PLOS Computational Biology

RESEARCH ARTICLE

PlasticEnz: An integrated database and screening tool combining homology and machine learning to identify plastic-degrading enzymes in meta-omics datasets

Anna Krzynowek, Jasper Snoeks, Karoline Faust*

Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, Laboratory of Molecular Bacteriology, KU Leuven, Leuven, Belgium

* karoline.faust@kuleuven.be

OPEN ACCESS

Citation: Krzynowek A, Snoeks J, Faust K (2026) PlasticEnz: An integrated database and screening tool combining homology and machine learning to identify plastic-degrading enzymes in meta-omics datasets. PLoS Comput Biol 22(1): e1013892. https://doi.org/10.1371/journal.pcbi.1013892

Editor: Eduardo Jardón-Valadez, Universidad Autonoma Metropolitana - Lerma, MEXICO

Received: June 2, 2025

Accepted: January 6, 2026

Published: January 26, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1013892

Abstract

PlasticEnz is a new open-source tool for detecting plastic-degrading enzymes (plastizymes) in metagenomic data by combining sequence homology-based search with machine learning techniques. It integrates custom Hidden Markov Models, DIAMOND alignments, and polymer-specific classifiers trained on ProtBERT embeddings to identify candidate depolymerases from user-provided contigs, genomes, or protein sequences. PlasticEnz supports 11 plastic polymers with ML classifiers for PET and PHB, achieving F1 > 0.7 on an independent test set. Applied to plastic-exposed microcosms and field metagenomes, the tool recovered known PETases and PHBases, distinguished plastic-contaminated from pristine environments, and clustered predictions with validated reference enzymes. *PlasticEnz* is fast, scalable, and user-friendly, providing a robust framework for exploring microbial plastic degradation potential in complex communities.

Author summary

Plastic pollution is a global problem, and one promising solution is to apply microbes to break them down. However, finding the enzymes responsible for this in complex environmental samples is not easy. We developed ***PlasticEnz***, a free and easy-to-use tool that helps researchers identify plastic-degrading enzymes or "plastizymes" in metagenomic data. *PlasticEnz* combines traditional sequence similarity search methods with machine learning models trained on previously known plastizymes. It works with protein sequences, contigs, or genomes with ML components optimised for classification of two common plastizymes: PETases and PHBases. We tested *PlasticEnz* on both controlled lab experiments and real-world samples from plastic-polluted soils and clean environments. The tool successfully identified known plastic-degrading enzymes

Data availability statement: All relevant data underlying the findings of this study, including the raw data used for model training, benchmarking, intermediate files, scripts to generate figures and tables are provided within the manuscript, supporting information files, and Zenodo repository (10.5281/zenodo.15395662). PlasticEnz is freely available at https://github.com/msysbio/PlasticEnz. All the raw data, training sets and scripts are available at 10.5281/zenodo.15395662.

Funding: AK recieved funding from Fonds Wetenschappelijk Onderzoek (FWO) PhD fellowship 1S03725N. https://fwo.be The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. AK recived salary from the funder.

Competing interests: The authors have declared that no competing interests exist.

and even helped distinguish between polluted and pristine sites. By making plastizyme detection more accessible, *PlasticEnz* enables researchers to better explore the microbial potential for plastic degradation, which could support future bioremediation efforts.

Introduction

Plastic pollution is a growing environmental problem posing serious risks to ecosystems, wildlife, and human health [1–4]. Despite this, only a small fraction of the millions of tons of plastic waste generated each year is recycled or reused, with most ending up in natural environments [5,6]. The capacity of some microorganisms to biodegrade plastic polymers has attracted considerable attention as a potential strategy aimed at mitigating plastic pollution [7–14]. In particular, polyester plastics (e.g., polyethylene terephthalate (PET), polylactic acid (PLA), polycaprolactone (PCL)), which are widely used in textile and packaging manufacturing, are of interest for potential microbial bioremediation due to the presence of repeated ester bonds that can be targeted by various extracellular depolymerases [11,13,15,16]. Recent advances in bioinformatics combined with decreasing cost of whole-genome sequencing, have greatly improved our ability to screen complex communities for novel enzymes. However, existing computational tools often lack specificity for plastic substrates and involve complex multi-step pipelines that might not be accessible for researchers without bioinformatics expertise.

Several publicly available databases now catalogue protein sequences and associated metadata for enzymes involved in plastic degradation [17–19]. Commonly used approaches to identify candidate plastic-degrading enzymes involve large-scale homology searches against these databases using fast sequence aligners such as DIAMOND [20], Bowtie2 [21], or Minimap2 [22]. Other methods rely on domain-based annotation, such as screening for specific functional domains using tools like InterProScan [23] or Pfam [24]. In addition, custom Hidden Markov Motifs (HMMs) [25,26] built from curated enzyme sequences and tailored to specific plastic polymers have been employed before to improve search specificity [27,28]. More recently, the application of machine learning (ML) to protein function prediction is being increasingly explored, enabling the discovery of novel enzyme candidates that lack clear homology to known proteins [29–35]. In this context, ML models have been applied to predict plastic-degrading enzymes, as demonstrated by Jiang et al. [36,37]. However, these models were not tested on real-world metagenomics datasets and are not readily available to be applied to the user's own data.

To address these limitations, we developed *PlasticEnz*, an open-access tool that combines homology-based search using custom HMMs with machine learning prediction to improve the detection of plastic-degrading enzymes. In this study, we describe the development, testing, and application of *PlasticEnz*. The tool accepts protein sequences, genomic assemblies, or contigs as input, and identifies candidate plastic-degrading enzymes using a combination of custom HMMs, DIAMOND-based

homology searches, and optional machine learning classification. Specifically, HMM-based screening is available for poly(3-Hydroxypropionate (P3HP), poly(butylene adipate-co-terephthalate) (PBAT), polybutylene succinate (PBS), poly(butylene succinate-co-adipate) (PBSA), polycaprolactone (PCL), poly(ethylene adipate) (PEA), polyethylene terephthalate (PET), polyhydroxybutyrate (PHB), poly(3-hydroxybutyrate-co-3-hydroxyvalerate) (PHBV), and polylactic acid (PLA); DIAMOND-based searches are implemented for PBS, PBSA, PCL, PES [polyethersulfone], PHBV, and PLA; and machine learning predictions using XGBoost (default) and a more sensitive neural network model are currently available for PET and PHB (S1 Table). This integrated approach allows *PlasticEnz* to flexibly detect enzyme homologs across a wide range of plastic polymers with adjustable sensitivity and specificity. To facilitate accessibility, *PlasticEnz* is implemented as a command-line tool with streamlined output formats that include HMMER/DIAMOND outputs such as bitscore and E-values, ML prediction scores and normalized gene abundances (TPM/RPKM), making the results easier to interpret and compare across diverse metagenomic datasets.

We applied *PlasticEnz* to both controlled microcosm experiments and diverse field metagenomes, demonstrating its ability to discriminate plastic-exposed from pristine environments. In the Laguna Madre microcosm, PlasticEnz identified a strong enrichment of PHB depolymerase homologs in PHA biofilms, while PETase signals remained consistently low across all treatments, aligning with the findings of the original study. In plastic-contaminated soil samples, those collected in Sewapura and Varamin exhibited the highest abundance and prediction scores for both PET and PHB depolymerases, whereas pristine Kamchatka acid hot springs communities yielded negligible hits. Benchmarking against our test set, the ML classifiers achieved F1 values above 0.7 for PET and PHB, with XGBoost maximizing precision and the neural network showing enhanced sensitivity. Sequence-level analyses further confirmed that predicted homologs clustered closely with experimentally validated reference enzymes.

PlasticEnz offers a streamlined and accessible solution for identifying plastic-degrading enzymes in metagenomic data by combining homology-based and machine learning approaches. By integrating curated reference data with predictive models in a single pipeline, it enables researchers regardless of computational background to explore microbial plastic degradation potential across a wide range of environments.

Results

PlasticEnz database

The *PlasticEnz* database contains 213 unique protein sequences associated with plastic polymer degradation pathways, extracted from 176 peer-reviewed studies. Each database entry includes detailed annotation of the enzyme's gene name, location (e.g., extracellular or cell-bound), enzyme classification, operon or gene cluster, targeted bond types, and catalytic domains. Enzymes are also linked to their associated reactions (substrate-product) and source organisms, for which marker genes or whole-genome sequences are provided. Furthermore, the database integrates cross-references to external databases (UniProt, NCBI). The comparison with other available *Plastizyme* databases is shown in Table 1.

Table 1. Comparison of stored information across available plastic-degrading enzyme databases.

Feature	PlasticEnz	PlasticDB [18]	PAZy [19]
Reactions annotated	Yes (substrate–product pairs)	No	Limited (some polymer types)
Microorganism sequences	Genome or 16S rRNA sequence provided	No	No
Links to other databases	NCBI, UniProt	None	GenBank, Uniprot, PDB
Enzyme location (cellular)	Specified	Not available	Specified for some entries
Enzyme isolation environment	Specified for some enzymes	Not available	Not available

https://doi.org/10.1371/journal.pcbi.1013892.t001

ML model evaluation

To assess the predictive performance of PlasticEnz, we compared three machine learning models (Neural Network, Random Forest, and XGBoost) for their ability to classify plastic-degrading enzymes across polymer substrates. This comparison aimed to determine which model best balances sensitivity and precision when identifying plastizyme candidates in metagenomic data (**Fig 1**). Overall, all models struggled to accurately classify PLA, PBAT, PBSA, PCL and PHA polymers. However, for PET and PHB, both the Neural Network and XGBoost models achieved F1 values above 0.7 (Table 2). The XGboost model demonstrated higher precision (0.95 and 1 for PET and PHB) than the Neural network (0.84 and 0.63 for PET and PHB), indicating a lower rate of false positive classifications (paired t-test, p-value < 0.01, t = -71.0). However, it

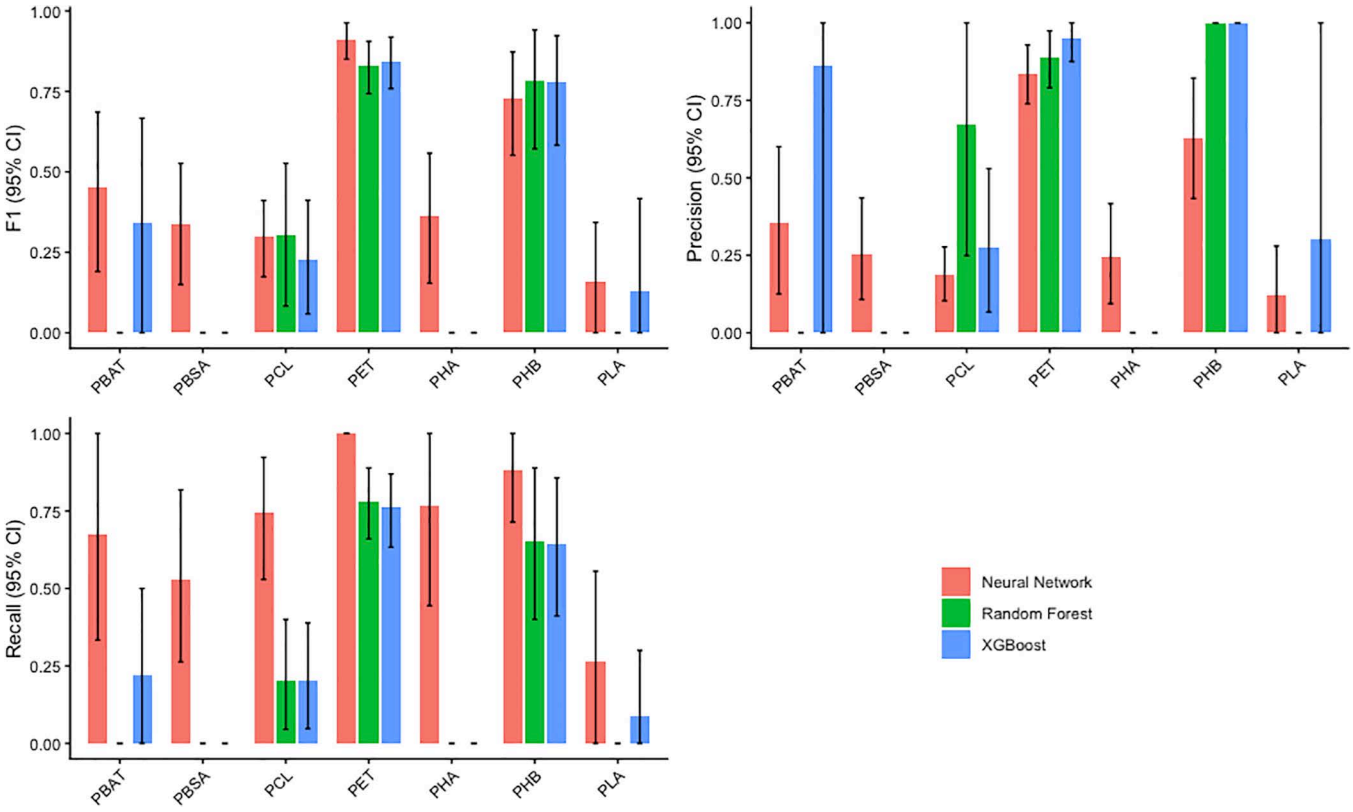

Fig 1. Bar charts depicting the mean bootstrapped (n = 1000) evaluation metrics (F1, Precision, and Recall) for classification of each plastic-degrading enzyme class.

https://doi.org/10.1371/journal.pcbi.1013892.g001

Table 2. Performance metrics of Neural Network and XGBoost models for PET and PHB classification. Numbers in the brackets represent 95% confidence intervals.

Model	Polymer	Precision	Recall	F1
Neural Network	PET	0.84 (0.74–0.93)	1.00 (1.00–1.00)	0.91 (0.85–0.96)
Neural Network	PHB	0.63 (0.43–0.82)	0.88 (0.71–1.00)	0.73 (0.55–0.87)
XGBoost	PET	0.95 (0.88–1.00)	0.76 (0.63–0.87)	0.84 (0.76–0.92)
XGBoost	PHB	1.00 (1.00–1.00)	0.64 (0.41–0.86)	0.78 (0.58–0.92)

https://doi.org/10.1371/journal.pcbi.1013892.t002

was overall more conservative in its identification, as reflected by lower recall (0.76 and 0.64 for PET and PHB) than Neural Network (1 and 0.88 for PET and PHB) (paired t-test, p-value < 0.01, t = 59.34).

The performance differences between the two models largely stem from the class imbalance in the training data, with many more true negatives than true positives. This led XGBoost to adopt a conservative prediction strategy, yielding high precision by minimizing false positives, which is useful when high-confidence predictions are desired. In contrast, the neural network achieved higher recall by detecting a broader range of potential degraders, though with more false positives. Both models are available in PlasticEnz to accommodate different research goals: XGBoost as the default high-confidence predictions and a more sensitive neural network (via the –sensitive flat) for broader candidate detection.

Different machine learning models are represented by distinct colors, and error bars indicate the 95% confidence intervals.

PlasticEnz tool workflow description, runtime, and performance

PlasticEnz identifies plastic-degrading enzymes from metagenomic data using a two-step search and optional machine learning classification. The tool accepts contigs, genomes, or protein sequences as input and screens them against a curated database using HMMER [25] and DIAMOND [20]. Users can specify the target polymer(s) with the --polymer flag. For PET and PHB, predictions can be refined using a machine learning classifier, namely XGBoost (default) or a more sensitive neural network (activated via --sensitive). If paired metagenomic reads or BAM files are supplied, the tool also estimates gene abundances and reports their raw as well as CPM, RPKM and TPM normalized counts. The output is a report containing information about the predicted plastic-degrading enzyme including its protein sequence, normalized abundance values, sequence similarity scores (E-value, bitscore) and ML prediction score (if applicable). The full workflow is depicted and described in **Fig 2**.

During runtime tests, the tool performed well across inputs of various types and sizes (S5 Table). For smaller inputs such as 1.2Mb single genomes or 100Mb protein files, the real runtime remained under 30 seconds. For medium-sized datasets, including genomes and proteins within 600–650Mb range, *PlasticEnz* workflow completed in under 4 minutes. The runtime for large datasets of nearly 1Gb remained practical, under 5 minutes for proteins and 35 minutes for contigs.

PlasticEnz matches or outperforms existing tools in PHB and PET enzyme annotation on the independent validation set

Benchmarking results against are summarized in **Table 3**.

For PET-degrading enzymes, PlasticEnz in Sensitive mode achieved the best overall performance across all metrics (precision = 0.98, recall = 0.96, F1 = 0.97, MCC = 0.95). PlasticEnz in Default mode performed comparably to eggNOG-mapper (F1 = 0.86 vs 0.90; MCC = 0.82 vs 0.87), whereas KEGG-based tools (KOfamKOALA and BlastKOALA) failed to identify any PET hydrolases, resulting in zero values for all metrics.

For PHB-degrading enzymes, all tools showed lower performance, reflecting the higher sequence diversity within this enzyme group. PlasticEnz in Sensitive mode again performed best (F1 = 0.67, MCC = 0.65). The Default mode, eggNOG-mapper, and KOfamKOALA achieved similar scores (all F1 = 0.58, MCC = 0.61), while BlastKOALA performed weakest (F1 = 0.30, MCC = 0.39).

Taken together, these benchmarking results demonstrate that while PlasticEnz in Default mode performs on par with established annotation tools such as eggNOG-mapper, the Sensitive mode consistently outperformed all other annotation tools, most notably for PET-degrading enzymes.

Application of PlasticEnz to PHB and PET-exposed benthic biofilm communities

We evaluated whether plastizyme candidates detected by PlasticEnz reflected expected environmental gradients and showed sequence similarity to experimentally validated enzymes. Specifically, we compared the abundance and confidence scores of predicted plastizymes (HMM E-value and bitscore, and ML prediction score) across PET, PHB, and control biofilm communities to assess whether putative plastic-degrading enzymes were more prevalent in polymer-exposed

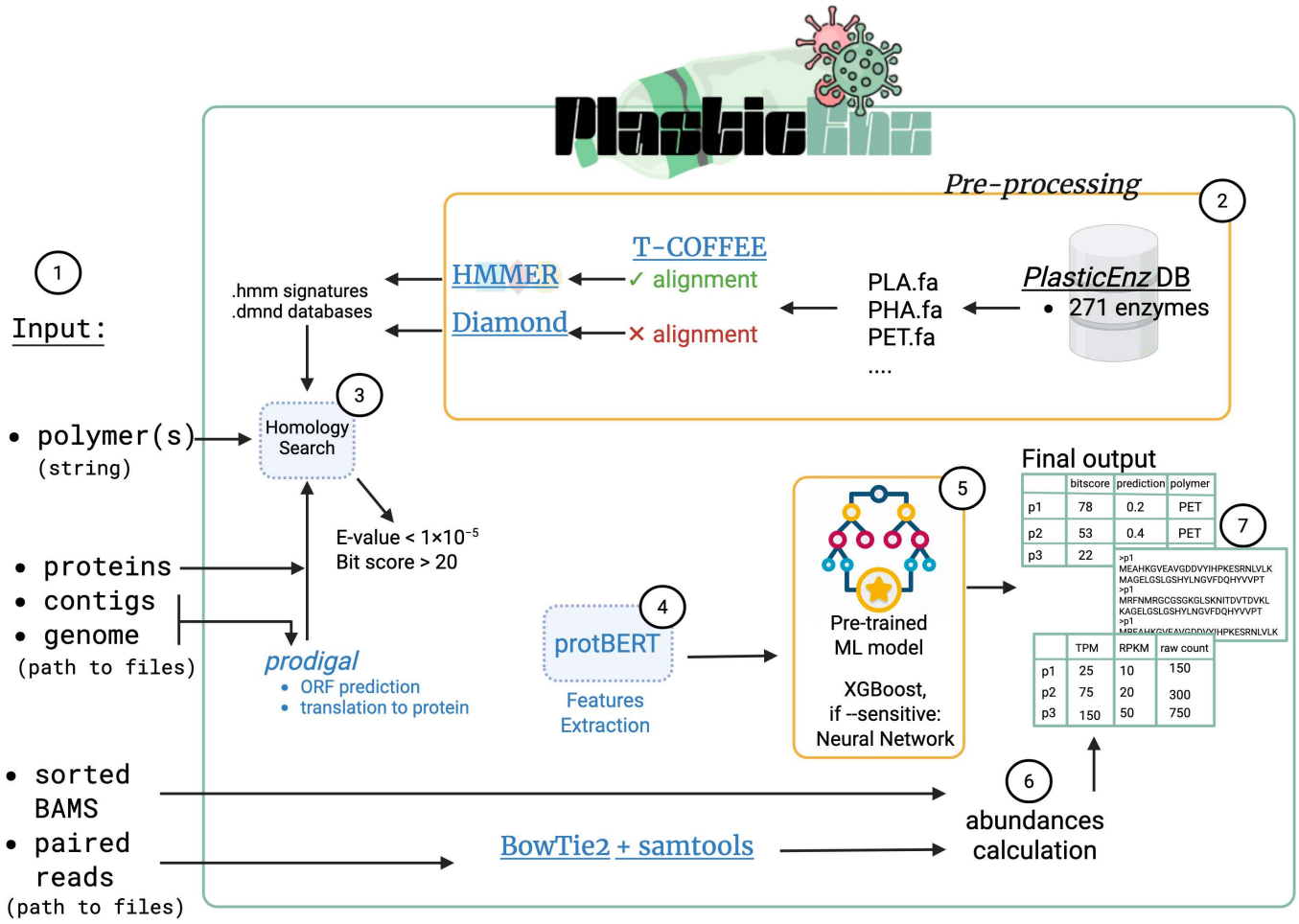

Fig 2. Overview of the PlasticEnz pipeline. PlasticEnz identifies candidate plastic-degrading enzymes in metagenomic datasets through a multi-step workflow. (**1**) The user specifies a target plastic polymer or combination of polymers using the --polymer flag and provides a path to one of the following: assembled contigs (--contigs), full genomes (--genome), or protein sequences (--proteins) in FASTA format. Optionally, paired-end sequencing reads or pre-aligned BAM files may be included to quantify gene abundance. (**2**) For nucleotide inputs (contigs or genomes), protein-coding genes are predicted using Prodigal [38] and translated to amino acid sequences. (**3**) The resulting proteins are screened against our custom-made HMM profiles using HMMER [25] or DIAMOND [20] (singleton sequences). Hits must pass default filters (E-value < 1×10^{-5}, bitscore > 20); HMMER hits are further filtered by bias score (must be < 10% of bitscore). (**4**) Sequences that pass this homology screen are embedded using ProtBERT [39] to generate contextualized feature vectors. (**5**) These embeddings are classified using one of two pre-trained machine learning models: XGBoost (default mode) or a neural network (sensitive mode, activated with --sensitive, optimized for recall). Predictions are returned as probabilities for each supported polymer class (currently PET and PHB). (**6**) If read data is provided, gene abundance is computed either via alignment-based quantification (Bowtie2 [21] + samtools [40]) or directly from sorted BAM files using internal scripts. (**7**) Final outputs include: a summary.csv file listing hits, homology scores, and ML prediction scores; a.fasta file with protein sequences of predicted homologs; and an optional abundance.csv file containing raw and normalized counts (RPKM, TPM, CPM). External tools and packages are marked in blue. This figure was created using BioRender and we have obtained full permission for its use in the publication.

https://doi.org/10.1371/journal.pcbi.1013892.g002

than in control biofilms, as observed in the original study. We further examined the functional relatedness of these predictions by comparing the top-scoring candidates with known plastizymes from the PlasticEnz database using phylogenetic trees and pairwise evolutionary distance analyses. For this evaluation, we applied PlasticEnz in both the default (XGBoost) and sensitive (Neural Network) modes to assembled contigs from PET, PHB, and ceramic biofilm communities, as well as seawater samples (H_2O), targeting PET and PHB as polymer variables.

Table 3. Comparative performance metrics of PlasticEnz vs reference tools annotations.

Model / Tool	Polymer	Precision	Recall	F1	MCC
PlasticEnz: Default (XGBoost)	PET	1.00	0.76	0.86	0.82
PlasticEnz: Sensitive (Neural Network)	PET	0.98	0.96	0.97	0.95
EggNOG	PET	1.00	0.82	0.90	0.87
KOfamKOALA	PET	0.00	0.00	0.00	0.00
BlastKOALA	PET	0.00	0.00	0.00	0.00
PlasticEnz: Sensitive (Neural Network)	PHB	0.90	0.53	0.67	0.65
PlasticEnz: Default (XGBoost)	PHB	1.00	0.41	0.58	0.61
EggNOG	PHB	1.00	0.41	0.58	0.61
KOfamKOALA	PHB	1.00	0.41	0.58	0.61
BlastKOALA	PHB	1.00	0.18	0.30	0.39

https://doi.org/10.1371/journal.pcbi.1013892.t003

First, we compared HMMER bitscores across all biofilm types. *PlasticEnz* identified 7 putative PETases in the PET biofilm community, with an average HMMER bitscore of 43.9 (SD: 9.6) (S4 Table). In comparison, the average HMMER bitscores for seawater and ceramic samples were 52.0 (SD: 12.8) and 41.7 (SD: 12.1), respectively. Next, we examined ML classifier prediction scores under both model settings. As expected, the sensitive model produced much higher average prediction values than the default model. For PET, seawater, and ceramic samples, average scores under the default model were 0.06 (SD: 0.05), 0.05 (SD: 0.04), and 0.04 (SD: 0.06), respectively, compared to 0.6 (SD: 0.2), 0.43 (SD: 0.2), and 0.3 (SD: 0.2) in the sensitive mode (Fig 3A). Finally, we quantified the number of proteins with high-confidence predictions, defined as the number of proteins reaching prediction scores > 0.7. In the default mode, no PETases in any sample met this threshold. However, in the sensitive mode, 3 PETases in the seawater sample (1.48 hits per million proteins) and 2 in the PET biofilm sample (1.45 hits per million proteins) passed this cutoff (Fig 3B). No high-confidence putative PETases were detected in the ceramic biofilms under either setting.

In the PHB biofilm community, *PlasticEnz* identified 826 putative PHB depolymerases with a high average HMM bitscore of 108.4 (SD: 91.1). In contrast, average HMM bitscores were substantially lower for the seawater (56.7, SD: 29.1) and ceramic communities (60.0, SD: 36.5) (S4 Table).

Classifier prediction scores for PHB homologs were higher than those observed for PET, across all modes. In the default (XGBoost) mode, average prediction scores were 0.2 (SD: 0.3) for the PHB biofilm, 0.07 (SD: 0.1) for seawater, and 0.09 (SD: 0.2) for ceramic samples (Fig 3C). Again, these values increased substantially under the sensitive (Neural Network) mode, reaching 0.6 (SD: 0.2) for PHB, and 0.5 (SD: 0.3) for both seawater and ceramic samples (Fig 3C).

Similarily to PETases, quantified the number of proteins with high-confidence predictions (score > 0.7). Under the default model, 64 proteins in the PHB biofilm (38.5 hits per million proteins), 3 in seawater (1.48 hits per million), and 5 in ceramic biofilms (3.69 hits per million) surpassed this threshold. In contrast, the sensitive mode yielded more high-scoring predictions: 363 in the PHB biofilm (219 hits per million), 61 in seawater (30.1 hits per million), and 45 in ceramic samples (33.1 hits per million) (Fig 3D).

To assess the sequence similarity between PlasticEnz-predicted PETases/ PHBases and experimentally validated plastizymes, we compared the top 10 predictions from each model (PHB-default, PHB-sensitive, and PET-sensitive) with sequences from the PlasticEnz database.

For the PHB-default model (Fig 4A), most predicted proteins showed close sequence alignment to several known enzymes, with an average evolutionary distance of 2.84 (range: 0.80–6.66). The closest matches included known PHB depolymerases from *Pseudomonas lemoignei* (PHB_SEED_3), *Alcaligenes faecalis* (PHB_SEED_4), and *Ralstonia pickettii* (PHB_SEED_5, PHB_SEED_26), all with distances below 1.6. The most dissimilar hits were to enzymes from *Cupriavidus necator* (e.g., PHB_SEED_41, _44, _36), with distances exceeding 4.5.

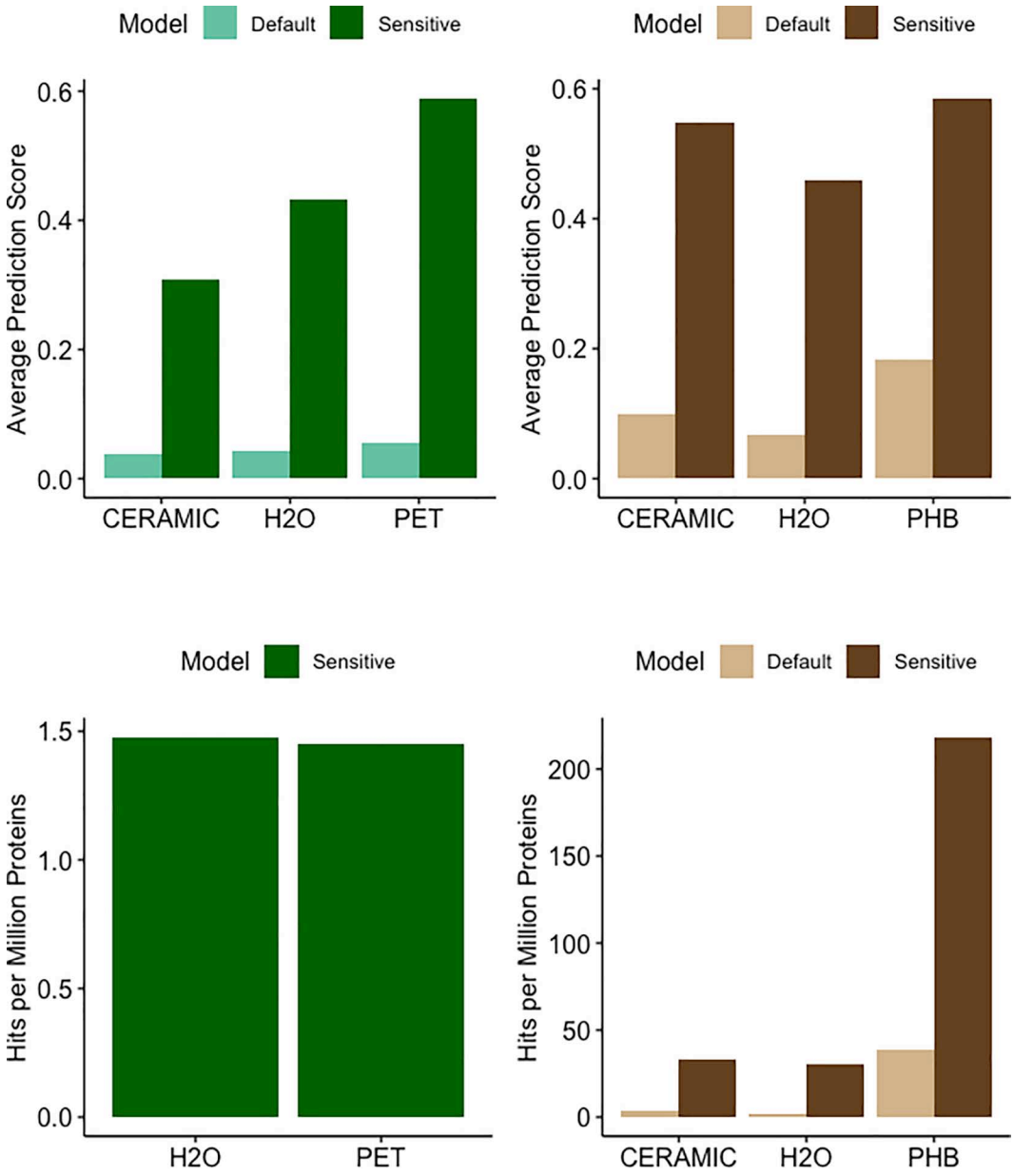

Fig 3. PlasticEnz predictions of PET- and PHB-degrading enzymes with prediction scores above 0.7 across samples from the Laguna Madre dataset (CERAMIC, PET, PHB, H_2O), shown for both the default and sensitive PlasticEnz models. Average prediction scores for PET- (**A**, Green) and PHB- (**C**, Brown) degrading enzyme candidates in microbial communities from PET, PHB, ceramic, and seawater (H_2O) samples. Results from the default PlasticEnz model are shown in lighter colours, and results from the sensitive model are shown in darker colours. Total abundances of PlasticEnz-predicted putative enzymes with above 0.7 prediction threshold for PET (**B,** Green) and PHB (**D,** Brown), normalized for sample depth and expressed as proteins per million.

https://doi.org/10.1371/journal.pcbi.1013892.g003

In the PHB-sensitive model (Fig 4B), predicted homologs were more diverse, with a higher average evolutionary distance of 3.78 (range: 0.51–10.00), indicating broader but less conserved matches. However, some sequences still aligned well with known depolymerases, including those from *Ralstonia pickettii* (PHB_SEED_5, _26), *Burkholderia cepacia* (PHB_SEED_18), and *Pseudomonas lemoignei* (PHB_SEED_24). The most distant predictions again mapped to

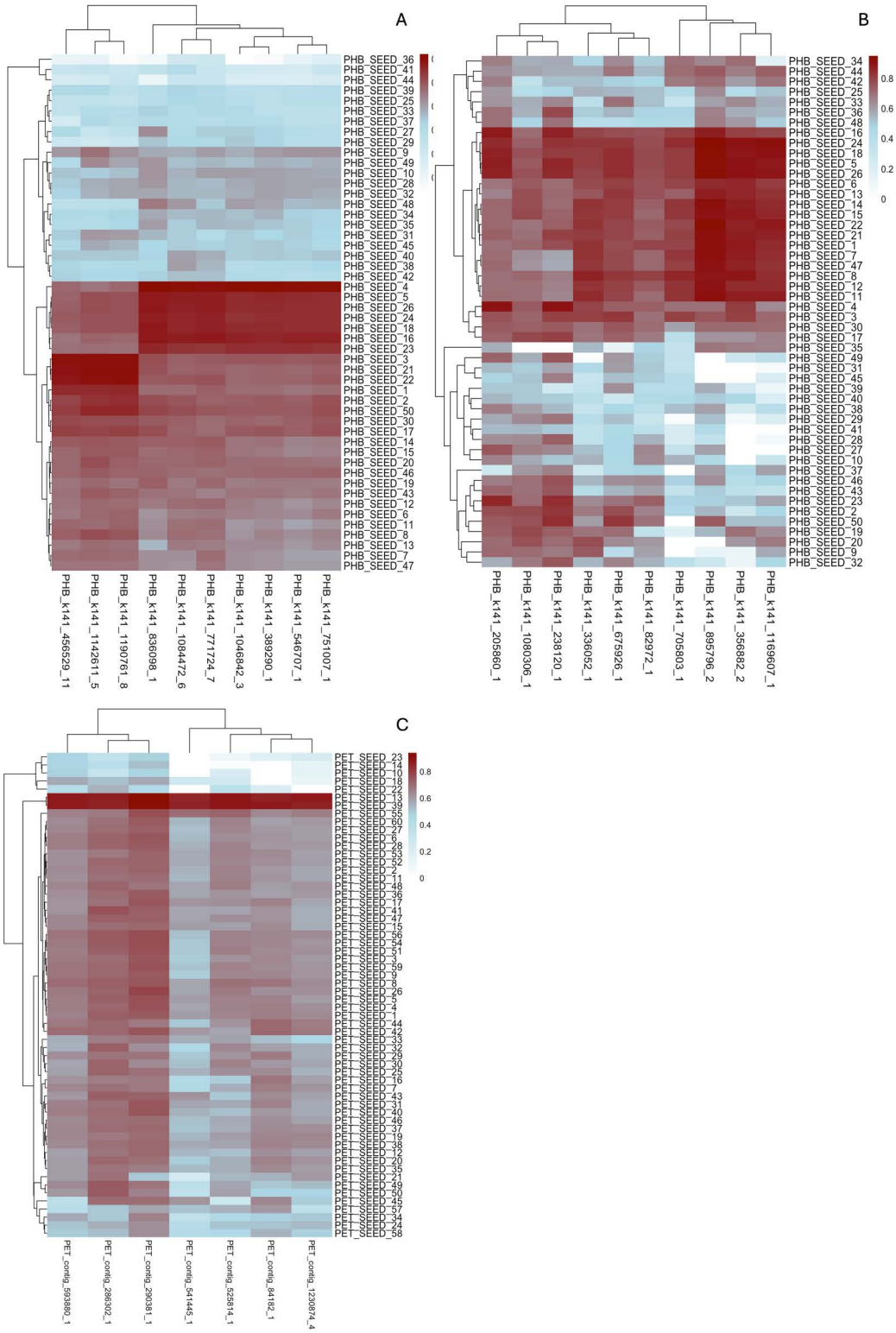

Fig 4. Protein sequence similarities between the top PlasticEnz predictions in the Laguna Madre dataset and their closest database reference sequences. Heatmaps display the sequence similarity between the top 10 highest-scoring predicted PETases and PHBases identified by *PlasticEnz* for PHB under default model **(A)**, PHB under sensitive model (**B**) and PET under sensitive mode **(C)**. *PlasticEnz*-identified plastizymes were compared

against their respective non-redundant reference enzymes from the database of confirmed plastic degrading enzymes. Pairwise evolutionary distances were computed using the LG substitution model and converted into similarity scores ranging from 0 (low similarity) to 1 (high similarity).

https://doi.org/10.1371/journal.pcbi.1013892.g004

Cupriavidus necator (PHB_SEED_41, _45, _39) and *Paracoccus denitrificans* (PHB_SEED_31), with distances above 6.3. No PETases were found under the default model, meanwhile the PET-sensitive model (Fig 4C) yielded the least conserved matches overall, with an average distance of 3.99 (range: 0.58–10.00). While most predictions showed weak similarity, a few aligned moderately well with reference enzymes such as PET_SEED_13 (*Bacillus subtilis*, p-nitrobenzylesterase), PET_SEED_39 (uncultured bacterium), and PET_SEED_7 (*Pseudomonas aestusnigri*), all showing average distances around 1.2–1.4. The most distant matches were against the fungal PETases from *Fusarium oxysporum* (PET_SEED_10, _14), *Fusarium solani* (PET_SEED_23), and *Humicola insolens* (PET_SEED_22), with distances exceeding 7.4. Furthermore, we compared averaged HMM bit scores and neural network prediction scores for contigs with the highest number of sequence matches to known plastizymes (e.g., *contig_593880_1*, *contig_290381_1*, *contig_286302_1*) and those with the fewest (e.g., *contig_1230874_4*, *contig_541445_1*, *contig_525814_1*). The difference in average HMM scores between the two groups was 48.7 vs. 41.2, meanwhile the ML model prediction scores showed 0.71 vs. 0.50.

Application of PlasticEnz to microbial communities from plastic-rich and pristine field samples

We applied PlasticEnz to metagenomes from plastic-contaminated soils and pristine hot springs to evaluate whether predicted plastizyme distributions corresponded with environmental exposure gradients. Specifically, we used *PlasticEnz* to identify putative plastic-degrading enzymes in metagenomic samples from plastic-contaminated urban soils (Varamin, Sewapura, Ghazipur) and compared the results to those from thermophilic sediment samples collected from pristine hot springs (Hot_springs_1, _2, _3).

We first examined the average HMM bitscores of PlasticEnz-identified plastizyme candidates across all polymers: P3HP, PBAT, PBS, PBSA, PCL, PEA, PES, PET, PHA, PHB, PHBV and PLA (Fig 5A). Plastizyme candidates from plastic contaminated soil samples consistently exhibited higher average HMMER bitscore values across all polymers. In contrast, candidates from hot spring samples had consistently lower HMMER bitscores, and for several polymers (e.g., P3HP, PEA in Hot_spring_3; PES, PET in all samples; PHB in Hot_spring_1) no candidates were detected at all.

Next, we quantified high-scoring plastizymes defined as candidates that exceeded HMMER bitscores of 100, 80, and 50, representing high, medium, and low sequence similarity to the *PlasticEnz* HMM motifs, respectively (Fig 6). Across all thresholds and polymer types, *PlasticEnz* consistently detected more high scoring plastizyme candidates in plastic-contaminated soils as opposed to hot spring samples. After normalizing for sequencing depth, Sewapura had the highest number of candidates, followed by Ghazipur and Varamin. As expected, relaxing the HMMER bitscore threshold increased the number of candidates. For example, the number of predicted PLA-degrading enzymes in hot spring samples rose from 3 (bitscore > 100), to 23 (bitscore > 80), and 99 (bitscore > 50). At the lowest threshold (bitscore > 50), the total number of predicted depolymerases also increased across other polymer classes in the pristine hot springs, reaching 4 for PBAT, 47 for PBSA, 14 for PCL, 4 for PHBV, and 46 for PLA.

Next, we examined the *PlasticEnz* ML prediction module by quantifying the number of plasizyme candidates classified as high-confidence PET or PHB depolymerases (prediction score > 0.7) under the default model (Fig 5B). As the highest prediction score for proteins in the hot spring samples was 0.0001, only plastic-contaminated sites were considered. As expected, across all contaminated sites, PHB depolymerases were more abundant than PETases. Sewapura exhibited the highest number of high confidence candidates, with 7.28 and 2.43 classified depolymerases per million proteins for PHB and PET, respectively. Varamin followed with 6.23 and 0.78 for PHB and PET while Ghazipuer showed the lowest abundance of PHB depolymerases (5.16 classified depolymerases per million proteins) and no PETases exceeding the 0.7 prediction threshold.

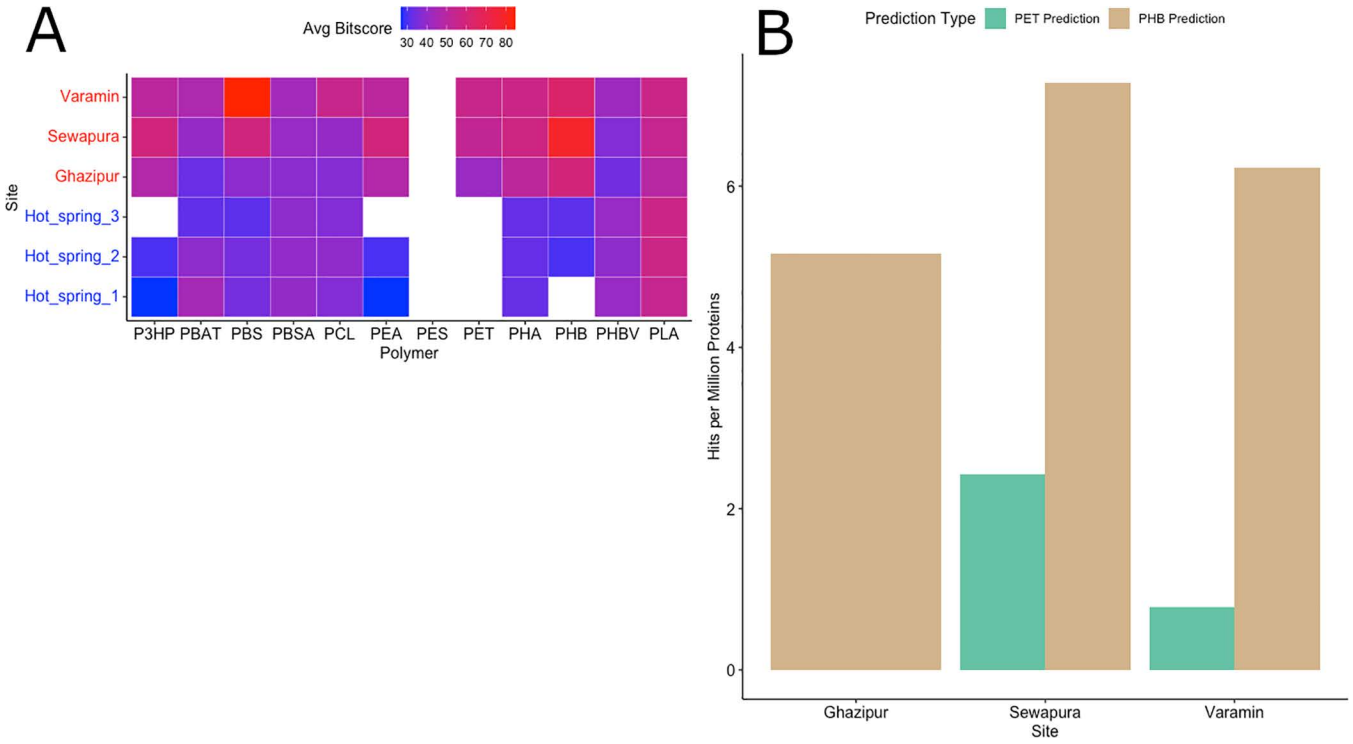

Fig 5. ***PlasticEnz*** **prediction results for plastic-contaminated urban soil (Ghazipur, Sewapura, Varamin) and pristine hot springs samples (Hot_spring_1–3). (A)** Heatmap showing average HMMER bitscores for PlasticEnz predicted plastic-degrading enzymes across sites and across polymers screened. Sites from pristine hot springs are in blue and plastic contaminated soil samples in red. **(B)** Abundances of proteins (expressed as hits per million proteins) predicted by *PlasticEnz* ML component (default mode) as putative PET (brown) or PHB (green) depolymerases at 0.7 prediction threshold.

https://doi.org/10.1371/journal.pcbi.1013892.g005

Finally, we assessed the sequence similarity between *PlasticEnz*-predicted PETases from contaminated sites and experimentally validated PET-degrading enzymes from the *PlasticEnz* reference database (Fig 7). Here, we compared two subsets: (i) PETases predicted with high confidence by the default machine learning component (prediction score > 0.7; Fig 7A), and (ii) PETases identified using the stringent HMM bitscore threshold (> 80; Fig 7B).

The ML-predicted candidates from Varamin, Sewapura, and Ghazipur clustered closely with multiple known PETases on the PCoA plot, exhibiting higher sequence similarity (PERMANOVA on the clusters, $p > 0.05$, $R^2 = 0.020$; beta-dispersion < 0.05) (Fig 7A **and** 7B). The ML-predicted candidates displayed strong similarity with diverse PETases from the databse, including PET_SEED_51 (*Nocardioidaceae*), PET_SEED_60 (*Marinactinospora thermotolerans*), PET_SEED_53 (*Saccharopolyspora flava*), PET_SEED_47 (*unknown bacterium*), and PET_SEED_48 (*Micromonosporaceae*) (Fig 7C). The mean pairwise distance between the predicted candidates and known plastizymes was 0.56 ± 0.12. In contrast, candidates with high HMMER bitscores formed a more distinct clade, with stronger homology to only two known PETases: PET_SEED_13 (p-nitrobenzylesterase from *Bacillus subtilis*) and PET_SEED_39 (unknown PET hydrolase from an uncultured bacterium) (Fig 7D). These sequences were more divergent from known PETases, as reflected by a higher mean pairwise distance (0.74 ± 0.04), and clear separation in the PCoA space (Fig 7B).

Discussion

Our study introduces *PlasticEnz*, a new bioinformatics tool designed to aid researchers in screening for potential plastic-degrading enzymes in complex metagenomic datasets. At the core of *PlasticEnz* is a carefully curated database containing experimentally validated plastic polymer degradation enzymes. This database allowed us to generate the

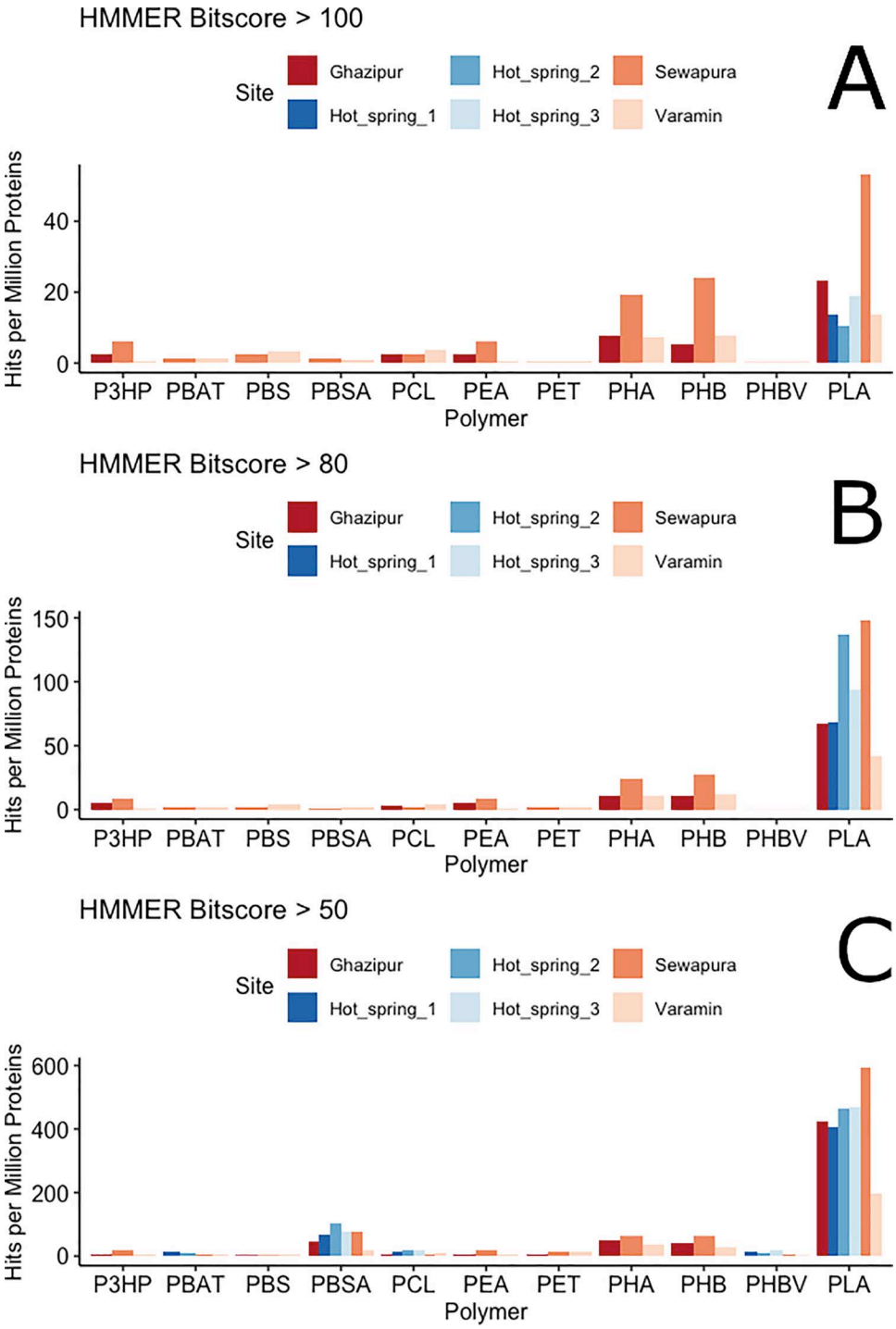

Fig 6. ***PlasticEnz*** **prediction results for bacteria from plastic-contaminated urban soils (Ghazipur, Sewapura, Varamin) and pristine hot springs samples (Hot_spring_1–3) (C–E) Bar plots showing the abundances of predicted plastic-degrading enzymes (expressed as hits per million proteins) for each screened polymer (X-axis) and site (colour) across varying HMM bitscore thresholds: > 100 (C), > 80 (D), and >50 (E).**

https://doi.org/10.1371/journal.pcbi.1013892.g006

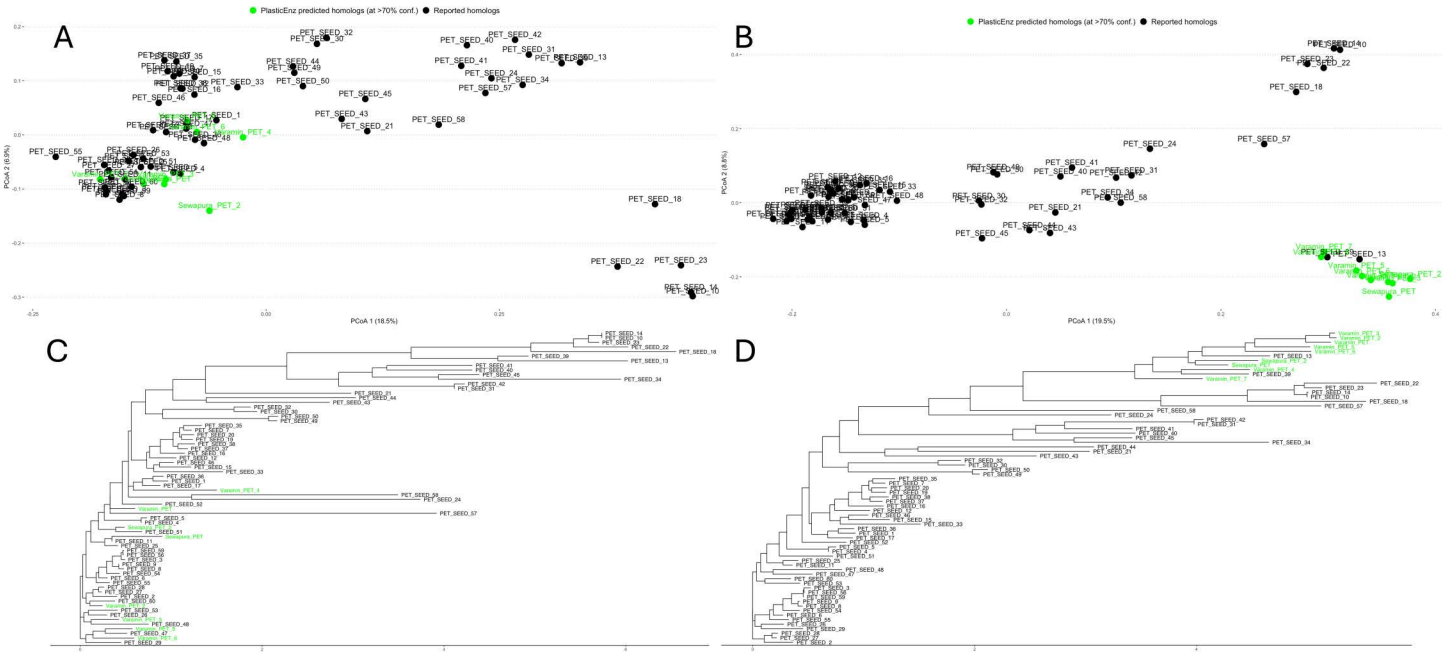

Fig 7. Evolutionary relatedness of *PlasticEnz* predicted PETases to known reference PET-degrading enzymes. (A, B) Principal Coordinates Analysis (PCoA) based on Fitch distances showing sequence-level similarities between *PlasticEnz*-predicted putative PETases (green) and known reference PETases (black) under high-confidence ML predictions (≥ 0.7) (**A**) or HMMER bitscore above 80 **(B)**. **(C, D)** Corresponding phylogenetic trees derived from Clustal Omega alignments and built with FastTree under the LG substitution model, showing clustering of predicted PETases (green) with reference enzymes (black).

https://doi.org/10.1371/journal.pcbi.1013892.g007

polymer-specific Hidden Markov Models (HMM) and provided a resource for training a machine learning (ML) classifier for PETases and PHBases annotation. Benchmarking against common protein annotation tools shows that PlasticEnz's strategy of combining custom and polymer-specific HMM motifs with pre-trained ML component outperforms common tools like KEGG mapper or eggNOG-mapper in annotation of PETases and PHBases.

A critical step in developing the machine learning component of PlasticEnz was the construction of clearly defined positive and negative datasets. For the positive set, i.e., ground truth sequences shown to degrade plastic polymers experimentally, we used the data from PlasticEnz and PlasticDB [18] databases. However, creating a reliable negative, defined as a set of sequences of distantly homologous proteins without proven ability to degrade plastic polymers was challenging due to the widespread presence of plastic contamination across all the planet's environments [41–43]. To overcome this, we used sequences from well-characterized bacteria that depend on hosts or live parasitically, making them unlikely to synthesise extracellular enzymes targeting plastic polymers. Furthermore, given the limited size of our dataset, we avoided deep learning architectures, which are prone to overfitting under data-scarce conditions [44,45]. Instead, we employed simpler, more interpretable models such as decision trees, gradient-boosted trees (XGBoost), and a shallow neural network (single hidden layer). To further reduce overfitting risk and improve generalizability, we incorporated regularization techniques such as early stopping and dropout for the neural network [46,47] and hyperparameter optimization for all models. In line with previous findings [37], which evaluated thirteen classifiers on plastic degradation enzyme classification problems, XGBoost emerged as the most suitable model. Overall, all models showed strong performance for two polymers: poly(ethylene terephthalate) (PET) and polyhydroxybutyrate (PHB), but consistently underperformed on others like PLA, PBAT, PBSA, and PCL. For these underrepresented classes, precision, recall, and F_1 scores were consistently low and exhibited wide confidence intervals. This is a common effect of class imbalance, where the number of

true positive examples is greatly outweighed by negative instances. As the size of the positive set increases, performance metrics such as precision and recall improve accordingly, which is supported by a positive correlation coefficient (Pearson, $r = 0.81$) between the F_1 score and the number of sequences in the positive set. These results highlight a key limitation in training ML models for plastic degradation classification: the limited number of ground truth sequences. As more enzymes are identified and incorporated into the training set, model performance is expected to improve accordingly. Therefore, to ensure reliable predictions, the *PlasticEnz* prediction module is limited to PET and PHB, the only polymers with sufficient training data. Users can choose between a default XGBoost model optimized for precision, or a more sensitive neural network model that emphasizes recall, allowing flexibility based on research priorities.

To showcase *PlasticEnz* capabilities, we applied it to two distinct experimental setups: (1) a controlled microcosm study conducted in a hypersaline lagoon (Laguna Madre), and (2) field samples from plastic-contaminated and pristine environments. In the Laguna Madre dataset, originally analyzed by Pinell et al., 2022 [48] metagenomic analysis focused on biofilm communities growing on PET, PHA, and ceramic substrates over 28 days. Consistent with the original findings, *PlasticEnz* did not detect significant enrichment of PET-degrading enzymes in PET biofilms compared to seawater or ceramic controls. No high-confidence PETase was found in any sample under the default mode, and even under the sensitive mode, PET biofilms did not show increased PETase predictions relative to seawater. This aligns with Pinell's interpretation that the remote location and low plastic exposure likely limited the selection pressure for PET-degrading enzymes in these communities. In contrast, the original study reported that communities extracted from PHA biofilms were significantly enriched in PHB depolymerases in comparison to control biofilms. Using *PlasticEnz*, we identified over 800 putative PHB depolymerases in these communities, with significantly higher average HMM bitscore and ML prediction scores relative to control samples. Notably, heatmaps showed that many of these predicted enzymes shared high sequence similarity to experimentally validated entries in the SEED database. However, *in-vitro* assays will be required to confirm their functional activity.

To further demonstrate the capabilities of *PlasticEnz*, we applied the tool to screen microbial communities from two contrasting environments. While plastic-contaminated urban soils of Varamin, Sewapura and Ghazipur served as representative examples of polymer-enriched ecosystems, identifying a truly pristine, plastic-free environment proved challenging due to the ubiquity of microplastics [27,28,41–43,49,50]. We selected the geothermal area of the Mutnovsky volcano in Kamchatka, a remote region with minimal anthropogenic activity, as a proxy for a low-contamination environment. This site is characterized by extreme physicochemical conditions, including broad gradients in temperature and pH [51]. Due to limited organic carbon sources, these communities are dominated by obligately or facultatively chemolithoautotrophic bacteria and the presence of hydrolytic enzymes capable of degrading complex plastic polymers was expected to be minimal [52,53]. As expected, plastic-contaminated soils contained putative plastic-degrading homologs with higher average HMM bitscores and greater numbers of depolymerases than pristine hot spring samples. The strongest enrichment was seen in Sewapura, followed by Ghazipur and Varamin. Notably, all sites, including hot springs, showed elevated PLA depolymerase signals. This is because most PLA-degrading enzymes belong to the serine protease family, a group of diverse and taxonomically widespread enzymes commonly found across bacterial, fungal, and archaeal taxa [8,54–56].

Next, we evaluated the *PlasticEnz* default prediction module using a classification threshold of 0.7. None of the putative PET or PHB homologs from the hot spring samples met this threshold; in fact, all prediction scores were below 0.1. Among the plastic-contaminated sites, PETase predictions exceeding the 0.7 confidence threshold were observed in communities from Sewapura and Varamin but not in Ghazipur. Meanwhile, high-confidence predictions of PHB depolymerases were common to all sites. This likely reflects underlying biological differences: PHB depolymerases are widespread due to the role of PHA polymers in microbial carbon and energy storage, whereas PET degradation requires evolved adaptations of esterases or cutinases [7,14]. As a result, PET-degrading enzymes remain rare and are typically associated with long-term plastic exposure.

Lastly, we conducted a comparison study by clustering *PlasticEnz*-identified putative PET homologs to previously reported and functionally validated PETases (SEED sequences). This analysis was split into two comparisons: one

including high-scoring homologs (HMM bitscore > 80), and a second including sequences assigned a prediction score greater than 0.7 by the XGBoost classifier (default mode). ML-predicted PETase homologs from plastic-contaminated sites clustered with several known reference PETases, while the high HMM score sequences formed a distinct clade associated with two particular SEED reference enzymes. These results demonstrate that the *PlasticEnz* prediction module can identify candidate PETases with strong evolutionary ties to a broader range of validated enzymes, extending beyond the reach of traditional homology-based methods like HMMs, especially in complex metagenomic datasets. However, it is also possible that the model may be biased toward well-represented sequence patterns in the training data, potentially reducing its ability to detect PETases that were underrepresented in the training set.

While *PlasticEnz* offers a streamlined and accessible framework for identifying putative plastic-degrading enzymes, several limitations should be considered. First, the presented analysis focused on high-confidence predictions based on stringent HMM bitscore and ML probability thresholds; users are encouraged to apply similar cutoffs to ensure reproducibility. Additionally, the speed of the pipeline scales with dataset size, and RAM requirements increase accordingly, particularly during ProtBERT embeddings and contigs translation to proteins. To address this, a --gpu option is provided to accelerate tokenization, and the tool allows the use of multiple cpu cores via the --cores flag. However, since prodigal is not optimised to execute on multiple cores, the users are advised to adjust computational resources as needed for bigger datasets (over 1Gb). Importantly, *PlasticEnz* relies on sequence homology, and its predictions remain putative until validated through *in vitro* assays. The accuracy and breadth of the tool are ultimately constrained by the number and diversity of experimentally verified plastic-degrading enzymes available for model training. As the field advances and more polymer-degrading enzymes are experimentally validated, *PlasticEnz* will be continuously updated with expanded HMM profiles and retrained machine learning classifiers, enhancing its predictive accuracy, polymer coverage, and utility for metagenomic screening in diverse environments.

Materials and methods

Curation and processing of plastic-degrading enzyme sequence data

We systematically collected data on experimentally confirmed plastic degradation enzymes, including their protein sequences, host organisms, reaction details (including catalysts and conditions), references to relevant publications, and cross-references to external databases such as UniProt [57], NCBI [58], and KEGG [59]. This comprehensive information was extracted from published literature and stored in a *PlasticEnz* SQL database (available within the *PlasticEnz* package). To enhance our dataset, we also incorporated data from PlasticDB [18], a publicly available database specializing in plastic degradation enzymes. In total, the two combined databases resulted in 422 protein sequences of diverse plastic-degrading enzymes. These protein sequences were then pooled together into combined fasta files based on their respective polymer substrate. To eliminate redundancy caused by the database merge, we clustered each combined fasta at 95% similarity using CD-HIT2 [60]. Following clustering, multiple sequence alignments (MSA) were performed for each grouped set of unique proteins using T-Coffee in expresso mode [61], which integrates protein structure. The resulting alignments were refined to maintain only those with average to good alignment score (TCoffee alignment score > 50). Sequences that failed to align adequately were pressed into the DIAMOND database [20] (S2 Table). MSAs were used to generate HMM profiles using the built-in *hmmbuild* function from the HMMER suite [25] (S2 Table).

Preparation of training and test data sets

To build the negative dataset, we used our previously generated HMM profiles to identify homologous sequences in bacterial genomes that are unlikely to be involved in plastic degradation activity. Specifically, we focused on representative Refseq genomes from well-studied host-dependent or parasitic bacteria (e.g., *Escherichia coli*, *Chlamydia trachomatis*, *Staphylococcus aureus*) (S3 Table). The resulting distant homologous sequences formed the negative dataset (410 sequences). For the positive dataset, we used our previous set of experimentally confirmed bacterial plastic-degrading

enzymes. Positive and negative datasets were combined and clustered at 95% identity with CD-HIT2 [60] to obtain 502 unique clusters. To avoid the presence of highly similar sequences between training and test datasets, we randomly split entire clusters into either the training set (80%, 606 protein sequences) or the test set (20%, 200 protein sequences). The feature matrix used for the model training contained one-hot encoded annotations indicating the specific plastic polymer(s) degraded by each enzyme.

Embeddings generation with ProtBERT

To generate protein embeddings, we used ProtBERT [39], a pre-trained transformer-based model specifically trained on protein sequences. ProtBERT and its tokenizer were loaded from the Hugging Face Transformers library. Sequences were preprocessed by trimming ambiguous residues ('X', 'x') from sequence ends, verifying that only standard amino acids (ACDEFGHIKLMNPQRSTVWY) were present, and formatting each sequence by inserting spaces between individual residues. ProtBERT generated contextual embeddings for each amino acid, which were then aggregated into a fixed-length embedding vector for each protein by mean pooling across the sequence length.

Machine learning model selection and evaluation

Full training procedures, hyperparameter settings, and performance metrics are detailed under '*Machine Learning Model Selection and Evaluation'* in S2 Data.

We evaluated three classifiers: neural network, Random Forest, and XGBoost, using precomputed protein embeddings and one-hot encoded features. All models were trained on a multi-label dataset for PET and PHB degradation using an 80/20 train/test split. Hyperparameters were optimized via grid search and validated on held-out data. Final models were retrained on the full training set and evaluated on the independent test set using precision, recall, and F1-score (S2 File). Model performance (F1-score, precision and recall) was compared between Neural Network and XGBoost classifiers using paired *t*-tests. Normality of metric distributions was confirmed using Shapiro–Wilk tests ($p > 0.05$). Mean, standard deviation, test statistics, and p-values were reported for each metric. Models were implemented and trained with PyTorch, scikit-learn, and XGBoost libraries. Prediction scores from XGBoost and neural network models are included in the PlasticEnz module and represent per-class probabilities for each polymer.

Benchmarking PlasticEnz

To evaluate the performance of PlasticEnz against existing protein tools, we benchmarked PlasticEnz in default (XGBoost) and sensitive (Neural Network) modes against widely used functional annotation platforms: EggNog-mapper (v. 2.1.12)[73], KOfamKOALA[74] and BlastcKOALA[75] (S1 Data). Since no public, gold-standard metagenomics dataset containing experimentally validated plastizymes and reliable negative examples currently exists, we conducted the benchmarking evaluation on the independent validation dataset (S2 and S3 and S4 Data files). This dataset comprises of protein sequences that were excluded from model training but utilized for classifier performance assessment in **Fig 1** & **Table 2**. In brief, proteins associated with PET and PHB degradation were extracted from the independent validation set and annotated using each platform under their default parameters. A correct annotation for PET-degrading enzymes was defined as assignment to K21104 (polyethylene terephthalate hydrolase), while correct annotations for PHB-degrading enzymes corresponded to K05973 or K03932 (poly(3-hydroxybutyrate) depolymerase and polyhydroxybutyrate depolymerase, respectively) and were classified as true positives. Any alternative KO assignment or absence of assignment for this set was considered a false negative. The negative control set consisted of proteins containing homologous polyester-degrading motifs but derived from bacteria known to lack plastic-degrading capability (e.g., parasitic or host-dependent organisms). If any of these proteins were incorrectly annotated as K21104, K05973, or K03932, they were classified as false positives; otherwise, they were considered true negatives. Following the same criteria established as during the bootstrap-based performance estimation, predictions with a PlasticEnz score

exceeding 0.5 for PET or PHB were classified as positive hits. Proteins with scores below 0.5 or without any annotation were considered negative (no hit).

Evaluation sets

A full description of datasets, processing steps, and analysis parameters is provided under '*Evaluation sets* ' in S1 File.

In brief, we analyzed paired-end Illumina metagenomes from a 28-day microcosm study by Pinnell et al. (2019) [NCBI: PRJEB15404] [48], which examined biofilm communities developed on PET, PHA, and ceramic beads in a hypersaline lagoon. Four sample groups were used: PET, PHA (PHB), ceramic, and seawater controls (H_2O). Raw reads were quality-filtered with Fastp [62] (default settings) and assembled using MEGAHIT (default settings) [63]. Assembled contigs were used directly as an input for *PlasticEnz*. Hit counts for PHB/PET depolymerases were normalized per million predicted proteins for comparison across samples. To assess functional similarity, top 10 enzyme predictions from PET-sensitive, PHB-default, and PHB-sensitive models were compared to *PlasticEnz*'s curated SEED reference set (CD-HIT, 90% identity). Sequences were aligned with *MUSCLE* [64], trimmed using *trimAl* (–automated1 setting) [65], and evolutionary distances were calculated with the LG model [66] in *phangorn* (R). Normalized pairwise distances were visualized as similarity heatmaps.

We validated *PlasticEnz* custom HMM motifs using datasets from environments with contrasting plastic exposure levels. Thermophilic hot spring sediment samples from Kamchatka, Russia (NCBI: PRJNA419239), served as a low-exposure control, while plastic-contaminated soil microbiomes from India and Iran, including Sewapura (NCBI: PRJNA1077790) [67], Varamin (NCBI: PRJNA924045) [68], and Ghazipur (NCBI: PRJNA388130), served as high-exposure test sites. Downloaded raw paired-end reads were processed identically to Laguna madre samples. High-confidence PETase predictions (prediction score > 0.7 (default model) or HMMER bitscore > 80) were aligned with SEED references using *Clustal Omega* [69]. Sequence similarity was assessed with Fitch distances [70] (*seqinr*) [71] and visualized using Principal Coordinates Analysis (*ape*) [72]. Phylogenies were generated with *FastTree* [73] using trimmed alignments (*trimAl* [65], gap threshold 0.5).

Analyses were performed in R (v4.2) and Python (v3.11.11). Visualizations were generated using *ggplot2*, *ggpubr*, *patchwork* and *pheatmap*.

Code availability and supporting information

PlasticEnz is freely available at https://github.com/msysbio/PlasticEnz

All relevant data underlying the findings of this study, including the raw data used for model training, benchmarking, intermediate files, scripts to generate figures and tables are provided within the manuscript, Supporting Information files and zenodo repository (10.5281/zenodo.15395662). No additional restrictions apply.

Supporting information

S1 Table. Supported polymers and available search strategies.
(XLSX)

S2 Table. Seed sequence counts for hMMER and DIAMOND searches.
(XLSX)

S3 Table. Description of positive and negative training sequences.
(XLSX)

S4 Table. PlasticEnz results for Laguna Madre microcosm study.
(XLSX)

S5 Table. Speed tests results.
(XLSX)

S1 File. Supplementary Materials & Methods.
(DOCX)

S2 File. Formulas for Precision, Recall and F1-Score.
(PDF)

S1 Data. Benchmarking values calculation.
(XLSX)

S2 Data. Protein sequencs used for benchmarking tests of negative set.
(FASTA)

S3 Data. Protein sequencs used for benchmarking tests of PET set.
(FASTA)

S4 Data. Protein sequencs used for benchmarking tests of PHB set.
(FASTA)

Acknowledgments

We thank Maxime Greffe and Hubert Krukowski for kindly testing the tool.

Author contributions

Conceptualization: Anna Krzynowek.

Funding acquisition: Anna Krzynowek, Karoline Faust.

Investigation: Karoline Faust.

Methodology: Anna Krzynowek.

Project administration: Anna Krzynowek, Karoline Faust.

Software: Anna Krzynowek, Jasper Snoeks.

Supervision: Anna Krzynowek, Karoline Faust.

Validation: Anna Krzynowek.

Visualization: Anna Krzynowek.

Writing – original draft: Anna Krzynowek.

Writing – review & editing: Anna Krzynowek, Karoline Faust.

References

1. MacLeod M, Arp HPH, Tekman MB, Jahnke A. The global threat from plastic pollution. Science. 2021;373(6550):61–5.
2. Borrelle SB, Ringma J, Law KL, Monnahan CC, Lebreton L, McGivern A. Predicted growth in plastic waste exceeds efforts to mitigate plastic pollution. Science. 2020;369(6510):1515–8.
3. Allouzi MMA, Tang DYY, Chew KW, Rinklebe J, Bolan N, Allouzi SMA, et al. Micro (nano) plastic pollution: The ecological influence on soil-plant system and human health. Sci Total Environ. 2021;788:147815.
4. Thushari GGN, Senevirathna JDM. Plastic pollution in the marine environment. Heliyon. 2020;6(8):e04709. https://doi.org/10.1016/j.heliyon.2020.e04709 PMID: 32923712

5. Kedzierski M, Frère D, Le Maguer G, Bruzaud S. Why is there plastic packaging in the natural environment? Understanding the roots of our individual plastic waste management behaviours. Sci Total Environ. 2020;740:139985.
6. Geyer R, Jambeck JR, Law KL. Production, use, and fate of all plastics ever made. Sci Adv. 2017;3(7):e1700782. https://doi.org/10.1126/sciadv.1700782 PMID: 28776036
7. Paloyan A, Tadevosyan M, Ghevondyan D, Khoyetsyan L, Karapetyan M, Margaryan A. Biodegradation of polyhydroxyalkanoates: current state and future prospects. Front Microbiol. 2025;16:1542468. https://doi.org/10.3389/fmicb.2025.1542468
8. Lim H-A, Raku T, Tokiwa Y. Hydrolysis of polyesters by serine proteases. Biotechnol Lett. 2005;27(7):459–64. https://doi.org/10.1007/s10529-005-2217-8 PMID: 15928850
9. Yoshida S, Hiraga K, Takehana T, Taniguchi I, Yamaji H, Maeda Y, et al. A bacterium that degrades and assimilates poly(ethylene terephthalate). Science. 2016;351(6278):1196–9. https://doi.org/10.1126/science.aad6359 PMID: 26965627
10. Jendrossek D, Knoke I, Habibian RB, Steinbüchel A, Schlegel HG. Degradation of poly(3-hydroxybutyrate), PHB, by bacteria and purification of a novel PHB depolymerase from Comamonas sp. J Environ Polym Degrad. 1993;1(1):53–63.
11. Qi X, Ma Y, Chang H, Li B, Ding M, Yuan Y. Evaluation of PET Degradation Using Artificial Microbial Consortia. Front Microbiol. 2021;12.
12. Yang J, Yang Y, Wu W-M, Zhao J, Jiang L. Evidence of polyethylene biodegradation by bacterial strains from the guts of plastic-eating waxworms. Environ Sci Technol. 2014;48(23):13776–84. https://doi.org/10.1021/es504038a PMID: 25384056
13. Sriyapai P, Chansiri K, Sriyapai T. Isolation and characterization of polyester-based plastics-degrading bacteria from compost soils. Microbiology. 2018;87(2):290–300.
14. Wei R, Zimmermann W. Microbial enzymes for the recycling of recalcitrant petroleum-based plastics: how far are we?. Microb Biotechnol. 2017;10(6):1308–22. https://doi.org/10.1111/1751-7915.12773
15. Yoshida S, Hiraga K, Takehana T, Taniguchi I, Yamaji H, Maeda Y, et al. A bacterium that degrades and assimilates poly(ethylene terephthalate). Science. 2016;351(6278):1196–9. https://doi.org/10.1126/science.aad6359 PMID: 26965627
16. Mergaert J, Swings J. Biodiversity of microorganisms that degrade bacterial and synthetic polyesters. J Ind Microbiol. 1996;17(5):463–9.
17. Plastics Meta-omic Database (PMDB). https://www.plasticmdb.org. Accessed 2025 April 30.
18. Gambarini V, Pantos O, Kingsbury JM, Weaver L, Handley KM, Lear G. PlasticDB: a database of microorganisms and proteins linked to plastic biodegradation. Database. 2022;2022:baac008.
19. Buchholz PCF, Feuerriegel G, Zhang H, Perez-Garcia P, Nover L-L, Chow J, et al. Plastics degradation by hydrolytic enzymes: The plastics-active enzymes database-PAZy. Proteins. 2022;90(7):1443–56. https://doi.org/10.1002/prot.26325 PMID: 35175626
20. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18(4):366–8. https://doi.org/10.1038/s41592-021-01101-x PMID: 33828273
21. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923 PMID: 22388286
22. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
23. Zdobnov EM, Apweiler R. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001;17(9):847–8. https://doi.org/10.1093/bioinformatics/17.9.847 PMID: 11590104
24. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Research. 2021;49(D1):D412–9.
25. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195 PMID: 22039361
26. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol. 1994;235(5):1501–31. https://doi.org/10.1006/jmbi.1994.1104 PMID: 8107089
27. Danso D, Schmeisser C, Chow J, Zimmermann W, Wei R, Leggewie C, et al. New Insights into the Function and Global Distribution of Polyethylene Terephthalate (PET)-Degrading Bacteria and Enzymes in Marine and Terrestrial Metagenomes. Appl Environ Microbiol. 2018;84(8):e02773–17. https://doi.org/10.1128/AEM.02773-17 PMID: 29427431
28. Zrimec J, Kokina M, Jonasson S, Zorrilla F, Zelezniak A. Plastic-Degrading Potential across the Global Microbiome Correlates with Recent Pollution Trends. mBio. 2021;12(5):e0215521. https://doi.org/10.1128/mBio.02155-21 PMID: 34700384
29. Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci U S A. 2019;116(28):13996–4001. https://doi.org/10.1073/pnas.1821905116 PMID: 31221760
30. Bileschi ML, Belanger D, Bryant DH, Sanderson T, Carter B, Sculley D, et al. Using deep learning to annotate the protein universe. Nat Biotechnol. 2022;40(6):932–7. https://doi.org/10.1038/s41587-021-01179-w PMID: 35190689
31. Kulmanov M, Khan MA, Hoehndorf R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2018;34(4). https://doi.org/10.1093/bioinformatics/btx684
32. Kim GB, Kim JY, Lee JA, Norsigian CJ, Palsson BO, Lee SY. Functional annotation of enzyme-encoding genes using deep learning with transformer layers. Nat Commun. 2023;14(1):7370. https://doi.org/10.1038/s41467-023-43216-z PMID: 37963869

33. Zheng L, Shi S, Lu M, Fang P, Pan Z, Zhang H, et al. AnnoPRO: a strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path encoding. Genome Biol. 2024;25(1):41. https://doi.org/10.1186/s13059-024-03166-1 PMID: 38303023
34. Bordin N, Dallago C, Heinzinger M, Kim S, Littmann M, Rauer C, et al. Novel machine learning approaches revolutionize protein knowledge. Trends Biochem Sci. 2023;48(4):345–59. https://doi.org/10.1016/j.tibs.2022.11.001 PMID: 36504138
35. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics. 2022;38(8):2102–10. https://doi.org/10.1093/bioinformatics/btac123
36. Jiang R, Yue Z, Shang L, Wang D, Wei N. PEZy-miner: An artificial intelligence driven approach for the discovery of plastic-degrading enzyme candidates. Metab Eng Commun. 2024;19:e00248.
37. Jiang R, Shang L, Wang R, Wang D, Wei N. Machine learning based prediction of enzymatic degradation of plastics using encoded protein sequence and effective feature representation. Environ Sci Technol Lett. 2023;10(7):557–64.
38. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119. https://doi.org/10.1186/1471-2105-11-119 PMID: PMID: 20211023
39. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans Pattern Anal Mach Intell. 2022;44(10):7112–27. https://doi.org/10.1109/TPAMI.2021.3095381 PMID: 34232869
40. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352
41. Lusher AL, Tirelli V, O'Connor I, Officer R. Microplastics in Arctic polar waters: the first reported values of particles in surface and sub-surface samples. Sci Rep. 2015;5:14947. https://doi.org/10.1038/srep14947 PMID: 26446348
42. O'Brien S, Rauert C, Ribeiro F, Okoffo ED, Burrows SD, O'Brien JW. There's something in the air: A review of sources, prevalence and behaviour of microplastics in the atmosphere. Sci Total Environ. 2023;874:162193.
43. Wang J, Liu X, Li Y, Powell T, Wang X, Wang G. Microplastics as contaminants in the soil environment: A mini-review. Sci Total Environ. 2019;691:848–57.
44. Montesinos López OA, Montesinos López A, Crossa J. Overfitting, Model Tuning, and Evaluation of Prediction Performance. Multivariate Statistical Machine Learning Methods for Genomic Prediction. Springer International Publishing. 2022:109–39. https://doi.org/10.1007/978-3-030-89010-0_4
45. Li H, Rajbahadur GK, Lin D, Bezemer C-P, Jiang ZM. Keeping Deep Learning Models in Check: A History-Based Approach to Mitigate Overfitting. IEEE Access. 2024;12:70676–89. https://doi.org/10.1109/access.2024.3402543
46. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–58.
47. Prechelt L. Early Stopping — But When? Lecture Notes in Computer Science. Springer Berlin Heidelberg. 2012:53–67. https://doi.org/10.1007/978-3-642-35289-8_5
48. Pinnell LJ, Turner JW. Shotgun metagenomics reveals the benthic microbial community response to plastic and bioplastic in a coastal marine environment. Front Microbiol. 2019. https://doi.org/10.3389/fmicb.2019.01252
49. Sajjad M, Huang Q, Khan S, Khan MA, Liu Y, Wang J. Microplastics in the soil environment: A critical review. Environ Technol Innov. 2022;27:102408.
50. Wong JKH, Lee KK, Tang KHD, Yap PS. Microplastics in the freshwater and terrestrial environments: prevalence, fates, impacts and sustainable solutions. Sci Total Environ. 2020;719:137512.
51. Bessonova EP, Bortnikova SB, Gora MP, Manstein YUA , Shevko AYA, Panin GL, et al. Geochemical and geo-electrical study of mud pools at the Mutnovsky volcano (South Kamchatka, Russia): Behavior of elements, structures of feeding channels and a model of origin. Applied Geochemistry. 2012;27(9):1829–43. https://doi.org/10.1016/j.apgeochem.2012.06.001
52. Kochetkova TV, Podosokorskaya OA, Elcheninov AG, Kublanov IV. Diversity of Thermophilic Prokaryotes Inhabiting Russian Natural Hot Springs. Microbiology. 2022;91(1):1–27. https://doi.org/10.1134/s0026261722010064
53. Garcia-Lopez E, Ruiz-Blas F, Sanchez-Casanova S, Peña Perez S, Martin-Cerezo ML, Cid C. Microbial communities in volcanic glacier ecosystems. Frontiers in Microbiology. 2022;13.
54. Urbanek AK, Mirończuk AM, García-Martín A, Saborido A, de la Mata I, Arroyo M. Biochemical properties and biotechnological applications of microbial enzymes involved in the degradation of polyester-type plastics. Biochim Biophys Acta Proteins Proteom. 2020;1868(2):140315. https://doi.org/10.1016/j.bbapap.2019.140315 PMID: 31740410
55. Kawai F, Nakadai K, Nishioka E, Nakajima H, Ohara H, Masaki K. Different enantioselectivity of two types of poly(lactic acid) depolymerases toward poly(l-lactic acid) and poly(d-lactic acid). Polym Degrad Stab. 2011;96(7).
56. Tripathi LP, Sowdhamini R. Genome-wide survey of prokaryotic serine proteases: analysis of distribution and domain architectures of five serine protease families in prokaryotes. BMC Genomics. 2008;9:549. https://doi.org/10.1186/1471-2164-9-549 PMID: 19019219
57. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480–9.
58. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research. 2005;33(suppl_1):D501–4. https://doi.org/10.1093/nar/gki025

59. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28(1):27–30.
60. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2.
61. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000;302(1):205–17.
62. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
63. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674–6. https://doi.org/10.1093/bioinformatics/btv033 PMID: 25609793
64. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7. https://doi.org/10.1093/nar/gkh340 PMID: 15034147
65. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.
66. Le S, Gascuel O. An improved general amino acid replacement matrix. Molecular Biology and Evolution. 2008;25(7):1307–20. https://doi.org/10.1093/molbev/msn067
67. Kumar A, Lakhawat SS, Singh K, Kumar V, Verma KS, Dwivedi UK. Metagenomic analysis of soil from landfill site reveals a diverse microbial community involved in plastic degradation. J Hazard Mater. 2024;480:135804.
68. Jahanshahi DA, Ariaeenejad S, Kavousi K. A metagenomic catalog for exploring the plastizymes landscape covering taxa, genes, and proteins. Sci Rep. 2023;13(1):16029. https://doi.org/10.1038/s41598-023-43042-9 PMID: 37749380
69. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 2011;7:539.
70. Fitch WM. An improved method of testing for evolutionary homology. J Mol Biol. 1966;16(1):9–16. https://doi.org/10.1016/s0022-2836(66)80258-9 PMID: 5917736
71. Charif D, Lobry JR. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. Biological and Medical Physics, Biomedical Engineering. Springer Berlin Heidelberg. 2007:207–32. https://doi.org/10.1007/978-3-540-35306-5_10
72. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20(2):289–90. https://doi.org/10.1093/bioinformatics/btg412 PMID: 14734327
73. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5(3):e9490. https://doi.org/10.1371/journal.pone.0009490