 46 GARSE-mined nature PURases in ref.[19]. We found that the majority (44 of 46) of them had a RAMER embedding with a high cos-sim (>0.95), and the active ones showed significantly higher similarities than those without detectable activity or available data (two-sided Wilcoxon rank-sum test, $P = 0.013$, **Fig. 3e**). Moreover, we observed a positive correlation between the activities

and the RAMER-derived cos-sim values (Pearson's correlation coordinate 0.413, activities on longitude scale, **Fig. 3f**).

The lid domains of PURases play a critical role in the catalysis, and the conformations of lid directly affect activities[19]. While Aes72 possesses conformation with high activity, we speculate that the RAMER-embedding might reflect the conformation, which led to the positive correlation. Therefore, to explore the lid conformation in a wider context, we conducted an *in silico* mining analysis of PURases using RAMER with Aes72 as the query (illustrated in **Fig. 3g**). From NCBI genome database, we retrieved proteins of 200–400 aa from *Actinomycetota* (a taxonomic group from which some PURases originate) in the NCBI genome database. Using RAMER embeddings, 23 million (53.66%) proteins were predicted as enzymes, and their cos-sim values to Aes72 were calculated. Interestingly, while the cos-sim distribution was generally skewed toward low values, we observed a small but observable enrichment of enzymes around 0.97 (**Fig. 3g**, bottom). A total of 5,313 *Actinomycetota* enzymes exhibited a cos-sim ≥0.95 cos-sim; among these, 5053 (95.1%) possessed the Ser-His-Glu catalytic triad and 230 (4.33%) possessed the Ser-His-Asp triad, both characteristic of the α/β-hydrolase fold-3 (ABH3) superfamily (**SFile. 7**).

The RAMER-derived UMAP projection shows that the 5313 enzymes were organized in several clusters, with the majority of validated enzymes in ref.[19] falling within some of the major clusters and the query Aes72 located at the center (**Fig. 3h** and **SFig. 12a-b**). We inspected the enzymes in the outer clusters and found that, compared to the predicted PETases in **Fig. 3d**, these enzymes were generally more similar in overall structure, with differences primarily manifested in the lid domain (**SFig. 12c**). We observed a divergence in lid conformations among these nature enzymes from *Actinomycetota*, including a crossed alpha-helix lid (the dark grey cluster in **Fig. 3h**, right), a single lateral alpha-helix lid (the purple cluster in **Fig. 3h**, right), and a shortened helical lid (the teal cluster **in Fig. 3h**, right). Next, we examined the enzymes adjacent to the reported ones and, intriguingly, they largely shared the typical lid conformations (**Fig. 3h**, left, and **SFig. 12c**). For example, GB6, which exhibits high activity, and its adjacent enzymes formed a V-shaped architecture, whereas the low-activity F4 and inactive GF4 had a closed lid. GF5 is a low-activity enzyme without a lid; adjacent to GF5, we identified another lidless enzyme, as well as one with a single-helix lid.

RAMER appears to capture the characteristics of the critical function lid domain of the PURases. To further validate this, we analyzed adenylate kinases (ADKs, EC 2.7.4.3). ADKs are conventionally classified into three types by their lid domains: lidless, H-bond lid, and Zn^{2+}-binding lid. Recently, Muir et al. experimentally assessed the catalytic activities of ADK orthologs across *Bacteria* and *Archaea* and demonstrated that among the three types, the lidless ADKs possess lower catalytic activities, and Zn^{2+} lid domains are associated with higher growth temperatures[40]. We obtained RAMER embeddings of the 175 ADK assessed in ref.[40] (**SFile. 6**). Remarkably, in the RAMER-derived UMAP projections, these ADKs formed three distinctly separated clusters, aligned precisely with the three ADK types (**Fig. 3i**, left; three exceptions due to annotation error in ref.[40], see **SFig. 13**). In contrast, embeddings from ESM, ProtT5, GearNet, and the CLEAN modules failed to distinguish among the three types of ADKs (**Fig. 3i,** right).

We further explored the correlation between RAMER embeddings and the catalytic activities of ADK reported in ref.[40]. However, we observed a rugged catalytic landscape within each ADK type, similar to that obtained with other PLMs (**SFig. 14**). For an exploratory analysis, we compared using RAMER and other PLMs for enzyme activity prediction using the methods and test set from the study of CataPro[17]. The results showed that RAMER had no advantage, suggesting that RAMER is more suitable for downstream classification tasks (**STab. 4**).

We therefore constructed a classifier for the three types based on RAMER embeddings, trained and validated on the 175 ADKs (see Methods, **SFig. 15a-d**). We then applied the classifier to the 180,742 enzymes assigned to 2.7.4.3 from the gcMeta and Mariana Trench sediment microbiomes. The visualization of RAMER-derived UMAP projection showed consistency with the classification (ARI = 0.524 and NMI = 0.558; **Fig. 3j**). Among the three types, Zn^{2+}-binding lid ADKs were the most abundant, followed by lidless ADKs at a comparable proportion, whereas the ADKs with H-bond lids were considerably less abundant in this set of 180,742 (**Fig. 3k**, upper). This result is consistent with the notion that the H-bond catalytic motif is evolutionarily more difficult to acquire than the Zn^{2+}-binding or lidless configurations, as it involves more complex residue interactions[40]. Analysis of environmental origins revealed that, compared to lidless ADKs, H-bond lid ADKs were less represented in extreme habitats, whereas Zn^{2+} lid ADKs were relatively overrepresented (**Fig. 3k**, bottom, and **SFig.**

15e).

Mining and validation of fluorinases and PETases by using RAMER

To experimentally validate the utility of RAMER in enzyme mining, we focused on two case studies: (*i*) fluorinases that are naturally rare and relatively conserved, to examine the consistency of RAMER annotations with those in existing databases such as UniProtKB/TrEMBL; and (*ii*) PETases from a newly released microbiome dataset of Mariana Trench sediments[36], to explore the conformational diversity of PET-degrading enzymes.

Fluorinases (EC 2.5.1.63) are relatively rare enzymes that predominantly originate from *Actinomycetota* bacteria. Fluorinases function as homotrimers, with each monomer consisting of ~300 aa. As illustrated in **Fig. 4a**, we retrieved 5.68 million 250–350 aa proteins from *Actinomycetota* from the UniProtKB, among which 3.22 million were identified as enzymes. This protein set includes eight known experimental validated nature fluorinases. Using RAMER, we predicted EC numbers for these sequences, and 103 of the 3.22 million enzymes were assigned to EC 2.5.1.63. Sixteen of them have a high reliance of ≥0.99 cos-sim, including all the eight nature's known fluorinases. The rest eight were all annotated as fluorinase in UniProtKB, albeit with an unreviewed status and the lowest annotation score (1/5).

Phylogenetic analysis revealed that these 16 sequences could be divided into three branches (**Fig. 4b**). We selected five candidates for crude enzyme validation and found that two (UniProt accessions A0A1G9FQX8 and A0ABP8UHB8) exhibited fluorinase activity, catalyzing the conversion of S-adenosyl-L-methionine (SAM) to 5′-fluoro-5′-deoxyadenosine (5'-FDA). It should be noted that the outcome could be affected by the incompleteness of the N-terminal region, as the recently reported FlA^{Smor} restored activity after N-terminal complementation[41]; in contrast, in our crude enzyme validation, the corresponding enzyme (A0A7Y7B3E7) did not yield detectable product. We further validated the fluorinase activity in a purified enzyme system, confirming the activity of these two candidates (**Fig. 4c**).

The fluorinases are generally conserved, particularly in the catalytic domain (**SFig. 16**). Thus, the RAMER-based mining of fluorinases indicates the consistency of RAMER annotations with those from other tools, such as the automatic annotation in UniProtKB. To further explore the utility of RAMER in enzyme mining, we conducted

a mining of PETases in the newly released microbiome dataset from Mariana Trench sediments[36] (**Fig. 4d**). From over 7,000 species-level metagenome-assembled genomes, 20.7 million proteins were annotated, of which 8.27 million were classified as enzymes based on RAMER embeddings. Using RAMER max-sep model, 104 enzymes were predicted as PETases in EC 3.1.1.101. Among these, 30 possessed the characteristic Gly-X-Ser-X-Gly catalytic motif, and 28 of them also retained the conserved Ser-His-Asp catalytic triad[42], as in *Is*PETase [43]. Then, we deliberately selected 21 enzymes that were not identified as EC 3.1.1.101 enzymes by other tools or that had low confidence scores (**SFile. 8**).

We screened the 21 candidates for ester-hydrolysing activity using p-nitrophenyl acetate (*p*NPA) as a model substrate (**Fig. 4e**). Most candidates were active, although activity varied markedly across the panel. FD40, FD53 and FD65 showed the highest rates (0.68, 0.52, and 0.60 μM/s), exceeding that of benchmark *Is*PETase. FD36 and PD63 showed intermediate activity, whereas several others were close to the benchmark. Considering that many of the 21 enzymes were not assigned to EC 3.1.1 by other models, RAMER selected the valid candidates with detectable ester-hydrolysing activity. We examined their RAMER-derived clusters and found that five were in cluster 2 defined in **Fig. 3a** (FD28, FD40, FD54, FD71, FD98) and the rest in cluster 1 (**SFig. 17a**). These in cluster 1 and cluster 2 had no significant difference in ester-hydrolysing activity (two-sided Wilcoxon rank-sum test, $P = 0.905$, **SFig. 17b**). In phylogenetic or MSTA clustering, these validated enzymes in RAMER-derived clusters 2 and 5 were distributed in multiple branches (**SFig. 17c-d**).

We next measured PET depolymerization for candidates that with relatively high ester-hydrolysing activity (FD36, FD40, FD53, FD65 and FD71) using GfPET film and quantified the total amount of terephthalic acid (TPA) and mono(2-hydroxyethyl) terephthalate (MHET) released after 24 h by HPLC (**Fig. 4f**). At 37°C, the candidates were already clearly stratified in performance, with FD65 showing the highest product titre (0.44 mM TPA+MHET), followed by FD36 (0.33 mM). Compared to *Is*PETase, these two PET hydrolases displayed 21.8- and 16.6-fold higher activity, respectively. At 50°C, product release increased markedly for all candidates. FD53, FD36 and FD71 were the most active, each reaching 2.8, 2.7 and 2.6 mM degradation products (TPA+MHET), corresponding to 5.2-, 5.0-, and 4.9-fold higher titres than *Is*PETase.

It should be noted that FD40 and FD71, which were assigned to the RAMER-derived cluster 2 of PETases (**Fig. 4g**), exhibited a relatively more distinct phylogeny and protein architectures from *Is*PETase than FD65, FD53, and FD36 (**SFig. 18** and **SFig. 19**), particularly a long β-strand (**Fig. 4h**). FD65, FD53, and FD36 belong to cluster-1 PETases, where other known PETases are located; they also differed from *Is*PETase in terms of absent or additional α-helices (denoted as green in **Fig. 4h)**. Together, with RAMER, we efficiently retrieved active PETases bearing unique structural conformations from more than 20 million proteins.

Discussion

With the pretrained model RAMER, we expanded the atlas of microbial enzymes with EC functional classification to 1.04 billion entries, constituting a resource for mining and exploring enzymes from the vast "microbial dark matter"[4]. The majority of level-4 EC classes become highly abundant with enzymes in the atlas, and RAMER enables a finer classification of enzymes, as seen in plastic-degrading enzymes and adenylate kinases, providing a functional-centric organization alternative to conventional sequence- or structure-based approaches and PLMs.

RAMER shows a substantial advantage in recall for predicting level-4 EC numbers of enzymes, which is particularly important for enzyme mining. We attribute this improvement to the reaction-aware multi-modal pre-training combined with a zero-shot prediction strategy. During training, we included sequence, structure and catalytic reaction information, all of which are biologically meaningful; we did not incorporate artificial annotations such as EC classes, Gene Ontology terms, or human language descriptions of enzymes. Ablation studies demonstrated the contribution of the structural and reaction modalities on top of sequences embedded by PLMs, and the zero-shot prediction using a background set of enzymes with known EC numbers outperformed supervised approaches. The zero-shot strategy is virtually indispensable for the billion-scale prediction task, since small inaccuracies in classifiers could lead to large numbers of false positives. As an illustration, no enzymes were assigned to EC 6.7 or EC 3.12 among the 1.04 billion entries using the zero-shot strategy.

The billion-scale enzyme atlas greatly expands the known space of enzymes, providing an enzymology perspective to interrogate microbial ecosystems across diverse environments. Enzymes account for approximately half of the microbial

proteins encoded in genetic materials. Microbial communities can vary substantially among habitats, such as those respectively dominated by archaea and bacteria. However, ensembles of enzymes encoded in genomes of microbiomes from various environments exhibit striking overall consistency in their profiles of EC classes. Compared to the average level, certain functional classes are under- or over-represented, yet the deviations are relatively modest, for which a Z-score of three is already considered large (e.g., EC 6.1.1 in groundwater microbiome). The phenomenon that the functional composition remains stable aligns well with the previous finding of decoupling of function and taxonomy in the global ocean microbiome[44]. Fridman et al. also reported that the intrinsic structure of metabolic networks strongly constrains their evolutionary trajectories, allowing only a few specific metabolic hierarchies to be selected and stably maintained[45]. Together, the consistent profiles of microbial enzyme classes across distinct environments suggest that core biochemical operations are under strong evolutionary constraints.

We are also intrigued by the high diversity of enzymes for specific catalytic reactions, as revealed by RAMER and the atlas. We inspected plastic-degrading enzymes, ADKs and fluorinases; they exhibit conserved overall scaffolds in their protein structures, consistent with the observation that global folds of enzymes are under fundamental evolutionary constraints[46]. Within these overall scaffolds, except for fluorinases which are highly conserved, we observed vivid examples in plastic-degrading enzymes that microbiomes possess various conformations, particularly in the functionally critical lid domains. RAMER appears to organize PURases and ADKs in a manner that reflects critical lid domains, which cannot be achieved using whole-protein structures. Namely, RAMER and the atlas could provide a broad and potentially function-centric organization of sequence space for specific enzymes, which might greatly facilitate the exploration of the sequence–catalytic functional landscape of enzymes.

On the other hand, RAMER embeddings are computationally much cheaper than AI protein structure prediction, and the protein structures in AFDB and ESMAtlas are relatively limited compared to the vast protein universe. A more feasible strategy is to first mine and cluster the enzymes of interest using RAMER, either from the atlas or from customized protein sequence datasets (together with validated ones), and then select representatives for protein structure prediction in further investigations.

The performance of functional annotations using RAMER and the zero-shot strategy could be further improved, as indicated in **Fig. 1h and 1i**, by expanding the background datasets of rare EC classes. Since the billion-scale atlas provides abundant candidate sequences, further expansion of the background set could be achieved through, for example, *in silico* structural analysis, literature curation of active enzymes, or additional experimental validations. The vast enzyme space is far from being sufficiently sampled, and it is of considerable interest to further apply AI to explore the microbial dark matter, which represents both a major challenge and a significant opportunity for advancing our understanding of microbial functional diversity[6,47]. RAMER contributes new tools and data to the existing toolbox and resource repertoire.

In this study, we did not investigate reaction- or substrate-based mining of enzymes, despite RAMER being a multi-modal model. This is because: (*i*) RAMER embeddings showed no major advantage over MolT5 for reactions, and reactions were scarce (~1:20) in training; (*ii*) RAMER's multimodal architecture aligns reaction and structure embeddings to sequences via separate losses, making it unsuitable for direct reaction-to-sequence mining. Very recently, several studies have employed dual-modal AI architectures to address this problem[48,49]. Cui et al. reported EZSpecificity, which employs a cross-attention-enhanced SE(3)-equivariant graph neural network to predict enzyme–substrate pairings from sequence and structure[50]. While these studies have validated mined enzymes for target reactions, further benchmarking is still needed.

Finally, beyond enzyme mining, RAMER and the atlas might be insightful for enzyme design and engineering. Nature enzymes have long-range evolutionary constraints: residues distant from the active site are under strong purifying selection, largely driven by catalytic function rather than by the maintenance of overall protein fold[51]. However, current approaches focus primarily on the catalytic pockets, such as directed evolution and the recent *de novo* AI design of serine hydrolases[52]. The atlas could facilitate harnessing nature's evolutionary diversity, shaped over billions of years of microbiome evolution, to advance rational and data-driven AI enzyme design.

References

1. Tenorio-Salgado, S., Villalpando-Aguilar, J. L., Hernandez-Guerrero, R., Poot-Hernández, A. C. & Perez-Rueda, E. Exploring the enzymatic repertoires of Bacteria and Archaea and their associations with metabolic maps. *Braz. J. Microbiol.* **55**, 3147–3157 (2024).

2. Griesemer, M., Kimbrel, J. A., Zhou, C. E., Navid, A. & D'haeseleer, P. Combining multiple functional annotation tools increases coverage of metabolic annotation. *BMC Genomics* **19**, 948 (2018).

3. Freilich, S. *et al.* The Complement of Enzymatic Sets in Different Species. *J. Mol. Biol.* **349**, 745–763 (2005).

4. Pavlopoulos, G. A. *et al.* Unraveling the functional dark matter through global metagenomics. *Nature* **622**, 594–602 (2023).

5. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).

6. Alanzi, A. R. Exploring Microbial Dark Matter for the Discovery of Novel Natural Products: Characteristics, Abundance Challenges and Methods. *J. Microbiol. Biotechnol.* **35**, (2025).

7. Abramson, J. *et al.* Addendum: Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **636**, E4–E4 (2024).

8. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

9. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. (2023).

10. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* **118**, e2016239118 (2021).

11. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).

12. Elnaggar, A. *et al.* ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).

13. Elnaggar, A. *et al.* Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. (2023).

14. Yu, T. *et al.* Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).

15. Liu, Y. *et al.* EnzymeCAGE: A Geometric Foundation Model for Enzyme Retrieval with Evolutionary Insights. Preprint at https://doi.org/10.1101/2024.12.15.628585 (2024).

16. Boorla, V. S. & Maranas, C. D. CatPred: a comprehensive framework for deep learning in vitro enzyme kinetic parameters. *Nat. Commun.* **16**, 2072 (2025).

17. Wang, Z. *et al.* Robust enzyme discovery and engineering with deep learning using CataPro. *Nat. Commun.* **16**, 2736 (2025).

18. Santos, C. A. *et al.* A metagenomic 'dark matter' enzyme catalyses oxidative cellulose conversion. *Nature* **639**, 1076–1083 (2025).

19. Chen, Y. *et al.* Glycolysis-compatible urethanases for polyurethane recycling. *Science* **390**, 503–509 (2025).

20. Yoon, P. H. *et al.* Structure-guided discovery of ancestral CRISPR-Cas13 ribonucleases. *Science* **385**, 538–543 (2024).

21. Fleming, J. *et al.* AlphaFold Protein Structure Database and 3D-Beacons: New Data and Capabilities. *J. Mol. Biol.* **437**, 168967 (2025).

22. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).

23. Edwards, C. *et al.* Translation between Molecules and Natural Language. in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (eds Goldberg, Y., Kozareva, Z. & Zhang, Y.) 375–413 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022). doi:10.18653/v1/2022.emnlp-main.26.

24. Zhang, Z. *et al.* Protein language models learn evolutionary statistics of interacting sequence motifs. *Proc. Natl. Acad. Sci.* **121**, e2406285121 (2024).

25. Zhang, Z. *et al.* Protein Representation Learning by Geometric Structure Pretraining. (2023).

26. Radford, A. *et al.* Learning Transferable Visual Models From Natural Language Supervision. in *International Conference on Machine Learning* (2021).

27. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018).

28. Healy, J. & McInnes, L. Uniform manifold approximation and projection. *Nat. Rev. Methods Primer* **4**, 82 (2024).

29. Bansal, P. *et al.* Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res.* **50**, D693–D700 (2022).

30. Song, Y. *et al.* Accurately predicting enzyme functions through geometric graph learning on ESMFold-predicted structures. *Nat. Commun.* **15**, 8180 (2024).

31. Kim, G. B. *et al.* Functional annotation of enzyme-encoding genes using deep learning with transformer layers. *Nat. Commun.* **14**, 7370 (2023).

32. Sanderson, T., Bileschi, M. L., Belanger, D. & Colwell, L. J. ProteInfer, deep neural networks for protein functional inference. *eLife* **12**, e80942 (2023).

33. Su, J. *et al.* A trimodal protein language model enables advanced protein searches. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-025-02836-0 (2025) doi:10.1038/s41587-025-02836-0.

34. Sun, Y. *et al.* gcMeta 2025: a global repository of metagenome-assembled genomes enabling cross-ecosystem microbial discovery and function research. *Nucleic Acids Res.* **54**, D724–D733 (2025).

35. Chen, J. *et al.* Global marine microbial diversity and its potential in bioprospecting. *Nature* **633**, 371–379 (2024).

36. Xiao, X. *et al.* Microbial ecosystems and ecological driving forces in the deepest ocean sediments. *Cell* **188**, 1363-1377.e9 (2025).

37. Baltoumas, F. A. *et al.* NMPFamsDB: a database of novel protein families from microbial metagenomes and metatranscriptomes. *Nucleic Acids Res.* **52**, D502–D512 (2024).

38. Bittner, N. *et al.* Novel urethanases for the enzymatic decomposition of polyurethanes. *Eur. Pat. Appl. EP3587570A1* (2020).

39. Wu, L., Wang, L., Wang, F., Hu, W. & Zhang, H. Deep Learning as a Compass for Industrial Biocatalysis: The Grase Framework Rewrites the Rules for Polyurethane Recycling. *Biotechnol. J.* **21**, e70222 (2026).

40. Muir, D. F. *et al.* Evolutionary-scale enzymology enables exploration of a rugged catalytic landscape. *Science* **388**, eadu1058 (2025).

41. Verma, R. K. *et al.* Unveiling the molecular basis of selective fluorination of SAM-dependent fluorinases. *Chem. Sci.* **16**, 10610–10619 (2025).

42. Joo, S. *et al.* Structural insight into molecular mechanism of poly(ethylene terephthalate) degradation. *Nat. Commun.* **9**, 382 (2018).

43. Yoshida, S. *et al.* A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science* **351**, 1196–1199 (2016).

44. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272–1277 (2016).

45. Takano, S., Vila, J. C. C., Miyazaki, R., Sánchez, Á. & Bajić, D. The Architecture of Metabolic Networks Constrains the Evolution of Microbial Resource Hierarchies. *Mol. Biol. Evol.* **40**, msad187 (2023).

46. Davidi, D., Longo, L. M., Jabłońska, J., Milo, R. & Tawfik, D. S. A bird's-eye view of enzyme evolution: chemical, physicochemical, and physiological considerations. *Chem. Rev.* **118**, 8786–8797 (2018).

47. Zha, Y., Chong, H., Yang, P. & Ning, K. Microbial Dark Matter: From Discovery to Applications. *Genomics Proteomics Bioinformatics* **20**, 867–881 (2022).

48. Rocks, J. W. *et al.* Dual-encoder contrastive learning accelerates enzyme discovery. *Proc. Natl. Acad. Sci.* **123**, e2520070123 (2026).

49. Zhou, Z. *et al.* Reaction-Conditioned Enzyme Discovery with Multimodal Deep Learning. *bioRxiv* 2026.03.09.710689 (2026) doi:10.64898/2026.03.09.710689.

50. Cui, H. *et al.* Enzyme specificity prediction using cross-attention graph neural networks. *Nature* **647**, 639–647 (2025).

51. Sharir-Ivry, A. & Xia, Y. Nature of Long-Range Evolutionary Constraint in Enzymes: Insights from Comparison to Pseudoenzymes with Similar Structures. *Mol. Biol. Evol.* **35**, 2597–2606 (2018).

52. Lauko, A. *et al.* Computational design of serine hydrolases. *Science* **388**, eadu2454 (2025).

Figure Captions

Figure 1. Architecture and performance of RAMER. **a**, Schematic of the RAMER pre-training framework. The model integrates three modalities: catalytic reactions encoded via a reaction-aware network, sequence embeddings, and enzyme protein structure embeddings. Projection layers were appended to each encoder to unify dimensions of the three modalities. Cos sim, cosine similarity. **b**, UMAP visualization of RAMER-derived embeddings for over 555,000 reviewed proteins in Swiss-Prot, colored by Enzyme Commission (EC) major classes (EC 1–7) for enzymes and other proteins without annotation of EC number. An inset shows a magnified local region to highlight class-specific clustering. **c-d**, Clustering performance of RAMER and other protein large language models and CLEAN. The analysis was based on 279 thousand Swiss-Prot reviewed enzymes, evaluated at EC levels 1–4. Adjusted Rand index is a supervised clustering metric (**c**) and Silhouette score is an unsupervised clustering compactness metric (**d**). **e**, Performance of RAMER and benchmarked tools at EC levels 1–3 based on the New-392 test set. Precision and recall are provided in **STab. 2**. **f-g**, Performance of RAMER and benchmarked tools at level-4 EC number prediction based on New-392 (**f**) and the home-made Ram-255 test set (**g**). **h**, Performance of level-4 EC number prediction related to enzyme abundance in the training set. Numbers of EC classes in abundance groups are denoted in brackets. **i**, Level-4 EC prediction performance related to sequence similarity between test and training sets. **j**, Prediction performance on level 1–3 EC numbers for enzymes without level-4 EC enzymes. **k**, Ablation study based on the New-392 test set for level-4 EC number prediction. RAMER (max-sep) uses the three modalities. Seq., sequence modality.

Figure 2. RAMER-derived billion-scale enzyme atlas from microbiomes. a, Schematic workflow for constructing the global microbial enzyme atlas. Microbiome datasets were collected from four major habitats (marine, freshwater, terrestrial, and extreme). **b**, Percentage of predicted enzymes in different protein datasets and microbial biomes. For reference databases, proteins from GOMC were filtered to 200–1024 aa. Datasets in gcMeta assigned the 'environments' and 'extreme' were included. Environmental microbiomes follow gcMeta annotation, together with datasets from Mariana Trench Sediment[36]. **c**, Distribution of level-1 EC major classes. **d**, Abundance distribution of the 1.04 billion enzymes across level-2 EC classes, plotted on a logarithmic scale. **e**, Correlation between the number of enzymes in each level-2 EC class in Swiss-Prot and the microbial enzyme atlas, visualized on a log-log scale. **f**, Pearson correlation coefficients (PCC) of level-3 EC functional profiles across databases and biomes. Pairwise PCC values were calculated between each dataset and the 24 gcMeta microbiomes, as well as among themselves, presented as boxplots (middle line, median). *P* values of two-sided Wilcoxon rank-sum tests are indicated. **g,** Heatmap of Z-score normalized level-3 EC class abundances across the environments, with hierarchical clustering shown on the left. Acidic environments (acid mine drainage, acid habitat) and hot environments (hot spring, hot habitat) are merged. Level-1 EC classes are labeled at the top, colored as in **c**. **h**, Relative abundances of level-4 EC

6.1.1.- classes in each environment, with groundwater highlighted by blue circles. **i**, UMAP visualization of RAMER embeddings for EC 6.1.1.1 enzymes, with up to 5,000 sampled from each environment. Groundwater-origin enzymes are highlighted in blue. Clustering indices (adjusted rand index, ARI; normalized mutual information, NMI) calculated for groundwater are denoted aside. **j**, Clustering indices for level-4 EC classes with >100,000 enzymes. The left and right panels are based on 25 individual environments and the four major habitats as in **a**.

Figure 3. RAMER-derived clustering and discovery of conformation diversity of PETases, PURases and adenosine kinases (ADKs). a, RAMER-based clustering of PETases in EC 3.1.1.101 class. The region containing known PETases in the UMAP of Swiss-Prot enzymes is magnified. The EC 3.1.1 enzymes in the billion-scale microbial enzyme atlas are downsampled and exhibited, colored by level-4 EC numbers. These predicted as EC 3.1.1.101 PETases were further downsampled and subjected to RAMER clustering. **b-c,** Maximum-likelihood phylogenetic (**b**) and multiple structural alignment (**c**) analyses of the selected PETases in **a**, colored as the clusters defined in **a**. The UMAP visualizations based on ESM and GearNet are shown at the left of trees. **d**, Representative structures of enzymes from each PETase cluster. Number of β-strands and protein length, with signal peptide trimmed are denoted. Catalytic triads are highlighted in different colors. aa, amino acid. **e-f**, Validation of RAMER for PURase function prediction. Correlations between RAMER-derived cosine similarity (cos-sim) to Aes72 and experimentally validated product formation (**e**) and specific catalytic activity (**f**) are shown, based on data from ref. 19. *P* values from two-sided Wilcoxon rank-sum tests are indicated. **g**, Scheme for PURase mining using RAMER. The cos-sim distribution is shown at bottom, with the high-similarity range magnified. **h**, RAMER-derived clustering of the PURase candidates with high cos-sim values. Validated PURases from ref.[19] are marked with stars. Representative structural conformations are shown inset, with lid domains highlighted in deep blue. More representatives are provided in **SFig. 12**. **i**, Organization of ADKs by RAMER and benchmark models visualized via UMAP. ADKs are colored by structural class. **j**, UMAP of enzymes assigned to EC 2.7.4.3 (ADKs) in the billion-scale atlas. ADKs classified as Zn^{2+} lid, H-bond, and lidless are colored as in **i**; with unclassified enzymes in gray. Clustering indices (Adjusted Rand Index, ARI; Normalized Mutual Information, NMI) for the three classes are indicated. **k,** Statistics of predicted ADK types in the atlas. Upper panel: absolute counts of each ADK class. Lower panel: habitat distribution of each ADK class across extreme, marine, terrestrial, and freshwater environments (see also in **SFig. 15**).

Figure 4. Mining fluorinases in *Actinomycetota* and PETases from Microbiome from Mariana Trench sediments. a, Schematic workflow for fluorinase mining. **b**, Maximum-likelihood phylogenetic tree of predicted fluorinases. Known fluorinases are indicated in parentheses. For crude enzyme validation, the absence of detectable product can be attributed to lack of expression, insolubility, or inactivity. **c**, Enzymatic activity validation of purified fluorinase candidates. **d**, Schematic workflow for mining

plastic-degrading enzymes from the Mariana Trench sediment microbiome. **e**, Carboxylic ester hydrolase activity validation via crude enzyme assays. **f**, PET-hydrolyzing activity validation of purified enzymes, with TPA+MHET production levels shown. **g**, Phylogenetic distribution and RAMER-derived cluster assignments of the validated PETases. **h**, Predicted structures of validated PETases. FD65 and FD53, as well as FD71 and FD40, exhibit high structural similarity, with representative structures shown. Structural domains distinct from *Is*PETase are highlighted in green.

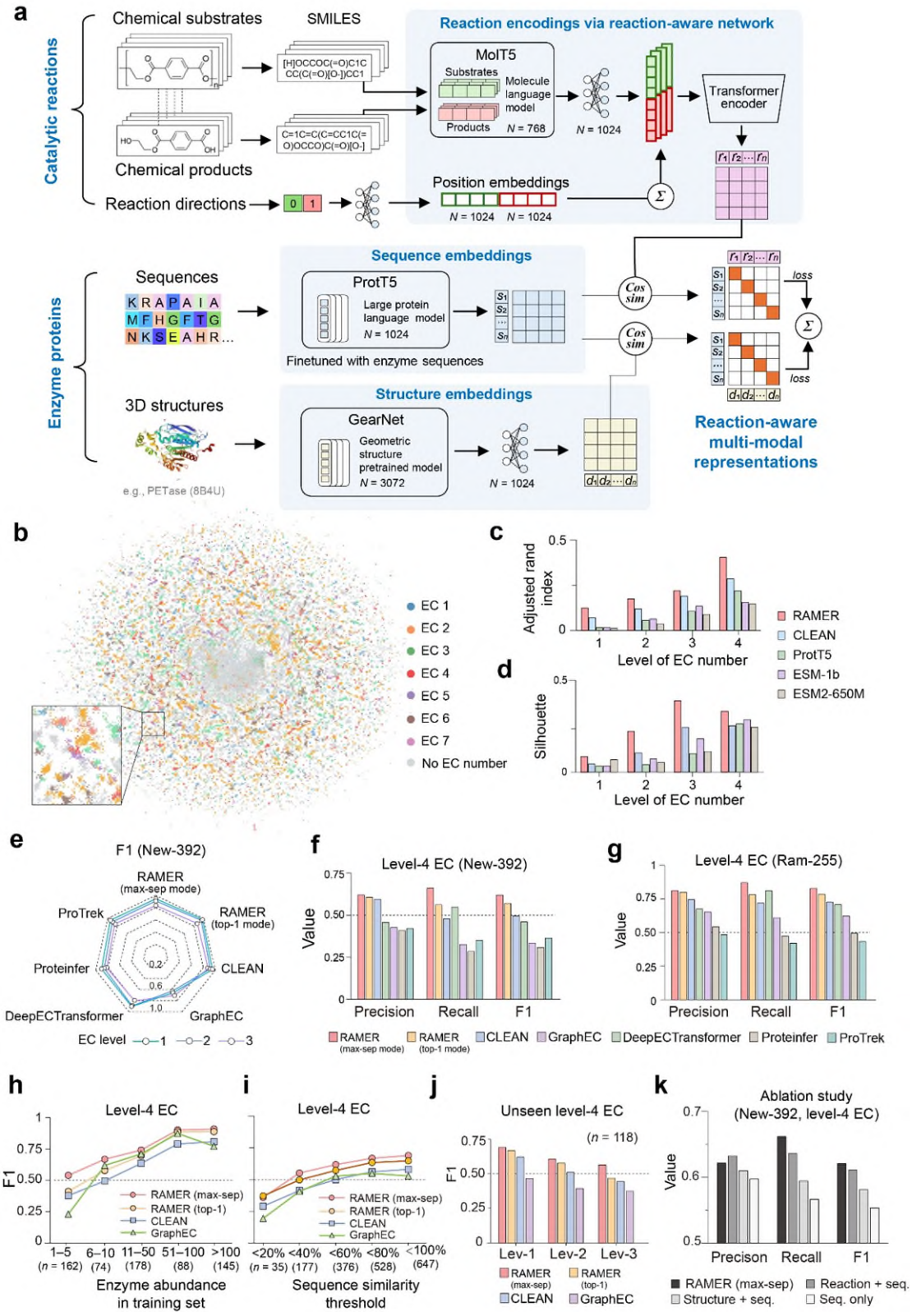


Figure 1. Architecture and performance of RAMER. **a**, Schematic of the RAMER pre-training framework. The model integrates three modalities: catalytic reactions encoded via a reaction-aware network, sequence embeddings, and enzyme protein structure embeddings. Projection layers were appended to each encoder to unify dimensions of the three modalities. Cos sim, cosine similarity. **b**, UMAP visualization

of RAMER-derived embeddings for over 555,000 reviewed proteins in Swiss-Prot, colored by Enzyme Commission (EC) major classes (EC 1–7) for enzymes and other proteins without annotation of EC number. An inset shows a magnified local region to highlight class-specific clustering. **c-d**, Clustering performance of RAMER and other protein large language models and CLEAN. The analysis was based on 279 thousand Swiss-Prot reviewed enzymes, evaluated at EC levels 1–4. Adjusted Rand index is a supervised clustering metric (**c**) and Silhouette score is an unsupervised clustering compactness metric (**d**). **e**, Performance of RAMER and benchmarked tools at EC levels 1–3 based on the New-392 test set. Precision and recall are provided in **STab. 2**. **f-g**, Performance of RAMER and benchmarked tools at level-4 EC number prediction based on New-392 (**f**) and the home-made Ram-255 test set (**g**). **h**, Performance of level-4 EC number prediction related to enzyme abundance in the training set. Numbers of EC classes in abundance groups are denoted in brackets. **i**, Level-4 EC prediction performance related to sequence similarity between test and training sets. **j**, Prediction performance on level 1–3 EC numbers for enzymes without level-4 EC enzymes. **k**, Ablation study based on the New-392 test set for level-4 EC number prediction. RAMER (max-sep) uses the three modalities. Seq., sequence modality.

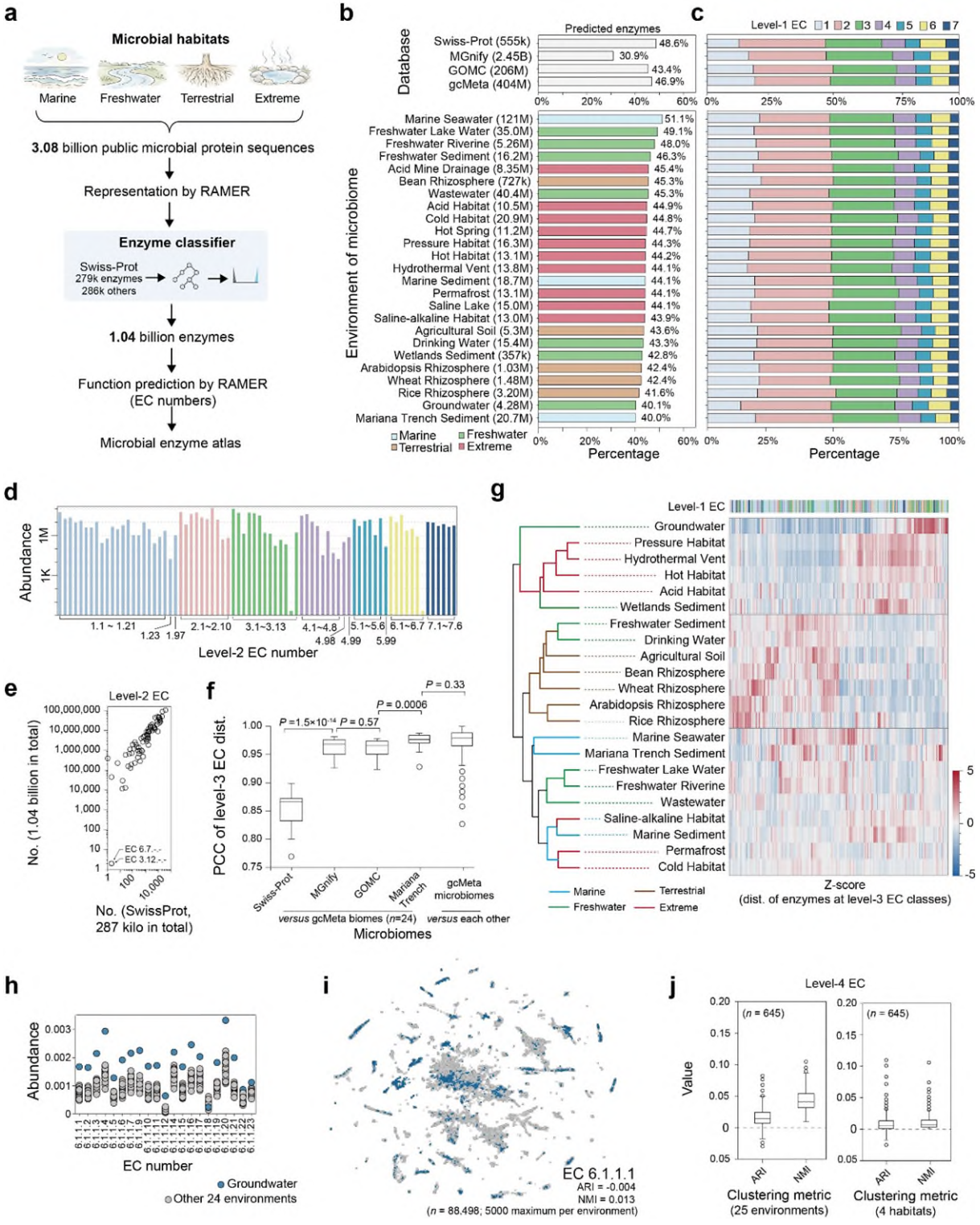


Figure 2. RAMER-derived billion-scale enzyme atlas from microbiomes. a, Schematic workflow for constructing the global microbial enzyme atlas. Microbiome datasets were collected from four major habitats (marine, freshwater, terrestrial, and extreme). **b**, Percentage of predicted enzymes in different protein datasets and microbial biomes. For reference databases, proteins from GOMC were filtered to 200–1024 aa. Datasets in gcMeta assigned the 'environments' and 'extreme' were included. Environmental microbiomes follow gcMeta annotation, together with datasets from

Mariana Trench Sediment[36]. **c**, Distribution of level-1 EC major classes. **d**, Abundance distribution of the 1.04 billion enzymes across level-2 EC classes, plotted on a logarithmic scale. **e**, Correlation between the number of enzymes in each level-2 EC class in Swiss-Prot and the microbial enzyme atlas, visualized on a log-log scale. **f**, Pearson correlation coefficients (PCC) of level-3 EC functional profiles across databases and biomes. Pairwise PCC values were calculated between each dataset and the 24 gcMeta microbiomes, as well as among themselves, presented as boxplots (middle line, median). *P* values of two-sided Wilcoxon rank-sum tests are indicated. **g,** Heatmap of Z-score normalized level-3 EC class abundances across the environments, with hierarchical clustering shown on the left. Acidic environments (acid mine drainage, acid habitat) and hot environments (hot spring, hot habitat) are merged. Level-1 EC classes are labeled at the top, colored as in **c**. **h**, Relative abundances of level-4 EC 6.1.1.- classes in each environment, with groundwater highlighted by blue circles. **i**, UMAP visualization of RAMER embeddings for EC 6.1.1.1 enzymes, with up to 5,000 sampled from each environment. Groundwater-origin enzymes are highlighted in blue. Clustering indices (adjusted rand index, ARI; normalized mutual information, NMI) calculated for groundwater are denoted aside. **j**, Clustering indices for level-4 EC classes with >100,000 enzymes. The left and right panels are based on 25 individual environments and the four major habitats as in **a**.

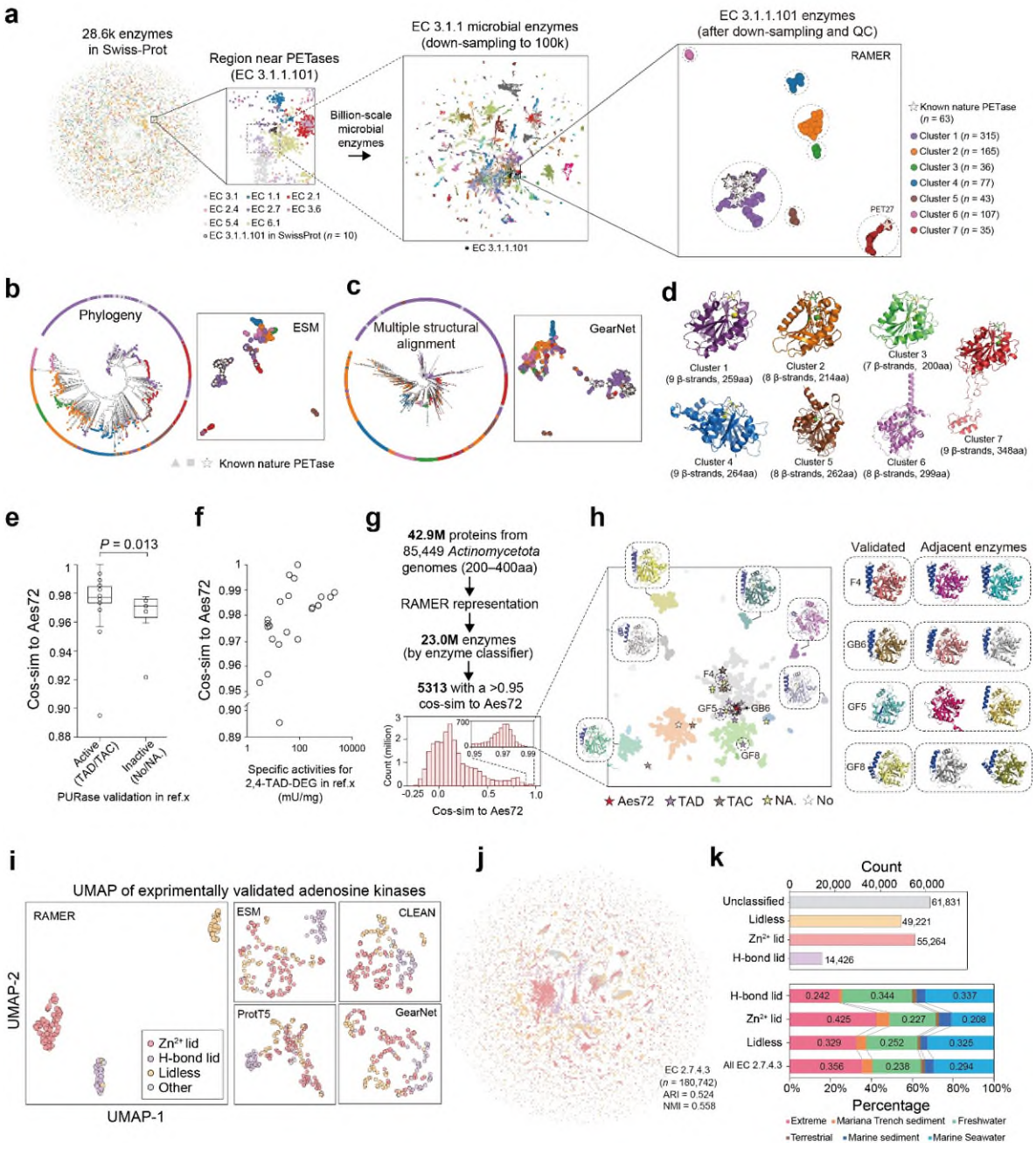


Figure 3. RAMER-derived clustering and discovery of conformation diversity of PETases, PURases and adenosine kinases (ADKs). a, RAMER-based clustering of PETases in EC 3.1.1.101 class. The region containing known PETases in the UMAP of Swiss-Prot enzymes is magnified. The EC 3.1.1 enzymes in the billion-scale microbial enzyme atlas are downsampled and exhibited, colored by level-4 EC numbers. These predicted as EC 3.1.1.101 PETases were further downsampled and subjected to RAMER clustering. **b-c,** Maximum-likelihood phylogenetic (**b**) and multiple structural alignment (**c**) analyses of the selected PETases in **a**, colored as the clusters defined in **a**. The UMAP visualizations based on ESM and GearNet are shown at the left of trees. **d**, Representative structures of enzymes from each PETase cluster. Number of β-strands and protein length, with signal peptide trimmed are denoted. Catalytic triads are highlighted in different colors. aa, amino acid. **e-f**, Validation of RAMER for PURase

function prediction. Correlations between RAMER-derived cosine similarity (cos-sim) to Aes72 and experimentally validated product formation (**e**) and specific catalytic activity (**f**) are shown, based on data from ref. 19. *P* values from two-sided Wilcoxon rank-sum tests are indicated. **g**, Scheme for PURase mining using RAMER. The cos-sim distribution is shown at bottom, with the high-similarity range magnified. **h**, RAMER-derived clustering of the PURase candidates with high cos-sim values. Validated PURases from ref.[19] are marked with stars. Representative structural conformations are shown inset, with lid domains highlighted in deep blue. More representatives are provided in **SFig. 12**. **i**, Organization of ADKs by RAMER and benchmark models visualized via UMAP. ADKs are colored by structural class. **j**, UMAP of enzymes assigned to EC 2.7.4.3 (ADKs) in the billion-scale atlas. ADKs classified as Zn^{2+} lid, H-bond, and lidless are colored as in **i**; with unclassified enzymes in gray. Clustering indices (Adjusted Rand Index, ARI; Normalized Mutual Information, NMI) for the three classes are indicated. **k,** Statistics of predicted ADK types in the atlas. Upper panel: absolute counts of each ADK class. Lower panel: habitat distribution of each ADK class across extreme, marine, terrestrial, and freshwater environments (see also in **SFig. 15**).

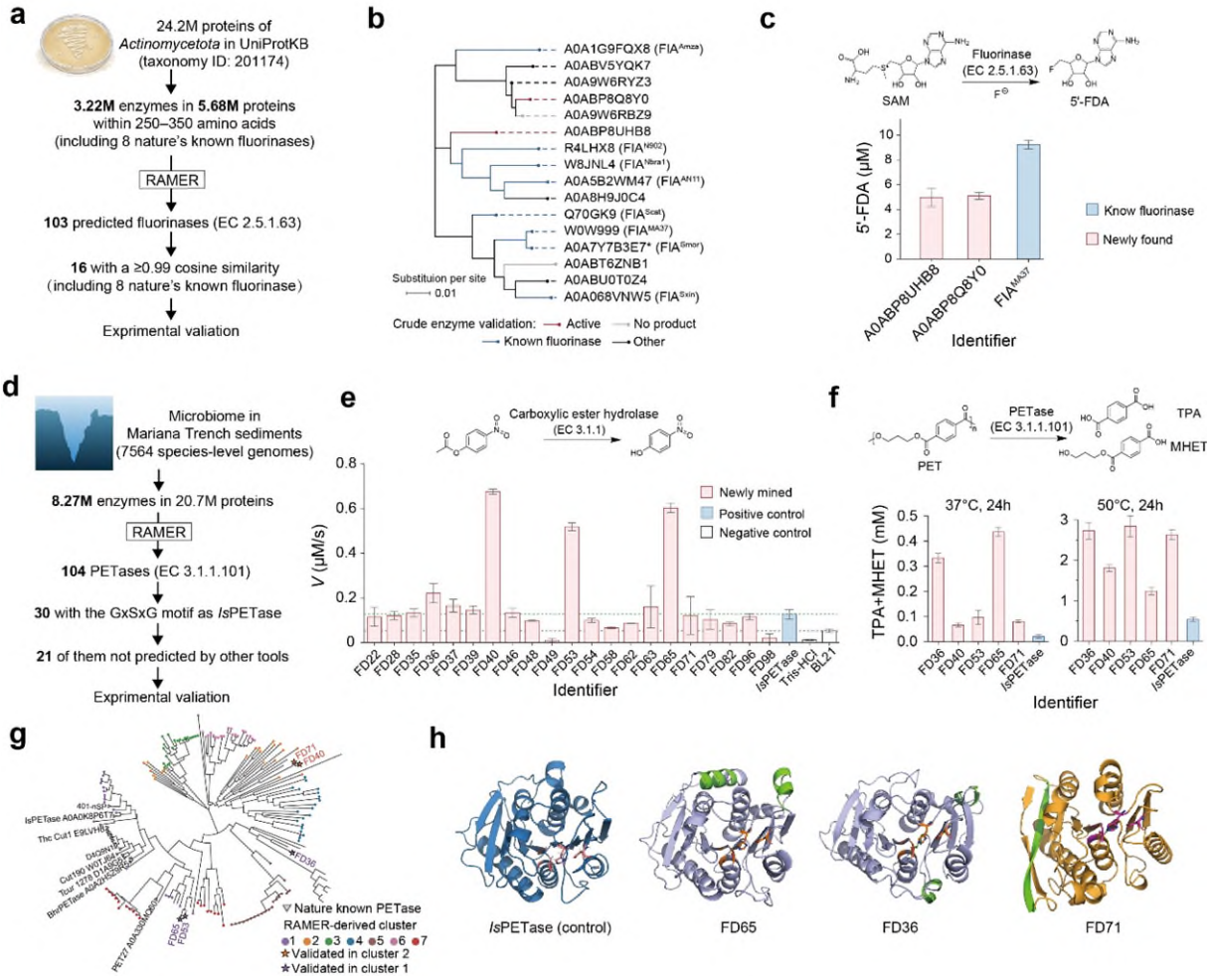


Figure 4. Mining fluorinases in *Actinomycetota* and PETases from Microbiome from Mariana Trench sediments. a, Schematic workflow for fluorinase mining. **b**, Maximum-likelihood phylogenetic tree of predicted fluorinases. Known fluorinases are indicated in parentheses. For crude enzyme validation, the absence of detectable product can be attributed to lack of expression, insolubility, or inactivity. **c**, Enzymatic activity validation of purified fluorinase candidates. **d**, Schematic workflow for mining plastic-degrading enzymes from the Mariana Trench sediment microbiome. **e**, Carboxylic ester hydrolase activity validation via crude enzyme assays. **f**, PET-hydrolyzing activity validation of purified enzymes, with TPA+MHET production levels shown. **g**, Phylogenetic distribution and RAMER-derived cluster assignments of the validated PETases. **h**, Predicted structures of validated PETases. FD65 and FD53, as well as FD71 and FD40, exhibit high structural similarity, with representative structures shown. Structural domains distinct from *Is*PETase are highlighted in green.

Materials and methods

Dataset preprocessing

Training dataset of enzyme sequence, structure, and catalytic reaction triplets

The dataset used for training RAMER comprises protein sequences, protein structures, and enzymatic reactions. Enzyme sequences were obtained from UniProtKB/Swiss-Pro[1] (https://www.uniprot.org, downloaded in Oct, 2024), the corresponding enzyme structures from AFDB (https://alphafold.com/)[2], and catalytic reactions from Swiss-Prot and the Rhea database (https://www.rhea-db.org/)[3]. The dataset construction is also illustrated in **SFig. 1b**.

The construction procedure was as follows. First, 259,184 reviewed protein entries with annotated catalytic reactions were downloaded from Swiss-Prot, and duplicates were removed. Quality control was conducted by removing sequences containing ambiguous amino acids, including B (Asx), Z (Glx), and X (unknown), and non-canonical amino acids, including U (selenocysteine), and O (pyrrolysine). Second, based on the accession numbers from Swiss-Prot, the corresponding AI-predicted structures were downloaded from AFDB; proteins not available in AFDB were removed. Third, enzymatic reaction dataset files, including 'Rhea-reaction-smiles.tsv' and 'Rhea-directions.tsv', were downloaded from the Rhea database. The Rhea entries with UniProtKB enzyme accessions, SMILES-formatted products and substrates, and reaction direction were retained. A small number of reactions are annotated as "undefined" or "bidirectional" in terms of reaction direction in Rhea, and they were treated as left-to-right (LR) reactions. Fourth, structures in PDB format that could not be embedded by the protein structural foundation model GearNet were further removed. Finally, to prevent data leakage, the validation sets used in this study (Price-149, New-389, and Ram-255) were removed from the training dataset. As a result, a total of 267,386 sequence-structure-reaction trios, comprising 200,024 unique enzyme sequences and 10,596 unique enzymatic reactions, were used for RAMER training.

Dataset for EC annotation knowledgebase

In this study, we utilized three versions of datasets containing EC annotations for a set of enzyme sequences, as a knowledgebase for EC number prediction. We downloaded the curated EC datasets employed in the CLEAN tool[4] (https://zenodo.org/records/7787510), and herein, we used the full dataset (split100.csv file, 227,362 enzymes, referred to as CLEAN100) and the partial dataset (split70.csv file, 74,487 enzymes, named as CLEAN70). Sequences in the partial dataset contained enzymes with a <70% sequence similarity.

During the construction of RAMER, CLEAN70 was used for zero-shot prediction of EC numbers to select the optimal epoch and evaluate the performance of the PLMs. In the benchmarking process, the full version CLEAN100 was used to compare RAMER with other tools; CLEAN100 was a common dataset also used by other tools. Furthermore, in the phase of applying RAMER to infer EC annotations for billion-scale protein sequences, we used an updated version of CLEAN100. Specifically, we updated the EC numbers of each protein in CLEAN100 by incorporating all new EC numbers from the latest Swiss-Prot (as of Sep, 2025). This updated version contains the same number of enzymes (n = 227,362) as CLEAN100, among which 10,726 enzymes had their EC numbers updated.

Test and validation datasets to enzyme function annotation

Several benchmarking datasets containing enzyme sequences and EC functional annotations were used in this study. The Price-149 dataset was used for epoch selection during RAMER training. Price-149 is a benchmark dataset consisting of 149 enzyme sequences distributed across 56 different EC numbers[5,6]. The New-392 benchmark dataset, as well as the home-made Ram-255 and unseen level-4 EC datasets, were used for evaluating the performance of EC number prediction. New-392 is a time-split benchmark dataset from UniProtKB/Swiss-Prot after April 2022, spanning 177 different EC numbers[4].

The Ram-255 and unseen level-4 EC datasets are additional time-split benchmark datasets curated in this study. The enzymes in the two datasets were retrieved from UniProtKB/Swiss-Prot database after March 30, 2023. Subsequently, quality control and filtering were performed as follows. First, entries with incomplete EC numbers, enzymes with a sequence length exceeding 1024 aa, and entries containing non-standard amino acids (X, U, O, B, Z) were removed. Second, a 40% sequence identity redundancy reduction was then performed using MMseqs2[7] by using the 'mmseqs easy-cluster' command with the setting '--min-seq-id 0.4' and all other parameters left as default. Third, the enzymes with 100% sequence identity to those in the training dataset or New-392 were removed. After these steps, we obtained 373 entries. Among them, 255 entries had labeled EC numbers that existed in the full CLEAN EC knowledgebase, and they constituted the Ram-255 dataset. The remaining 118 entries had complete level-4 EC numbers that were not included in the knowledgebase, forming the unseen level-4 EC dataset.

RAMER architecture and pre-training

RAMER is a pre-trained tri-modal model that generates representations (embeddings) of enzyme protein sequences. These embeddings can be used for downstream tasks, including clustering analysis, classification of enzymes versus non-enzymes, and EC number prediction. For EC number prediction, RAMER employs a zero-shot strategy, in which a background dataset containing known sequence–EC relationships is used, and predictions are made by finding sequences whose RAMER-derived embeddings show the highest similarity to the query.

The training dataset of RAMER consists of triplets of sequences, structures, and enzymatic reactions. For each modality, a corresponding foundation or representation module was used to generate sequence, structure, and reaction embeddings, respectively. A dual contrastive learning architecture[8] was then employed, with enzyme sequences serving as the central modality. Through training, information from enzyme structures and catalytic reactions was incorporated into the sequence representation network, thereby forming the overall enzyme representation model of RAMER. Details of the RAMER construction are described below and illustrated in **Fig. 1a** and **SFig. 1a**.

Selection of PLMs for the protein sequence representation module

We evaluated four PLMs: ProtT5 (prot_t5_xl_uniref50, available via http://huggingface.co/Rostlab/prot_t5_xl_uniref50), ESM-1b (esm1b_t33_650M_UR50S, http://huggingface.co/facebook/esm1b_t33_650M_UR50S), ESM2-650M (esm2_t33_650M_UR50D, http://huggingface.co/facebook/esm2_t33_650M_UR50D), ESM2-3B (esm2_t36_3B_UR50D, https://huggingface.co/facebook/esm2_t36_3B_UR50D). Both their original versions and fine-tuned versions of enzyme sequences from the training dataset were assessed.

A parameter-efficient fine-tuning technique Low-Rank Adaptation (LoRA)[9] was adopted for fine-tuning the four PLMs. LoRA freezes the pretrained model weights and injects trainable low-rank decomposition matrices into each Transformer layer, significantly reducing memory usage and computational overhead while preserving the model's general capabilities[9]. For both ProtT5 and the ESM series models, the last six layers of the PLM backbone were finetuned using LoRA on the enzyme sequences from RAMER's training set. Additionally, a linear projection layer ($Proj(\cdot)$) with a dimension of 1024 was used for alignment with structures and enzymatic reactions.

The formulation of sequence representation, S, was as follows:

$$S_0 = \mathcal{F}_{PLM}(Seq; \Theta_{PLM} + \Delta\Theta) \qquad (1)$$

$$S = Proj(S_0; W_p) \qquad (2)$$

The Price-149 dataset was used as a benchmark for performance evaluation. EC number prediction was performed using a zero-shot strategy with the CLEAN70 knowledgebase. For each query sequence, the enzyme in the knowledgebase with the highest cosine similarity to the query embedding was identified, and its EC number was assigned as the prediction. The CLEAN70 set was used instead of CLEAN100 because its smaller size reduced computational cost, as embeddings had to be computed eight times (once for each of the eight models evaluated).

Prediction performance was quantified using precision, recall, F1 score, and overall accuracy. Among the non-fine-tuned PLMs, the original ESM2-3B achieved the highest F1 score. However, its fine-tuned version showed decreased performance, particularly a substantial drop in recall. In contrast, while ProtT5 exhibited the lowest performance among the four non-fine-tuned models, its fine-tuned version achieved improved recall and precision, resulting in a slightly higher F1 score than the original ESM2-3B, ranking second among the eight models. The fine-tuned ProtT5 also showed higher overall accuracy, which may be partially attributed to class imbalance in the Price-149 test set. Based on these results, we selected ProtT5 as the enzyme representation backbone model for RAMER.

Protein structural representation module with GearNet

We adopted GearNet as the structure representation module of RAMER. GearNet is a pretrained protein structure encoder pre-trained on approximately one million protein structure[10] (https://github.com/DeepGraphLearning/GearNet). The structure representation module consisted of a GearNet backbone and a projection layer to unify the embedding dimension. Specifically, each enzyme structure obtained from AFDB was first preprocessed into a structural graph based on C_α atom coordinates and distance thresholds using an in-house script (https://github.com/Ming-Ni-Group/RAMER/tree/main/gearnet_process). The resulting graphs were then fed into the GearNet backbone to extract a whole-protein embedding by taking the output of the final graph-level pooling layer. The output dimension of the GearNet embedding was 3,072. To align it with the sequence representation, a single linear layer $Proj(\cdot)$ was applied to project this structural embedding to 1,024 dimensions.

During multimodal contrastive learning alignment training, the parameters of the GearNet backbone were frozen, and only the projection layer was trainable. This allowed the structural representation to guide the alignment of the sequence representation without interfering with the pretrained structural encoder.

The formulation of protein structural representation, *G*, is as follows:

$$G_0 = \mathcal{F}_{GearNet}(Str; \Theta_{GearNet}) \quad (3)$$

$$G = Proj(G_0; W_p) \quad (4)$$

Enzymatic reaction representation module with a reaction-aware network

The reaction representation module consisted of a substrate/product representation sub-module and a reaction-aware network sub-module. Substrates and products were embedded using MolT5[11], a chemical language model pre-trained on ~100 million unlabeled SMILES strings and natural language text (Colossal Clean Crawled Corpus, C4) via self-supervised learning.

First, MolT5 generated a 768-dimensional embedding for each chemical. For reactions involving multiple substrates or multiple products, the embeddings were summed separately for substrates and for products. These summed embeddings were then concatenated to obtain a preliminary representation of the catalytic reaction. The reaction-aware network concatenated this preliminary representation with a positional encoding indicating reaction direction, where 0 denoted substrates and 1 denoted products. This concatenated vector was added to the substrate-product embedding. Subsequently, the result was passed through two transformer layers and a projector linear with two linear layers, a GELU activation function, and a LayerNorm layer (**SFig. 1**), finally outputting a 1024-dimensional representation. During training, MolT5 was frozen, while gradient updates were enabled for the reaction-aware learning network.

The substrate (h_{Sub}) and the product (h_{Prod}) representations are as follows:

$$h_{Sub} = M_{Mol}(Sub; \Theta_{Mol}) \quad (5)$$

$$h_{Prod} = M_{Mol}(Prod; \Theta_{Mol}) \quad (6)$$

where M_{Mol} represents the MolT5 backbone, Sub represents reaction substrate, and $Prod$ represents reaction product.

Then, a projection layer consisting of one linear layer, denoted as *Proj*(·), was applied to project the reaction representation to 1024 dimensions:

$$Z_{Sub} = Proj(h_{Sub}; W_p) \quad (7)$$

$$Z_{Prod} = Proj(h_{Prod}; W_p) \quad (8)$$

The whole reaction representation *r* was obtained through summation,

$$r = \sum_{i=0}^{m} Z_{Sub_i} + \sum_{j=0}^{n} Z_{Prod_j} \quad (9)$$

where i and *j* were the counts of substrates and products, respectively.

The reaction-aware network, $\mathcal{G}_{react}$, took *r* with added positional encoding ($PE(pos)$) as input. The reaction representation *R* is as follows:

$$R_0 = \mathcal{G}_{react}(r + PE(pos); \Theta_{react} + \Delta\Theta) \quad (10)$$

$$\mathrm{R} = \mathrm{P}(R_0, W_p) \quad (11)$$

The projector layer $\mathrm{P}(x)$ includes two linear layers, a GELU activation function $GELU(x)$,

and a LayerNorm layer $LayerNorm(x)$, as follows:

$$\mathrm{P}(R_0, W_p) = W_2 \cdot LayerNorm(GELU(\mathrm{W}_1 + \mathrm{b}_1) + \mathrm{b}_2) \quad (12)$$

where W_2 and W_1 are the wights of linear layers, and $\mathrm{b}_1$and b_2 are the bias of linear layers.

Modality fusion via multimodal contrastive learning and loss function

RAMER adopted multimodal contrastive learning to align sequence, structural, and reaction modalities to generate multimodal representations in a joint. For RAMER training, the enzyme sequences served as the anchor, while the structures (encoded by GearNet) and reactions (encoded by MolT5-based reaction network) are aligned to it via a contrastive learning architecture based on the triplet training set. This enables the sequence representation to capture both structural and functional information.

The loss employed for training RAMER was defined as the weighted sum of two contrastive losses (Information Noise-Contrastive Estimation, InfoNCE): one between reactions and sequences, $\mathcal{L}_{\mathrm{NCE}}^{\mathrm{rec}}$, and another between structures and sequences $\mathcal{L}_{\mathrm{NCE}}^{\mathrm{str}}$. The total loss is

$$\mathcal{L} = \alpha\mathcal{L}_{\mathrm{NCE}}^{\mathrm{str}} + \beta\mathcal{L}_{\mathrm{NCE}}^{\mathrm{rec}} \quad (13)$$

where $\alpha = 0.4$ and $\beta = 0.6$ are balancing coefficients. The individual losses are defined as

$$\mathcal{L}_{\mathrm{NCE}}^{\mathrm{str}} = -log\frac{exp(sim(S^n,R^n)/\tau)}{\sum exp(sim(S^n,R^n)/\tau)} \quad (14)$$

$$\mathcal{L}_{\mathrm{NCE}}^{\mathrm{rec}} = -log\frac{exp(sim(S^n,G^n)/\tau)}{\sum exp(sim(S^n,G^n)/\tau)} \quad (15)$$

Here, $sim(S^n, R^n)$ denotes the cosine similarity between the reaction embedding and the PLM embedding of the n-th sample in a batch, and $sim(S^n, G^n)$ denotes the cosine similarity between the structure embedding (from GearNet) and the PLM embedding of the same sample. In each training epoch, the embedding of each enzyme serves as the anchor, with its corresponding structure and reaction embeddings as positive samples, while all other structures and reactions in the batch are treated as negative samples. The temperature factor τ decreased with each epoch. This design stabilizes training in early epochs and gradually encourages the model to more accurately distinguish between positive and negative samples as training proceeds.

RAMER model parameters and training

The final RAMER model contained 1,512,905,443 total parameters, of which 37,040,128 were trainable. The protein language model (PLM) module accounted for 1,211,845,632 parameters, with 3,703,808 learnable parameters. Training was performed using the AdamW optimizer with a base learning rate of 1×10^{-4}. A learnable temperature parameter was introduced with a separate learning rate of 2×10^{-5}.

To prevent model collapse during contrastive learning training, we used the EC hit rate on the external Price-149 dataset as the criterion for early stopping. For each training epoch, we extracted the representations of the current model and computed the cosine similarity between Price-149 sequences and the CLEAN100 knowledgebase to assign EC numbers. Among models trained from scratch, the one achieving the best EC hit rate on Price-149 was selected. The final best model was obtained at epoch 19 (**SFig. 1c**).

Classification of enzymes and non-enzyme proteins

To construct the enzyme atlas, we first trained a machine-learning classifier to distinguish enzymes from non-enzyme proteins using annotated proteins in Swiss-Prot. The raw dataset was downloaded from the Swiss-Prot section of the UniProt knowledgebase (release version 2025.9) in the UniProt flat-file format (UniProt.dat). For each protein entry, the DE (description), DR (database cross-reference), and CC (comments) fields were examined. Specifically, a protein was classified as an enzyme if any of these fields contained the "EC=" pattern (EC number). Sequences longer than 1024 aa were then removed to accommodate the maximum input length of the PLMs. The final dataset comprised 269,868 enzyme entries and 285,608 non-enzyme entries. Subsequently, a stratified random split was performed to generate training, validation, and test sets while preserving the original class proportion, using an 8:1:1 ratio. Specifically, the enzyme classifiers were constructed on a training set of 449,935 samples, a validation set of 49,993 samples, and a test set of 55,548 samples.

We found that the difference between enzymes and non-enzyme proteins was clearly evident based on RAMER-derived embeddings, as well as embeddings from ESM2-620M and ProtT5 (both without fine-tuning) (**SFig. 7**). The classifiers were trained using XGBoost (version 1.6.2) with an early-stopping strategy on the validation set and a learning rate of 0.05. After obtaining the model's output probabilities, a threshold-based screening (threshold = 0.36) was applied to select the optimal classification threshold.

EC number prediction by using RAMER

After extracting enzymes from the proteome via the enzyme classifier, we utilized the RAMER-derived embeddings to predict EC numbers. We compared the outcomes of zero-shot strategy approaches and contrastive learning approaches, and we found that zero-shot strategy outperformed contrastive learning (**STab. 1**). The details are depicted below.

A zero-shot prediction approach

Before annotating EC numbers for query protein, we used the background representation set ($\boldsymbol{D}$) of a knowledgebase (CLEAN70, CLEAN100, or the updated CLEAN100). $\boldsymbol{D}$ consists of the RAMER-derived embeddings $\boldsymbol{e_i}$ of proteins with their known EC annotations $\boldsymbol{EC_i}$:

$$\boldsymbol{D} = \{(\boldsymbol{e_i}, \boldsymbol{EC_i})\} \quad (15)$$

$$\boldsymbol{EC}(\boldsymbol{e_i}) = \boldsymbol{EC_i} \quad (16)$$

For a given query protein sequence, we first obtained its RAMER embedding $\boldsymbol{q}$ and then computed cosine similarity between $\boldsymbol{q}$ and each $\boldsymbol{e_i}$ within $\boldsymbol{D}$. The EC number(s) of $\boldsymbol{q}$ were directly assigned based on the EC number of the most similar enzyme in $\boldsymbol{D}$ (the 'top-1 mode' for EC prediction). Simultaneously, a set of up to five enzymes was determined via a maximum separation algorithm (the 'max-sep mode'). It should be noted that enzymes in $\boldsymbol{D}$ can possess multiple EC numbers; therefore, the top-1 mode can output more than one EC number for a single query.

Top-1 mode. This mode outputs the EC number(s) of the single most similar enzyme in D as the final prediction:

$$EC_{pred} = \mathrm{EC}\left(\underset{e_i \in D, EC(e_i)=EC_i}{\operatorname{argmax}} cosine(q, e_i)\right), 1 \le i \le N \quad (17)$$

where N is the total number of enzymes in $\boldsymbol{D}$.

Max-sep mode. The maximum separation algorithm, originally adopted in CLEAN's EC number prediction, identifies a set of high-confidence candidate EC numbers rather than returning only the top-1 result, thereby providing more candidates for subsequent functional analysis. In RAMER's max-sep mode, we first sorted the background set D by cosine similarity in descending order, yielding:

$$S(1) \geq S(2) \geq \cdots \geq S(N) \qquad (18)$$

We then selected the proteins associated with the top 10 unique EC labels as the candidate set $\boldsymbol{C}$. The size of $\boldsymbol{C}$ is variable and is denoted as $\boldsymbol{M}$. Next, for the sorted candidates in $\boldsymbol{C}$, we define the set of adjacent similarity gaps $\boldsymbol{R}$ as:

$$R = \{\ r_i = \mid \cos(q, e_{i+1}) - \cos(q, e_i) \mid\ \}, 1 \leq i < M \qquad (19)$$

Each element $r_i \in R$ quantifies the absolute drop in cosine similarity between consecutive entries in the sorted list of C. The breakpoint is selected as the index with the maximum gap, i.e.:

$$k = \arg\max(r_i) \qquad (20)$$

The value $\boldsymbol{k}$ is dynamic. Finally, the EC numbers of the top k enzymes in the original sorted list $\boldsymbol{D}$ (or equivalently, those in $\boldsymbol{C}$ before the breakpoint) are selected as predictions, with the maximum number of predictions capped at 5:

$$EC_{\text{pred}} = \{EC(e_0), EC(e_1), \ldots, EC(e_k)\}, k \leq 5 \qquad (21)$$

A contrastive learning prediction approach

We adopted the contrastive learning losses as used in CLEAN[4]: Triplet Margin Loss[4] and Supcon-Hard Loss (a modified version of the supervised contrastive loss[12]. The model architecture followed CLEAN, consisting of two blocks (each with a linear layer, BatchNorm layer, and dropout) followed by an output linear layer, except that we replaced the input enzyme embeddings (originally derived from ESM-1b) with RAMER-derived embeddings. The definitions of the two losses are as follows:

Triplet Margin Loss (margin $\alpha = 1$) encourages intra-class compactness and inter-class separation by using an anchor, a positive from the same EC class, and a negative from a different EC class:

$$\mathcal{L}_{\text{TM}} = \| z_a - z_p \|_2 - \| z_a - z_n \|_2 + \alpha \qquad (22)$$

Supcon-Hard Loss (temperature $\tau = 0.1$, embeddings L2-normalized) is a modified supervised contrastive objective that uses multiple positives (same EC class) and hard mined negatives (other classes) per batch:

$$\mathcal{L}_{\text{Sup}} = \sum_{e \in E} \frac{-1}{|P(e)|} \sum_{z_p \in P(e)} \log \frac{\exp(z_e \cdot z_p / \tau)}{\sum_{z_a \in P(e) \cup N(e)} \exp(z_e \cdot z_a / \tau)} \qquad (23)$$

In each epoch, for every EC class $e \in E$, we trained the model using a single anchor z_e. The set $N(e)$ consisted of hard negative examples mined relative to the cluster center of class e and the anchor z_e. The set $P(e)$ contained one or more positive samples from the same EC class e.

Evaluation metrics for EC number prediction and unsupervised clustering

Metrics used to evaluate the performance of EC number prediction include precision, recall, F1-score (F1), and overall accuracy (ACC). These metrics are defined as follows:

$$\text{Precision} = \sum_{i=1}^{C} \frac{TP_i + FN_i}{N} \cdot \frac{TP_i}{TP_i + FP_i} \quad (23)$$

$$\text{Recall} = \sum_{i=1}^{C} \frac{TP_i + FN_i}{N} \cdot \frac{TP_i}{TP_i + FN_i} \quad (24)$$

$$F1 = \sum_{i=1}^{C} \frac{TP_i + FN_i}{N} \cdot \frac{2TP_i}{2TP_i + FP_i + FN_i} \quad (25)$$

$$\text{ACC} = \frac{\sum_{i=1}^{C} TP_i}{N} \quad (26)$$

where TP_i, FP_i, and FN_i denote true positives, false positives, and false negatives for EC class i in the multi-label setting, respectively. C is the number of EC classes, and N is the total number of samples. The EC classes considered can be at any level from level-1 to level-4.

In addition, we used EC hit rate for epoch selection during RAMER training. It is defined as the ratio of correctly annotated enzymes to the total number of instances:

$$\text{EC hit rate} = \frac{\sum_{i=1}^{C} TP_i}{N} \quad (27)$$

where C and N are as defined above.

RAMER and other pre-trained models can be used to cluster enzymes with their respective embeddings in an unsupervised manner. To assess the accuracy, robustness, and separability of these unsupervised clustering results, we employed both external and internal clustering validation metrics. For external evaluation, where true class labels were known and used as a reference, we used three widely adopted metrics: the Adjusted Rand Index (ARI)[13], Normalized Mutual Information (NMI), and Purity. For internal validation, which did not require prior label information, we used the Silhouette coefficient[14].

Microbial protein sequence source and preprocessing for the enzyme atlas

The atlas of enzymes was extracted from over three billion protein sequences downloaded from databases including MGnify[15], Global Catalogue of Metagenomics (gcMeta)[16], Global Ocean Microbiome Genomic Catalogue (GOMC)/Global Ocean Microbiome Protein Catalog (GOPC)[17], and a recently reported microbiome dataset from the Mariana Trench Sediment[18].

The protein sequences from MGnify were downloaded via https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2024_04/. Protein sequences from GOMC/GOPC were downloaded via https://ftp.cngb.org/pub/SciRAID/microbiomics/MDB0000002/GOPC.geneset.pep.fa.gz. The metagenome-assembled genomes (MAGs) of 24 environments in gcMeta (as listed in **STab. 3**) were obtained, and protein-coding genes were annotated using Prodigal (v2.6.3)[19] with the parameter "-p meta". For the MAGs derived from the Mariana Trench Sediment, genes were annotated using InterProScan (v5.59-91.0)[20] with the parameter "--f TSV,GFF -goterms -pathways".

We removed any sequences longer than 1024 amino acids and predicted EC numbers using the RAMER top-1 mode to construct the billion-scale enzyme atlas.

Besides, we also analyzed protein sequence datasets of bacterial origin from NMPFamsDB (https://pavlopoulos-lab.org/NMPFamsDB/)[21,22]. NMPFamsDB hosts novel metagenomic protein clusters with no or weak hits to Pfam[23] or reference genomes. We first extracted enzymes from these datasets using our enzyme classifier, and then predicted their EC numbers using RAMER top-1 mode. These enzymes were not included in the enzyme atlas.

Analyses and visualization of protein sequences and structures

Visualization of embeddings via dimensionality reduction

UMAP[24] (Uniform Manifold Approximation and Projection) was used for visualization analysis of embeddings of proteins and reactions with default parameters, except that we searched for optimal "n_neighbour" and "min_dist" values. The large datasets were used curl.manifold[25] (v24.10.00) library (https://umap-learn.readthedocs.io/en/latest/) on GPU. TMAP (Tree-based Mapper)[26], a method for visualizing high-dimensional data based on minimum spanning tree, was also applied. We used the tmap package (v1.0.6, https:// tamp.readthedocs.io/en/latest/) for this purpose.

CLEAN is not a pretrained model for embeddings, and we employed the last layer embeddings of CLEAN for UMAP analysis.

Protein structure prediction and analysis

Protein structures were predicted based on protein sequences by using AlphaFold3[27] with both a locally deployed version and the webserver (https://alphafoldserver.com/). Protein structures were analyzed and visualized using PyMOL v2.5.0[28].

Phylogenetic analysis of protein sequences

Multiple sequence alignment was performed using MAFFT[29] (version 7.525) with the auto strategy, which automatically selects the most appropriate alignment algorithm based on dataset size and characteristics. Phylogenetic trees were constructed by using FastTree (version 2.1.11)[30] with the LG amino acid substitution model (parameter: -lg). The CAT approximation for site-specific rate heterogeneity is enabled by default. The webserver iTOL v7[31] (https://itol.embl.de/) was used for tree visualization. Phylogenetic analyses were also performed by using Mega (v12)[32] with the MUSCLE multiple-alignment[33] and the maximum likelihood method.

Multiple sequence alignment of mined PETases and fluorinases was also visualized with ESPript 3.0[34], with conserved residues and secondary structure elements annotated.

Multiple structural alignment (MSTA)

The structural distance matrix was constructed based on the TM-score used TM-align (v20220412)[35] using the following transformation:

$$d(i,j) = -ln(TM - score(i,j)) \quad (28)$$

A structural phylogenetic tree was constructed using the neighbor-joining (NJ) method and using the DistanceTreeConstructor class from the BioPython package (v1.86). The resulting

trees were visualized using iTOL v7(https://itol.embl.de/).

Classifier of three types of adenylate kinases (ADKs)

To analyze the enzymes assigned to EC 2.7.4.3, we constructed a three-class classification model to distinguish three types of ADKs by their lid domains: lidless, H-bond lid, and Zn^{2+}-binding lid. The 175 experimentally validated ADKs (listed in SFile. 5) were used for training and testing. Based on the result shown in **Fig. 3i**, we reclassified "others" into "H-bond". From each of the three classes, 15 samples were sampled as the test set, and the remaining 140 ADKs were used as the training set. The ADK lidtype classifier was trained using XGBoost (version 1.6.2) with RAMER embeddings as input, and a grid search was used to select the best model, whose performance is shown in **SFig. 15a-d**.

Experimental validations of enzymes

Fluorinases

Enzyme expression and purification

All genes were codon-optimized and cloned into the pET-28a (+) vector between *NdeI* and *XhoI* sites. Fluorinase variants were synthesized by BGI Genomics (Wuxi, China). Recombinant plasmids were transformed into *E. coli* BL21 (DE3).

Cells were cultured in LB medium containing kanamycin (100 μg/mL) at 37 °C and 200 rpm until OD600 reached ~0.8. Protein expression was induced with 0.4 mM IPTG, followed by incubation at 16 °C and 150 rpm for 20 h.

Cells were harvested by centrifugation (8000×g, 10 min) and resuspended in lysis buffer (200 mM sodium phosphate, 10% glycerol, pH 8.0). Lysozyme (80 mg), PMSF (1 mM), DNase I (1 mg), and imidazole (30 mM) were added prior to sonication (4 °C, 25 min). After centrifugation (12,000×g, 30 min, 4 °C), the supernatant was collected.

The lysate was incubated with 3 mL Ni-NTA resin pre-equilibrated with lysis buffer for 30 min at 4 °C, loaded onto a gravity column, and washed with buffer containing 20 mM imidazole. Target proteins were eluted with 200 mM imidazole, concentrated using an Amicon centrifugal filter, desalted, and stored at −80 °C. Protein concentration was determined by absorbance at 280 nm.

Crude enzyme preparation and activity assays

Single colonies were cultured in LB medium with kanamycin at 37 °C to OD600 = 0.6–0.8. Seed cultures were diluted 1:100 into fresh LB medium. For 24-well deep-well expression, 3 mL cultures were grown at 37 °C (800 rpm) for 2.5 h, induced with 0.5 mM IPTG, and expressed at 16 °C (600 rpm) for 20 h. Cells were harvested and resuspended in 200 μL BugBuster Master Mix (Novagen), incubated at room temperature for 10 min, and centrifuged. Supernatants were used as crude enzyme solutions.

Activity assays (200 μL) contained 100 μM SAM, 75 mM KF, 100 μL crude enzyme, and 50 mM HEPES buffer (pH 7.8). Reactions were performed at 37 °C for 1 h, quenched at 95 °C for 5 min, and centrifuged (12,000×g, 30 min, 4 °C). All assays were conducted in triplicate.

Purified enzyme activity assays

Purified enzyme assays were conducted under identical conditions, with enzyme concentration adjusted to 100 μg/mL. All reactions were performed in triplicate.

HPLC analysis

Reaction mixtures were analyzed using HPLC (Shimadzu) with UV detection at 260 nm and a C18 column (4.6 × 250 mm, 5 μm). The mobile phase consisted of solvent A (0.05% acetic acid in water) and solvent B (acetonitrile) at a flow rate of 1 mL/min. The gradient program was as follows: 5% B for 5 min, increased to 30% B over 5 min, held at 30%, and then returned to 5% B over 10 min. Product identity (5′-fluorodeoxyadenosine, 5′-FDA) was confirmed by comparison with authentic standards.

PETases

Enzyme expression and purification

PETase genes were commercially synthesized by BGI Research after codon optimization for expression in *Escherichia coli*. To facilitate recombinant expression, the sequences encoding the N-terminal signal peptides were removed during gene design. The resulting coding sequences were cloned into the pET22b vector via the NdeI and XhoI restriction sites and transformed into chemically competent *E. coli* BL21 (DE3) cells (TransGen Biotech) for heterologous expression. Single colonies were inoculated into 96-well plates containing 800 μL LB medium supplemented with 100 μg/mL ampicillin and incubated at 37 °C and 800 rpm for 7 h. Subsequently, 200 μL of each culture was transferred into a new 96-well plate containing 800 μL GMML medium supplemented with 100 μg/mL ampicillin and 0.1 mM IPTG, followed by incubation for an additional 18 h. The cells were then harvested by centrifugation, and the pellets were resuspended in 80 μL Tris–HCl (pH 9.0) supplemented with 8 μL TieChui™ *E. coli* Lysis Buffer. After incubation at 4 °C for 10 min, the lysates were centrifuged, and the supernatants were collected for subsequent activity measurement.

Crude enzyme preparation and hydrolytic activity assays

Hydrolytic activity toward p-nitrophenyl acetate (*p*NPA; Aladdin, China) was assayed in a total reaction volume of 100 μL containing 2 μL crude enzyme, 88 μL of 0.1 M Tris–HCl buffer (pH 9.0), and 10 μL pNPA solution (2 mM), corresponding to a final substrate concentration of 0.2 mM. The reaction was carried out at 37 °C for 3 min and then immediately quenched by the addition of 100 μL absolute ethanol. The release of p-nitrophenol was monitored by measuring absorbance at 405 nm using a Spark microplate reader (Tecan, Männedorf, Switzerland). To enable quantitative determination of *p*NPA hydrolysis activity, a standard curve of *p*-nitrophenol was established under the same detection conditions. Briefly, p-nitrophenol was diluted to final concentrations of 0.01, 0.02, 0.05, 0.1, and 0.2 mM, each concentration was analyzed in quadruplicate.

Purified enzyme activity assays

Protein purification was conducted as previously described[36]. Amorphous GfPET film (ES301445, Goodfellow) was punched into circular discs with a diameter of 7 mm and an average weight of 11.3 mg. The GfPET film was incubated with 0.04 mg/mL of purified enzyme in 500 μl of Tris-HCl buffer (pH 9.0). The reaction mixture was incubated at 37 °C or 50 °C for 24 h, and immediately quenched by the addition of 500 μL absolute acetonitrile. Then the hydrolysis products were analyzed by high-performance liquid chromatography (HPLC) using a Shimadzu (Kyoto, Japan) DGU-20A3R system equipped with an autosampler and a UV

detector. Separation was achieved on a GL Sciences Inertsil ODS-3 column (5 μm, 4.6 × 150 mm) maintained at 40 °C, with UV detection at 260 nm. The mobile phase consisted of 0.1% formic acid in water (solvent A) and acetonitrile (solvent B), delivered at a flow rate of 1 mL/min. A linear gradient was applied, with solvent B decreasing from 50% to 20% over 10 min. Calibration curves for MHET and TPA were generated using standard solutions at concentrations of 0.2–2 mM by plotting peak area against concentration. All reaction mixtures were filtered through 0.22 μm membrane filters prior to injection.

Reproducibility

All enzymatic assays were performed with at least three independent biological replicates, each measured in technical triplicate where applicable. Data are reported as mean ± standard deviation (SD).

Data availability

The billion-scale enzyme atlas, including protein identifiers, predicted EC numbers, and environmental origins, is available on Zenodo via https://zenodo.org/doi/10.5281/zenodo.20021021. The microbial enzymes from the 25 environments, from MGnify, and from GOPC are separately uploaded. The 778 predicted PETases with their RAMER-derived cluster identifiers, and the 5313 mined enzyme with similar RAMER-derived embedding with the PURase Aes72, are provided in the supplementary files.

Code availability

The code used for training and inference of RAMER is accessible on GitHub at https://github.com/Ming-Ni-Group/. The machine-learning models for the two-class classifier (enzymes vs. non-enzymes), the three-class classifier for the three types of ADKs, and other custom scripts are also available in the repository. The RAMER model weights, background sets for EC prediction, partially intermediate processed data, as well as training and evaluation datasets, are publicly accessible at https://huggingface.co/PengJiaMa123/RAMER.

Acknowledgement

This work was supported by National Key Research and Development Program of China (No.2024YFA1307700). We thank Teng-xiang Yu for assistance in data uploading and validation.

References

1. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).

2. Fleming, J. *et al.* AlphaFold Protein Structure Database and 3D-Beacons: New Data and Capabilities. *J. Mol. Biol.* **437**, 168967 (2025).

3. Bansal, P. *et al.* Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res.* **50**, D693–D700 (2022).

4. Yu, T. *et al.* Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).

5. Sanderson, T., Bileschi, M. L., Belanger, D. & Colwell, L. J. ProteInfer, deep neural networks for protein functional inference. *eLife* **12**, e80942 (2023).

6. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018).

7. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

8. Mikhael, P. G., Chinn, I. & Barzilay, R. CLIPZyme: Reaction-Conditioned Virtual Screening of Enzymes. Preprint at https://doi.org/10.48550/arXiv.2402.06748 (2024).

9. Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models. *Iclr* **1**, 3 (2022).

10. Zhang, Z. *et al.* Protein Representation Learning by Geometric Structure Pretraining. (2023).

11. Edwards, C. *et al.* Translation between Molecules and Natural Language. in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (eds Goldberg, Y., Kozareva, Z. & Zhang, Y.) 375–413 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022). doi:10.18653/v1/2022.emnlp-main.26.

12. Khosla, P. *et al.* Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **33**, 18661–18673 (2020).

13. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).

14. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

15. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).

16. Sun, Y. *et al.* gcMeta 2025: a global repository of metagenome-assembled genomes enabling cross-ecosystem microbial discovery and function research. *Nucleic Acids Res.* **54**, D724–D733 (2025).

17. Chen, J. *et al.* Global marine microbial diversity and its potential in bioprospecting. *Nature* **633**, 371–379 (2024).

18. Xiao, X. *et al.* Microbial ecosystems and ecological driving forces in the deepest ocean sediments. *Cell* **188**, 1363-1377.e9 (2025).

19. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

20. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

21. Pavlopoulos, G. A. *et al.* Unraveling the functional dark matter through global metagenomics. *Nature* **622**, 594–602 (2023).

22. Baltoumas, F. A. *et al.* NMPFamsDB: a database of novel protein families from microbial metagenomes and metatranscriptomes. *Nucleic Acids Res.* **52**, D502–D512 (2024).

23. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

24. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv Prepr. ArXiv180203426* (2018).

25. Nolet, C. J. *et al.* Bringing UMAP closer to the speed of light with GPU acceleration. in vol. 35 418–426 (2021).

26. Probst, D. & Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminformatics* **12**, 12 (2020).

27. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

28. DeLano, W. L. Pymol: An open-source molecular graphics tool. *CCP4 Newsl Protein Crystallogr* **40**, 82–92 (2002).

29. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

30. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS One* **5**, e9490 (2010).

31. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).

32. Kumar, S. *et al.* MEGA12: molecular evolutionary genetic analysis version 12 for adaptive and green computing. *Mol. Biol. Evol.* **41**, msae263 (2024).

33. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

34. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).

35. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

36. Li, X. *et al.* Integrating rational design and machine learning to engineer a novel plastic depolymerase to break down PBAT-based mulch film for sustainable recycling. *J. Hazard. Mater.* 140581 (2025).

Supplementary Materials

Supplementary Tables

STab. 1. A zero-shot approach was more favorable than a contrastive learning approach.

STab. 2. Prediction performance at EC Level 1–4 across New392 and Ram255 validation datasets, corresponding to Fig. 1f-g.

STab. 3. Summary of datasets of microbial protein sequences, corresponding to Fig. 2a-b.

STab. 4. Enzyme activity prediction performance across representation models.

STab. 5. EC number prediction of PETase candidates from Mariana Trench microbiomes.

Supplementary Figures

SFig. 1. Detailed architecture of the RAMER framework and training dataset construction.

SFig. 2. The comparison and selection of protein language models (PLMs) for protein sequence embedding employed in RAMER.

SFig. 3. Visual clustering of enzymatic and non-enzymatic proteins in Swiss-Prot with different protein representation models.

SFig. 4. Clustering performance evaluation of RAMER, ESM, ProtT5, and the embedding modules of CLEAN.

SFig. 5. Embeddings of catalytic reactions with RAMER and MolT5 and clustering performance evaluation.

SFig. 6. Prediction performance stratified by training set enzyme abundance and sequence homology.

SFig. 7. A enzyme *versus* non-enzyme binary classifier.

SFig. 8. Distribution of protein and enzyme sizes across datasets.

SFig. 9. EC class distribution of bacterial unclassified ‘dark matter’ proteins from NMPFamsDB.

SFig. 10. Over- and under-represented EC enzyme classes associated with distinct environmental microbiomes.

SFig. 11. Sequence and structural similarities of the known and predicted PETases, corresponding to Fig. 3a right panel.

SFig. 12. Mining of candidate PUases, corresponding to Fig. 3h.

SFig. 13. Structural and sequence analysis of previously misannotated adenylate kinase (ADK) homologs.

SFig. 14. RAMER-derived prediction of catalytic activity and growth temperature of adenylate kinase (ADK) enzymes.

SFig. 15. Adenylate kinase (ADK) subtype classifier and distribution across environmental microbiomes.

SFig. 16. Alignment of mined fluorinases and known fluorinases.

SFig. 17. RAMER-based clustering, phylogeny and structural analysis of experimentally validated PETase candidates, corresponding to Fig. 4e-f.

SFig. 18. Alignment of mined PETases in Marina Trench sediment microbiome and *Is*PETase.

SFig. 19. ESM- and GearNet-based clustering of PETases mined in Marina Trench sediment microbiome.

Supplementary Files

SFile 1. Ram-255 validation set.

SFile 2. The 118 enzymes released after 2023 without level-4 EC numbers.

SFile 3. Abundances and ratios of billion-scale enzymes across level-1 to level-4 EC classes.

SFile 4. Abundances, ratios, and Z-score of enzymes from gcMeta and Mariana Trench sediments across 25 environments.

SFile 5. Enzymes predicted as EC 3.1.1.101.

SFile 6. Reported nature PETases and adenylate kinases.

SFile 7. Mined PURases with high cosine similarity to the PURases Aes72.

SFile 8. Candidate PETases.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTablesandFigures.docx