Computational discovery and design of potential microbial enzymes for plastics degradation

Celina Zhou

Ladue Horton Watkins High School, St. Louis, MO

May 2025

1 Introduction

Plastic pollution has become a pressing global concern as plastic production continues to rise. Conventional end-of-life management methods—landfilling, incineration, and limited recycling—are insufficient to address the mounting waste. Microbial enzymes offer a promising, bio-based alternative for plastic degradation and recycling, supporting the development of a circular economy (**?**). These enzymes facilitate natural and environmentally friendly breakdown of plastics into harmless compounds.

Microbial enzymes play a key role in plastic degradation by catalyzing the hydrolysis or oxidation of polymer bonds, converting plastics into smaller, less harmful molecules. Several families of enzymes, particularly hydrolases and oxidoreductases, have been identified for their ability to degrade different types of plastics.

Hydrolases are among the most extensively studied plastic-degrading enzymes. PETase and MHETase, discovered in Ideonella sakaiensis 201-F6, are highly effective in degrading polyethylene terephthalate (PET), a common synthetic plastic. PETase cleaves ester bonds in PET, producing terephthalic acid (TPA) and ethylene glycol (EG), which are further metabolized into carbon dioxide and water. Cutinases, found in fungi and bacteria such as Fusarium solani pisi and Thermobifida fusca, belong to the α/β-hydrolase superfamily and can degrade PET, poly(butylene succinate) (PBS), and other polyesters. Microbial pre-treatment can enhance PET degradation by reducing surface hydrophobicity and improving enzyme accessibility. Lipases—such as Lipase B (CALB) from Candida antarctica—and esterases are also capable of breaking down high-molecular-weight polymers and synthetic copolyesters like poly(butylene adipate-co-terephthalate) (PBAT).

Oxidoreductases, including laccases and peroxidases, catalyze oxidation reactions that assist in plastic degradation. Laccases, copper-dependent enzymes found in both fungi and bacteria, have demonstrated activity in degrading polyamide (PA), polyethylene (PE), and polypropylene (PP). Bacterial laccases are typically more stable across wide pH and temperature ranges. Peroxidases—commonly found in lignin-degrading fungi—use hydrogen peroxide to oxidize various substrates. The addition of manganese peroxidase has been shown to enhance PE degradation in certain fungal systems.

The efficiency of microbial plastic degradation depends on multiple environmental and material factors. Conditions such as temperature, pH, light, and humidity significantly influence microbial activity and enzyme function. For instance, UV photolysis can serve as an effective pre-treatment to promote polymer hydrolysis, while optimal moisture levels support microbial growth. Plastic properties, including molecular weight, surface area, and chemical structure, also affect biodegradability. Plastics with heteroatomic backbones (e.g., PET, PUR) are generally more degradable than those with carbon-carbon backbones (e.g., PE, PP, PS, PVC). Additionally, enzyme stability and pH sensitivity can limit degradation efficiency, though advances in protein engineering and site-directed mutagenesis have produced enzyme variants with enhanced activity and thermal stability. microbial enzymes represent a sustainable and scalable strategy for mitigating plastic pollution. Researchers are leveraging computational and biotechnological tools tp enhance enzyme performance, address current limitations and expand their applicability. In my work, I applied computational approaches to identify novel plastic-degrading enzymes from environmental metagenomic datasets in the NCBI database. Furthermore, I used AI-based models to design artificial enzymes with potential plastic-degrading activity.

2 Results and discussions

2.1 assembly and annotation of metagenomics data

Metagenomics is an advanced approach that relies on the direct sequencing of DNA from entire microbial communities, bypassing the need to culture individual organisms—an often challenging or impossible task for many microbial species. By analyzing total environmental DNA, metagenomics provides comprehensive insights into microbial community composition (which organisms are present), functional potential (what metabolic processes they perform), and evolutionary relationships. This approach has been instrumental in elucidating microbial contributions to pollutant degradation in contaminated environments such as soil and sediment. Furthermore, metagenomic studies have led to the discovery of numerous novel enzymes, metabolic pathways, and previously uncharacterized microorganisms. In this study, Illumina short-read sequencing data from 89 environmental metagenomic samples were downloaded from the NCBI database, including 23 coastal water, 26 marine water, 18 river water, and 22 soil samples. Each dataset was assembled using MEGAHIT with default parameters, as described in Section 3.2. The primary objective of this work was to evaluate the feasibility of metagenomic approaches for discovering novel plastic-degrading enzymes and identifying previously unknown proteins with potential plastic-degrading activity; therefore, detailed assembly quality assessment was not performed. In total, more than 420,000 contigs were assembled across the 89 samples. Gene prediction and functional annotation of the assembled contigs were conducted using Prodigal (Section 3.3), a widely used microbial genome annotation tool. Although fungal DNA may have been present in some samples, eukaryotic annotation was not performed due to computational resource limitations. Overall, the analysis yielded over 1.7 million predicted protein sequences longer than 50 amino acids from the combined metagenomic assemblies. After filtering bacteria housekeeping and proteins that definitely have no environmental remediation functions, about 700 thousands proteins sequence left.

2.2 Identification of co-occuring proteins and discovery of putative plastics degrading enzymes

Co-occurrence network analysis of the annotated metagenomic datasets identified 127 protein-coding genes that appeared in more than five environmental samples, though not necessarily the same five. Among these, 19 proteins exhibited strong sequence or domain similarity to enzymes known to participate in plastic degradation pathways. These include putative esterases, polyesterases, alkane hydroxylases, laccases, and monooxygenases, all of which are enzyme classes previously implicated in the breakdown of synthetic polymers such as polyethylene terephthalate (PET) and polyhydroxyalkanoates (PHA). An illustrative example is shown in Figure 1, where a candidate protein shares close sequence similarity with MCC4266688.1, a member of the PHB depolymerase family of esterases, suggesting potential involvement in biopolymer hydrolysis. The consistent co-occurrence of these genes across multiple distinct habitats suggests their potential ecological relevance in polymer degradation processes.

In addition to these characterized candidates, 30 proteins with unknown or unannotated functions were identified within the co-occurrence network. Determining the functions of these uncharacterized proteins represents a critical next step toward expanding the known repertoire of plastic-degrading enzymes. Structural and sequence-based analyses, coupled with experimental validation, will be necessary to elucidate their biochemical roles.

2.3 AI-generated putative PETase enzymes

A Deep Protein Language Model (DPLM) was applied to generate novel protein sequences with potential polyethylene terephthalate (PET)–degrading activity. Using experimentally validated PETase sequences as reference inputs, the model produced a set of putative PETase-like enzyme candidates. From this dataset, the top 10 protein sequences with lengths comparable to that of the PETase from *Ideonella sakaiensis were selected for further structural analysis (Table 1). These candidates exhibited diverse amino acid compositions and sequence identities, indicating the model's ability to explore novel regions of protein sequence space while maintaining overall structural feasibility.*

putative$_proteins$	length
protein1	291
protein2	295
protein3	295
protein4	298
protein5	302
protein6	305
protein7	309
protein8	312
protein9	312
protein10	315

Table 1: List of the 18 virus receptors and 22 viruses

The three-dimensional (3D) structures of the 10 AI-generated proteins were predicted using AlphaFold2. Structural comparison with the experimentally determined PETase from I. sakaiensis revealed that four of the predicted proteins displayed high overall structural similarity to PETase enzymes 2. These results suggest that the four candidates may represent putative novel PETase variants, warranting further computational and experimental investigation to confirm their substrate-binding properties and catalytic potential in plastic degradation.

3 Methods

3.1 Data Collection

Publicly available environmental metagenomic sequencing datasets were retrieved from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). Samples were selected from diverse ecological habitats, including soil, seashore, lakeshore, riverside, and marine environments, to capture a broad range of microbial communities potentially exposed to plastic contamination. Metadata associated with each dataset (sampling site, sequencing platform, and environmental context) were collected to ensure data quality and ecological diversity.

3.2 Metagenomic Assembly

Raw sequencing reads were quality-checked using FastQC and trimmed with Trimmomatic to remove low-quality bases and adapter sequences. The filtered reads were subsequently assembled into contigs using MEGAHIT (v1.2.9) and SPAdes (v3.15.5) under default metagenomic assembly parameters. Contigs shorter than 500bp were excluded from further analysis. The resulting contigs were stored in FASTA format for downstream annotation.

3.3 Gene Prediction and Annotation

Open reading frames (ORFs) were predicted from assembled contigs using Prodigal (v2.6.3) Functional annotation of predicted protein-coding genes was performed using Prokka (v1.14.6), incorporating BLASTp searches against the NCBI non-redundant (NR) protein database and Pfam domain assignments. The resulting annotations were filtered to retain only protein-coding sequences with confident functional predictions.

3.4 Gene Filtering

To remove common bacterial housekeeping genes and focus on potentially novel or environment-specific functions, annotated genes showing more than 70% sequence identity and more than 80% coverage with Escherichia coli and Bacillus subtilis and Mycobacterium tuberculosis proteins (based on BLASTp comparison against the E. coli K-12 MG1655, B. subtilis PS832 and M. tuberculosis H37Rv Siena reference genomes) were excluded. The remaining non-conserved genes were retained for subsequent co-occurrence and functional analyses.

3.5 Co-Occurrence Network Construction

A gene co-occurrence network was constructed to examine functional associations among the remaining protein-coding genes. Gene presence–absence matrices were generated across all metagenomic samples. Pairwise co-occurrence scores were calculated using the Jaccard similarity coefficient, and statistically significant associations (p ¡ 0.01, permutation test) were retained.

3.6 Identification of Potential Plastic-Degrading Genes

Functional inference of candidate plastic degrading genes was conducted by comparing all retained proteins against a curated database of experimentally validated plastic-degrading enzymes, including PETase, MHETase, cutinases, laccases, and peroxidases. Sequence similarity searches were performed using BLASTp with an E-value threshold of $1\mathrm{e}-10$ *and identity more than 30%. Candidate* *sequences were further analyzed using HMMER (v3.4) against PFAM and TIGRFAM profiles corresponding to hydrolase and oxidoreductase enzyme families. Genes showing structural homology or domain conservation with known plastic-degrading enzymes were designated as putative plastic-degradation candidates.*

3.7 Applying AI to design novel putative plastic-degraging enzymes

Protein sequences of experimentally validated plastic-degrading enzymes, including PETase, MHETase, cutinases, and other related hydrolases and oxidoreductases, were retrieved from the National Center for Biotechnology Information (NCBI) protein database. Redundant sequences were removed using CD-HIT with a 90% sequence identity threshold to obtain a nonredundant reference dataset. These curated sequences were incorporated as additional training data into two general protein generative AI frameworks—ProteinMPNN and DPLM (Deep Protein Language Model)—to fine-tune their capacity for generating proteins with plastic-degrading potential. The fine-tuned models were then queried to generate a large pool of novel protein sequences predicted to possess catalytic functions similar to known plastic-degrading enzymes.

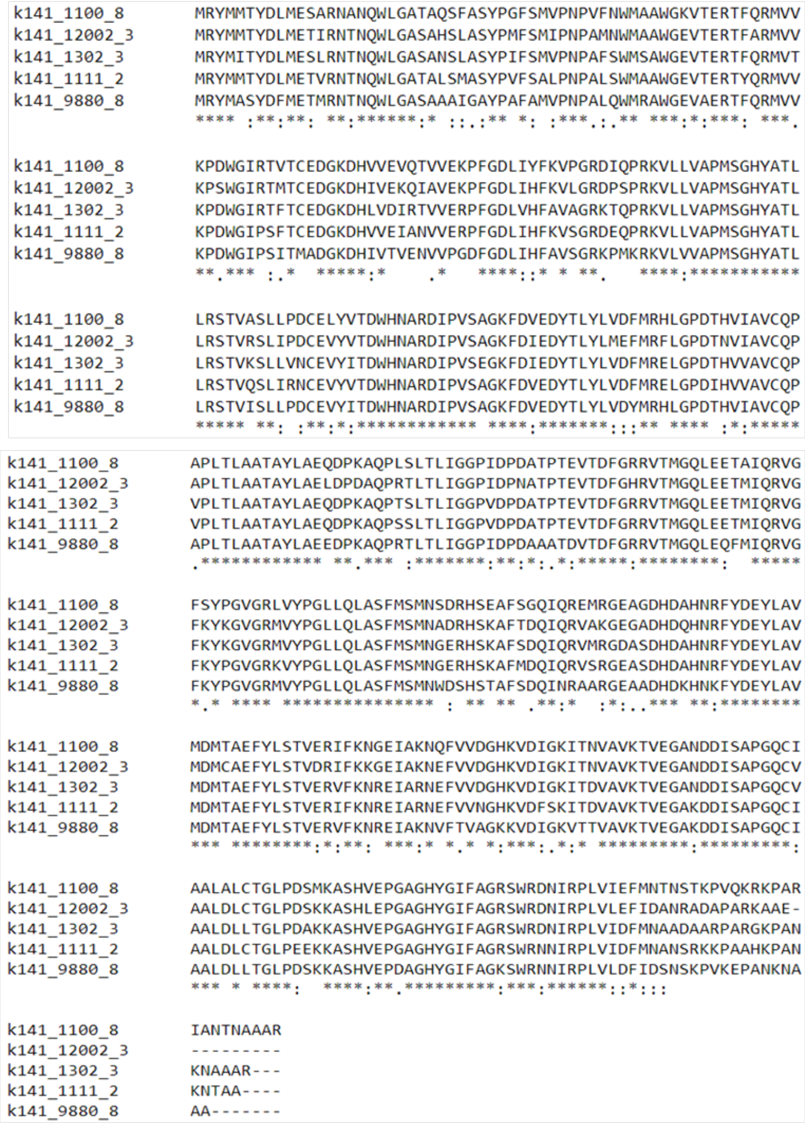
The newly generated sequences were subsequently analyzed using AlphaFold2 to predict their three-dimensional (3D) structures. The predicted protein models were aligned and compared with experimentally validated plastic-degrading enzyme structures obtained from the Protein Data Bank (PDB) using TM-align and RMSD calculations to assess structural similarity and potential functional relevance. Proteins exhibiting high structural correspondence with known PETase or related hydrolases were further examined for conserved catalytic residues and substrate-binding motifs. Based on these comparative analyses, candidate novel plastic-degrading enzymes were identified for future experimental validation.

4 Conclusion and Future Directions

Co-occurrence analysis of environmental metagenomic sequencing datasets provides a powerful strategy for uncovering novel plastic-degrading enzymes. While the majority of these datasets in the NCBI database were generated for unrelated research purposes, systematic re-analysis allows the identification of previously uncharacterized microbial genes and communities with potential plastic-degrading capabilities. Complementing this approach, artificial intelligence–based protein design enables the generation of novel enzymes with predicted functional properties. Well-trained deep protein language models (DPLMs) can produce protein sequences that adopt three-dimensional structures similar to native PETase, as confirmed by AlphaFold2, suggesting potential catalytic activity against synthetic polymers.

Looking forward, the integration of metagenomic co-occurrence analysis with AI-driven protein design represents a promising framework for expanding the repertoire of plastic-degrading enzymes. Future work will involve experimental validation of candidate enzymes, optimization of their catalytic efficiency, and exploration of synergistic enzyme consortia for enhanced plastic degradation. This combined computational and experimental strategy holds substantial potential to contribute to sustainable biotechnological solutions for mitigating global plastic waste.

Section References



```
k141_1100_8       MRYMMTYDLMESARNANQWLGATAQSFASYPGFSMVPNPVFNWMAAWGKVTERTFQRMVV
k141_12002_3      MRYMMTYDLMETIRNTNQWLGASAHSLASYPMFSMIPNPAMNWMAAWGEVTERTFARMVV
k141_1302_3       MRYMITYDLMESLRNTNQWLGASANSLASYPIFSMVPNPAFSWMSAWGEVTERTFQRMVT
k141_1111_2       MRYMMTYDLMETVRNTNQWLGATALSMASYPVFSALPNPALSWMAAWGEVTERTYQRMVV
k141_9880_8       MRYMASYDFMETMRNTNQWLGASAAAIGAYPAFAMVPNPALQWMRAWGEVAERTFQRMVV
                  **** :**:**: **:******:* ::.:** *: :***.:.** ***:*:***: ***.

k141_1100_8       KPDWGIRTVTCEDGKDHVVEVQTVVEKPFGDLIYFKVPGRDIQPRKVLLVAPMSGHYATL
k141_12002_3      KPSWGIRTMTCEDGKDHIVEKQIAVEKPFGDLIHFKVLGRDPSPRKVLLVAPMSGHYATL
k141_1302_3       KPDWGIRTFTCEDGKDHLVDIRTVVERPFGDLVHFAVAGRKTQPRKVLLVAPMSGHYATL
k141_1111_2       KPDWGIPSFTCEDGKDHVVEIANVVERPFGDLIHFKVSGRDEQPRKVLLVAPMSGHYATL
k141_9880_8       KPDWGIPSITMADGKDHIVTVENVVPGDFGDLIHFAVSGRKPMKRKVLVVAPMSGHYATL
                  **.*** :.*  *****:*    .*   ****::* * **.   ****:***********

k141_1100_8       LRSTVASLLPDCELYVTDWHNARDIPVSAGKFDVEDYTLYLVDFMRHLGPDTHVIAVCQP
k141_12002_3      LRSTVRSLIPDCEVYVTDWHNARDIPVSAGKFDIEDYTLYLMEFMRFLGPDTNVIAVCQP
k141_1302_3       LRSTVKSLLVNCEVYITDWHNARDIPVSEGKFDIEDYTLYLVDFMRELGPDTHVVAVCQP
k141_1111_2       LRSTVQSLIRNCEVYVTDWHNARDIPVSAGKFDVEDYTLYLVDFMRELGPDIHVVAVCQP
k141_9880_8       LRSTVISLLPDCEVYITDWHNARDIPVSAGKFDVEDYTLYLVDYMRHLGPDTHVIAVCQP
                  ***** **: :**:*:************ ****:*******:::** **** :*:*****
```

```
k141_1100_8       APLTLAATAYLAEQDPKAQPLSLTLIGGPIDPDATPTEVTDFGRRVTMGQLEETAIQRVG
k141_12002_3      APLTLAATAYLAELDPDAQPRTLTLIGGPIDPNATPTEVTDFGHRVTMGQLEETMIQRVG
k141_1302_3       VPLTLAATAYLAEQDPKAQPTSLTLIGGPVDPDATPTEVTDFGRRVTMGQLEETMIQRVG
k141_1111_2       VPLTLAATAYLAEQDPKAQPSSLTLIGGPVDPDATPTEVTDFGRRVTMGQLEETMIQRVG
k141_9880_8       APLTLAATAYLAEEDPKAQPRTLTLIGGPIDPDAAATDVTDFGRRVTMGQLEQFMIQRVG
                  .************ **.*** :*******:**:*:.*:*****:********:  *****

k141_1100_8       FSYPGVGRLVYPGLLQLASFMSMNSDRHSEAFSGQIQREMRGEAGDHDAHNRFYDEYLAV
k141_12002_3      FKYKGVGRMVYPGLLQLASFMSMNADRHSKAFTDQIQRVAKGEGADHDQHNRFYDEYLAV
k141_1302_3       FKYKGVGRMVYPGLLQLASFMSMNGERHSKAFSDQIQRVMRGDASDHDAHNRFYDEYLAV
k141_1111_2       FKYPGVGRKVYPGLLQLASFMSMNGERHSKAFMDQIQRVSRGEASDHDAHNRFYDEYLAV
k141_9880_8       FKYPGVGRMVYPGLLQLASFMSMNWDSHSTAFSDQINRAARGEAADHDKHNKFYDEYLAV
                  *.* **** *************** : ** ** .**:*  :*:..*** **:********

k141_1100_8       MDMTAEFYLSTVERIFKNGEIAKNQFVVDGHKVDIGKITNVAVKTVEGANDDISAPGQCI
k141_12002_3      MDMCAEFYLSTVDRIFKKGEIAKNEFVVDGHKVDIGKITNVAVKTVEGANDDISAPGQCV
k141_1302_3       MDMTAEFYLSTVERVFKNREIARNEFVVDGHKVDIGKITDVAVKTVEGANDDISAPGQCV
k141_1111_2       MDMTAEFYLSTVERIFKNREIARNEFVVNGHKVDFSKITDVAVKTVEGAKDDISAPGQCI
k141_9880_8       MDMTAEFYLSTVERVFKNREIAKNVFTVAGKKVDIGKVTTVAVKTVEGAKDDISAPGQCI
                  *** ********:*:**: ***:* *.* *:***:.*:* *********:*********:

k141_1100_8       AALALCTGLPDSMKASHVEPGAGHYGIFAGRSWRDNIRPLVIEFMNTNSTKPVQKRKPAR
k141_12002_3      AALDLCTGLPDSKKASHLEPGAGHYGIFAGRSWRDNIRPLVLEFIDANRADAPARKAAE-
k141_1302_3       AALDLLTGLPDAKKASHVEPGAGHYGIFAGRSWRDNIRPLVIDFMNAADAARPARGKPAN
k141_1111_2       AALDLCTGLPEEKKASHVEPGAGHYGIFAGRSWRNNIRPLVIDFMNANSRKKPAAHKPAN
k141_9880_8       AALDLLTGLPDSKKASHVEPDAGHYGIFAGKSWRNNIRPLVLDFIDSNSKPVKEPANKNA
                  *** * ****:  ****:**.*********:***:******::*:::

k141_1100_8       IANTNAAAR
k141_12002_3      ---------
k141_1302_3       KNAAAR---
k141_1111_2       KNTAA----
k141_9880_8       AA-------
```

Figure 1: A homologous protein Closely related to MCC4266688.1 , an PHB depolymerase family esterase, appears in 5 metagenomics samples

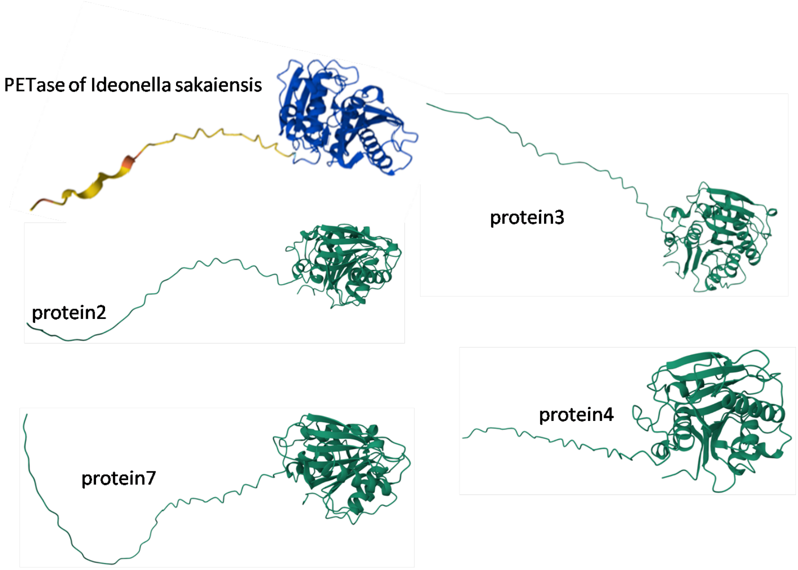


Figure 2: Structures of 4 AI generated proteins are similar to that of PETase