scientific reports

Check for updates

OPEN

# MiGPC: a comprehensive catalog of enzybiotics from environmental metagenomes

Donya Afshar Jahanshahi[2], Arad Ariaeenejad[2,3], Arman Hasannejad[1,3], Mohammad Reza Zabihi[2], Mohammad Reza Ghaffari[1], Shohreh Ariaeenejad[1]✉ & Kaveh Kavousi[2]✉

**Antimicrobial agents play a vital role in human and environmental health, with applications spanning medicine, food preservation, agriculture, and biotechnology. Among them, enzybiotics enzyme-based antimicrobials have emerged as powerful alternatives to conventional antibiotics due to their targeted mechanisms and lower propensity for resistance. Beyond their medical relevance, enzybiotics have emerging applications in food preservation, animal health, and agriculture, thereby broadening their industrial and environmental value. To support the discovery and characterization of these versatile biomolecules, we present the first genome-resolved metagenomic gene and protein targeted enzybiotic catalog focused on enzybiotics, derived from diverse environmental microbiomes. The Microbial Enzybiotic Gene and Protein Catalog (MiGPC), integrates 15 whole-metagenome datasets from oceans, soils, fecal samples, vegetation, and plastic-contaminated environments, capturing a wide ecological spectrum. Enzybiotic sequences were compiled through a hybrid strategy combining public database mining and manual literature curation, yielding over 136,000 enzybiotic sequences, 7654 metagenome-assembled genomes (MAGs), and ~100 million unique genes and proteins. MiGPC integrates taxonomic and enzybiotic gene profiles, offering a robust platform for the discovery, annotation, and ecological mapping of antimicrobial enzymes. Functional analyses using KEGG and eggNOG revealed that approximately 62% of the genes remained uncharacterized, highlighting a rich source of potentially novel functions. Glycoside hydrolases and glycosyl transferases were the most prevalent CAZyme families, while the dominant enzybiotic-producing taxa belonged primarily to the *Pseudomonadota* and *Bacillota* phyla. Statistical modeling uncovered two major ecological clusters that distinguished polluted from relatively pristine environments. MiGPC enables high-throughput screening of previously unexplored metagenomes, facilitating the identification of novel antimicrobial agents from under characterized ecosystems. Overall, MiGPC represents a landmark resource that will support multi-omics research, microbial ecology, and the development of next-generation biotechnological solutions based on enzybiotics.**

**Keywords** Enzybiotics, Antimicrobial enzymes, Metagenomic catalog, Genome-resolved metagenomics, Functional annotation, Microbial ecology

**Abbreviations**

| | |
|---|---|
| MiGPC | Microbial enzybiotic gene and protein catalog |
| MDR | Multidrug-resistant |
| MAGs | Metagenome assemble genome |
| GH | Glycoside hydrolases |
| AA | Auxiliary activity |
| CBM | Carbohydrate-binding modules |
| GT | Glycosyl transferase |
| CE | Carbohydrate esterase |

[1]Department of Systems and Synthetic Biology, Agricultural Biotechnology Research Institute of Iran (ABRII), Agricultural Research Education and Extension Organization (AREEO), Karaj, Iran. [2]Laboratory of Complex Biological Systems and Bioinformatics (CBB), Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran. [3]Arad Ariaeenejad and Arman Hasannejad contributed equally to this work. ✉email: sh.ariaee@abrii.ac.ir; shohreh.ariaeenejad@gmail.com; kkavousi@gmail.com; kkavousi@ut.ac.ir

| | |
|---|---|
| PL | Polysaccharide lyase |
| AMR | Antimicrobial resistance |
| PIGC | Pig integrated gene catalog |
| PDEC | Plastic-degrading enzyme catalog |
| MeSH | Medical subject headings |
| GTDB-tk | Genome taxonomy database toolkit |
| GMM | Gaussian mixture model |
| MeTarEnz | Metagenomic targeted enzyme miner |
| CAZy | Carbohydrate-active enzymes |
| PE | Polyethylene |
| PET | Polyethylene terephthalate |
| PS | Polyester |
| ROS | Reactive oxygen species |

Antimicrobial agents play a central role in safeguarding public health, food security, and ecosystem stability. While traditionally associated with the treatment of infectious diseases, these agents particularly enzyme-based antimicrobials such as enzybiotics are increasingly recognized for their diverse applications beyond clinical medicine, including food preservation, livestock protection, and agricultural sustainability[1,2,3,4]. These specialized enzymes, often derived from natural microbial sources, offer targeted mechanisms of action that minimize collateral damage to beneficial microbes and reduce the risk of resistance development in the host.

The widespread use and, in some cases, misuse of conventional antibiotics have historically contributed to unintended consequences such as antimicrobial resistance (AMR), which continues to compromise the effectiveness of many therapeutic interventions[5,6,7]. However, this challenge also underscores the urgent need for innovative and ecologically grounded antimicrobial strategies with broad applicability across human, animal, and environmental systems[8].

In this context, exploring microbial ecosystems, especially through metagenomic approaches, opens new avenues for identifying novel bioactive compounds and enzymes with multifunctional roles. Environmental microbiomes, shaped by both natural and anthropogenic factors, harbor immense genetic potential for antimicrobial activity, providing valuable insights into microbial interactions, ecological defense mechanisms, and evolutionary adaptations. Such discoveries not only inform clinical innovation but also support sustainable solutions for food preservation, environmental remediation, and agricultural biocontrol. Recently, antimicrobial enzymes, especially enzybiotics, have attracted increasing attention as promising alternatives to traditional antibiotics[6,9]. These lytic enzymes, often derived from bacteriophages or other natural sources, selectively degrade bacterial cell walls, resulting in rapid and specific bacterial lysis[9,10,11]. Unlike broad-spectrum antibiotics, enzybiotics have a lower risk of inducing resistance owing to their targeted mechanism of action.

Originally introduced by Nelson et al. (2001), the term *enzybiotic* referred to phage-derived enzymes, but it has since expanded to include various antimicrobial proteins, such as lysins, bacteriocins, lysozymes, and biofilm-disrupting enzymes like DNase I, amylases, and proteinases[12,13,14, 15,16,17,18], and interfering with quorum sensing pathways[19,20].

Despite their potential, enzybiotics face several limitations that hinder their clinical application, including poor solubility, limited structural stability, and potential immunogenicity. To overcome these challenges, research is focusing on enzyme engineering, such as domain reshuffling and protein fusion, as well as innovative delivery systems, such as nanozyme-based antibacterials (nanozybiotics) to improve their pharmacological properties.

Although many individual enzybiotics have been characterized, a systematic assessment of their genetic diversity and environmental distribution remains lacking. This gap in knowledge limits our ability to understand their natural reservoirs[21] and hampers the discovery of novel candidates. Metagenomics offers a powerful approach for accessing uncultured microbial diversity, unlocking new reservoirs of bioactive enzymes with therapeutic and industrial potential[22].

The ecological context of AMR provides essential insights into how resistance mechanisms originate, persist, and spread across different environments. Microbiomes exposed to antibiotic stressors, such as those in clinical, agricultural, or polluted environments can help identify resistance pathways and potential interventions, including antimicrobial peptides, CRISPR-based antimicrobials, and microbiota engineering[23]. However, antibiotic treatment can disrupt microbial community structures, which may inadvertently promote resistance gene transfer[24]. Studies of resistomes across environments such as the gut, soil, and atmosphere have revealed the widespread, dynamic, and multifactorial nature of AMR dissemination[25,26].

Metagenomic reference catalogs are critical resources for enhancing the resolution, reproducibility, and comparability of microbiome studies. By compiling non-redundant gene and genome datasets from specific ecosystems, these catalogs support functional profiling, taxonomic classification, and the discovery of novel enzymes and lineages. Landmark examples include the Human Gut Gene Catalog (> 10 million genes)[27], Cold Seep Gene Catalog (147 million genes, 3,164 MAGs)[28,29], and several host-specific catalogs such as the iMGMC (mouse gut), Chicken Gut Gene Catalog (16.6 million genes, 12,339 genomes)[30], and Pig Gut Gene Catalog (PIGC) (17.2 million genes, 6,339 MAGs)[31].

Despite these advances, a dedicated catalog of enzyme-associated metagenomes is lacking. A notable exception is the Plastic-Degrading Enzyme Catalog (PDEC), which highlights how targeted catalogs can identify valuable biocatalysts for environmental and industrial applications[32]. To address this gap, we introduce the Microbial Enzybiotic Gene and Protein Catalog (MiGPC) the first genome-resolved, integrated reference targeted enzybiotic catalog focused on enzybiotics. Built from 15 diverse metagenomic datasets including ocean water, soils, gut microbiomes, vegetation, and plastic-contaminated environments this catalog was constructed using a hybrid approach that combined public database mining (e.g., NCBI, BRENDA) and manual literature curation,

resulting in over 136,000 antimicrobial enzyme sequences (both predicted and experimentally validated). These sequences were assigned environmental contexts through BLAST-based mapping of metagenomic assemblies. MiGPC integrates taxonomic and enzybiotic gene catalogs, enabling the comprehensive profiling of antimicrobial agents across ecosystems. Statistical modeling and machine learning analyses revealed two distinct ecological clusters based on the co-abundance of enzybiotic-producing genera. Altogether, MiGPC offers a valuable resource for investigating the diversity, function, and ecological distribution of enzybiotics, supporting the development of novel antimicrobial strategies to combat the global antibiotic resistance crisis.

## Materials and methods

### Metagenomic sample collection and sources

*Enzybiotic dataset*

To construct a robust enzybiotic reference dataset, we employed two complementary strategies that integrated database mining and literature curation (Fig. 1).

In the first approach, we accessed the protein sequence section of major public repositories, including NCBI, UniProt, and BRENDA. Keywords such as *"Antibacterial enzyme"*, *"Antimicrobial enzyme"*, *"Antibacterial laccase"*, *"Endolysin"*, *"Enzybiotic"*, and *"Antibacterial oxidase"* were used to search for sequences with confirmed or predicted antimicrobial activity. In cases where keyword searches yielded a large number of non-specific hits (e.g., with "endolysin"), additional filters such as restricting by organism (bacterial sources), evidence level, or annotation tags were applied to improve specificity. Only sequences with antimicrobial properties, supported by experimental evidence or curated annotations, were retained. All selected sequences were downloaded in FASTA format and screened for redundancy and completeness.

In the second approach, we first identified antimicrobial enzymes from peer-reviewed publications indexed in databases such as PubMed, Google Scholar, and Scopus. Search terms included combinations of "enzybiotic", "metagenomics", "antibacterial enzyme", "inhibition of bacterial growth", and "novel enzyme" (see Fig. 1). After retrieving the articles, we manually curated the reported enzyme names and their corresponding accession numbers. These identifiers were used to retrieve precise sequence data from the NCBI protein database. This approach ensured high confidence in the antimicrobial activity of the selected enzymes by focusing on experimentally validated entries.

To improve recall and ensure accuracy in both methods, we also utilized controlled vocabularies and indexing tools such as Medical Subject Headings (MeSH), which allowed more precise filtering in literature and database searches.

Ultimately, this process led to the identification of 19 antimicrobial enzymes (Table 1), spanning two major enzyme classes: hydrolases and oxidoreductases. These enzymes were further categorized into functional subgroups (e.g., amidases, proteases, lipases, and laccases) and linked to their respective CAZy families, when applicable. Notable examples include N-acetylmuramoyl-L-alanine amidase, zoocin A, PersiLac1, and marinocine, which are widely recognized for their antimicrobial activity and structural uniqueness (see Table 1 for full list, enzyme classes, and references). To ensure high functional and biological relevance, the selection of these 19 antimicrobial enzymes was based on specific functional and sequence-based criteria:

1. Functional Criteria:

- Experimentally validated antimicrobial activity: Enzymes were included if they had experimentally proven antimicrobial activity, as documented in peer-reviewed publications. These activities include cell lysis, inhibition of bacterial growth, or biofilm disruption.

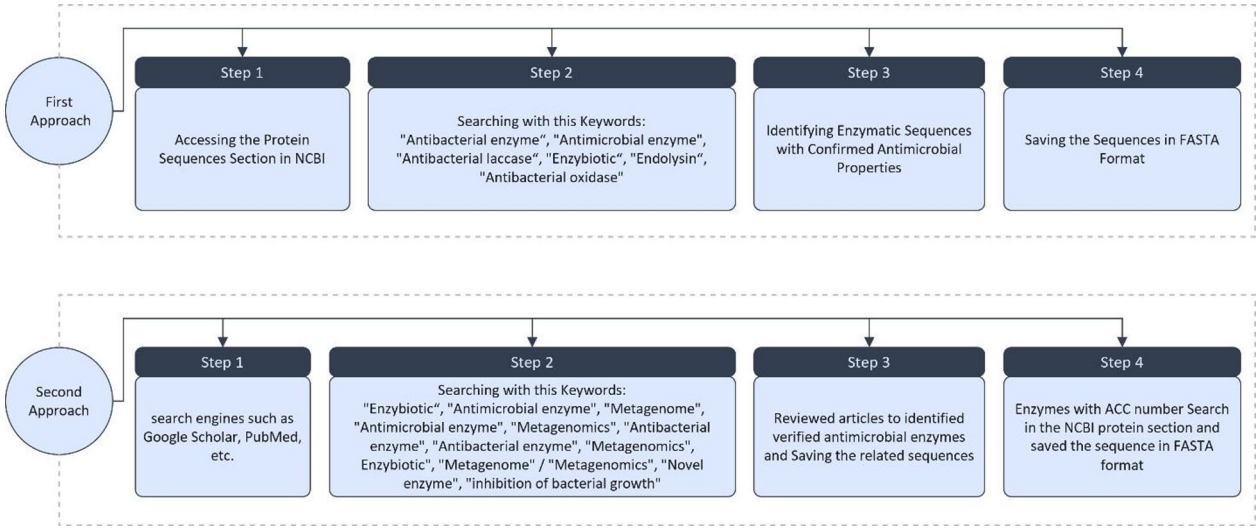


**Fig. 1**. Schematic representation of the two complementary strategies used to retrieve enzybiotic sequences from databases and literature.

| Name of Enzymes | Enzyme Class | Sub Classes | Cazy Family | Reference |
|---|---|---|---|---|
| lytic transglycosylase | Hydrolase | LTs | Glycoside Hydrolase | [33] |
| N-acetylmuramoyl-L-alanine amidase | Hydrolase | Amidases | Carbohydrate Esterases | 33 |
| N-acetylmuramoyl-L-alanine amidase | Hydrolase | Amidases | Carbohydrate Esterases | [34] |
| metallopeptidase | Hydrolase | Proteases | Auxiliary Activities | [35] |
| peptidase | Hydrolase | Proteases | Auxiliary Activities | [33] |
| PersiProtease1 | Hydrolase | Proteases | Auxiliary Activities | [36] |
| lipolytic enzyme | Hydrolase | Lipases | Auxiliary Activities | [33] |
| peptidase | Hydrolase | M20 Family Peptidase | Auxiliary Activities | [33] |
| Persilac1 | Oxidoreductases | Multy-copper oxidase | Auxiliary Activities | [37] |
| Persilac2 | Oxidoreductases | Multy-copper oxidase | Auxiliary Activities | [38] |
| Persilac3 | Oxidoreductases | Multy-copper oxidase | Auxiliary Activities | [39] |
| FAD-dependent L-amino acid oxidase | Oxidoreductases | Oxidase | – | [19] |
| autolytic | Oxidoreductases | Oxidase | – | [40] |
| antibacterial protein | Oxidoreductases | Oxidase | – | [41] |
| marinocine antimicrobial protein | Oxidoreductases | Oxidase | – | [42] |
| elastinolytic metalloprotease | Hydrolase | Proteases | Auxiliary Activities | [43] |
| lysostaphin | Hydrolase | Proteases | Auxiliary Activities | [44] |
| N-acetylmuramoyl-L-alanine amidase | Hydrolase | Amidases | Auxiliary Activities | [45] |
| zoocin A endopeptidase | Hydrolase | Proteases | Auxiliary Activities | [46] |

**Table 1**. List of major antimicrobial enzymes identified in the present study, categorized by enzyme class, subclass, and associated CAZy family "Note: The three entries labeled as N-acetylmuramoyl-L-alanine amidase refer to distinct antimicrobial enzymes originating from different organisms and supported by independent references."

- Curated functional annotations: Enzymes with functional annotations in authoritative databases (e.g., UniProt, NCBI Protein, BRENDA) were selected if they explicitly described antimicrobial, bacteriolytic, or bactericidal activities, supported by peer-reviewed evidence.

2. Sequence-Based Criteria:

- Exact match to enzyme names: Only sequences that matched a precisely defined enzyme name and accession number in the literature were retained.
- Full-length, non-fragmented sequences: Only complete, full-length sequences were included, ensuring the integrity of the data.
- Redundancy removal: To avoid duplicates, redundant sequences were removed using CD-HIT clustering with a 90% identity threshold.
- Multi-database consistency: Sequences were retained only if functional annotations were consistent across at least two major databases (e.g., NCBI and UniProt).
- Prioritizing high-confidence entries: In cases with multiple sequence variants, Swiss-Prot or RefSeq entries were given preference due to their high annotation reliability.

This systematic curation process led to the selection of 19 high-confidence enzybiotics that represent diverse functional classes of antimicrobial enzymes. These enzymes provide a reliable reference for downstream analyses and sequence alignment.

## Metagenomic samples collection

In metagenomic studies, selecting ecologically diverse and functionally relevant environments is essential to maximize the discovery of novel antimicrobial genes and enzymes[47].The primary aim of this study was to construct a genome-resolved reference catalog of microbial genes, proteins, taxa, and enzybiotics. To achieve this, environments likely to harbor enzybiotic-producing microorganisms were prioritized. The selection process began with a sequence similarity search: curated antimicrobial enzyme sequences were aligned against the NCBI protein database using BLAST. This analysis yielded a list of bacterial taxa exhibiting high homology with known enzybiotic proteins. Ten bacterial classes with strong sequence similarities were identified as potential antimicrobial enzymes producers.

To determine the natural habitats of these bacterial taxa, two complementary strategies were applied:

1. A systematic literature review of previously published metagenomic studies to identify environments associated with the identified bacterial classes.
2. Targeted taxonomic queries using the NCBI Taxonomy Browser to cross-reference bacterial classifications with environmental metadata.

The second approach (NCBI Taxonomy) proved particularly effective in associating taxa with specific ecosystems, such as marine sediments, coral reef microbiomes, brine pools, contaminated soils, estuaries, and extreme environments, offering high-resolution insights into their ecological distributions.

By integrating findings from both strategies and eliminating redundant entries, we assembled a preliminary list of 38 ecologically diverse environments with high enzybiotic potential. From this list, 15 representative whole-metagenome sequencing datasets were selected based on the following criteria:

- Public availability and raw FASTQ accessibility.
- High-quality sequencing data (Phred score ≥ 30 after trimming, read length ≥ 100 bp, and no evidence of substantial contamination based on quality control analysis).
- Ecological distinctiveness (e.g., polluted vs. pristine, aquatic vs. terrestrial).

The final selection spanned a wide range of biomes, including plastic-contaminated terrestrial and aquatic environments (polyethylene [PE], polyethylene terephthalate [PET], and polyester [PS]), hot springs, river estuaries, alpine vegetation, human gut microbiomes, brine pools, and groundwater systems.

Table 2 summarizes the selected metagenomic datasets, along with their corresponding SRA accession numbers and project IDs. These datasets were downloaded from public repositories such as NCBI SRA and MG-RAST, and subsequently used for gene prediction, taxonomic profiling, and enzybiotic identification.

## Gene prediction and functional annotation

Processing raw metagenomic data began with quality control using FastQC v0.11.8 with the following command: fastqc input_file.fastq to assess base quality and sequence composition. Adapter sequences and low-quality reads (Phred score < 20) were trimmed using Trimmomatic v0.39 with the following parameters: -phred33 SLIDINGWINDOW:4:20 MINLEN:36, followed by additional filtering with SOAPaligner2 with the following command: -a -v 5 -p 4 -i input_file.fq -o output_file.sam to eliminate residual low-quality or ambiguous reads[48]. High-quality reads were then assembled into contigs using MEGAHIT v1.2.9 [49] with the following parameters: --kmin-1pass, -m 60e + 10, --k-min 27, --k-max 127, --k-step 10, --min-contig-len 300, and -t 40 To estimate contig coverage and support downstream analyses, reads were mapped back to assembled contigs using BWA[50] v0.7.17 with the following parameters: mem -t 8 -M -R, and alignment files were processed using Samtools v1.11 to produce BAM-formatted outputs. Gene prediction was performed on assembled contigs using MetaGeneMark[51], identifying open reading frames (ORFs) and putative protein-coding sequences. Gene and protein prediction was performed using MetaGeneMark with the following command:/gmhmmp -a -m MetaGeneMark_v1.mod input.fa -A prot.fa -o output.fa. This parameter set was used for gene identification and annotation in metagenomic samples. To build a non-redundant gene and protein catalog, predicted gene sequences were clustered using CD-HIT v4.8.1 at a 90% sequence identity threshold ) -c 0.9, -M 0, and -T 0(. Individual catalogs from each metagenome were merged and re-clustered to generate a unified gene and protein catalog across all samples[52,53].

For functional annotation, two complementary tools were employed:

- KofamKOALA, for assignment of KEGG Orthology (KO) numbers based on HMM profiles;
- eggNOG-mapper v2.0.1 with the following command: eggNOG-mapper -i input.fasta -o output_directory --cpu 8 --output_format tsv, for classification into Clusters of Orthologous Groups (COGs) and broader functional categories.

| Environment Sample | SRR Number | Project Number |
|---|---|---|
| Seawater metagenome | SRR6441339 | PRJNA428417 |
| polyethylene terephthalate (PET) contaminated Cow Gut metagenome | SRR28167100 | PRJNA1081682 |
| River Estuary metagenome | SRR26623185 | PRJNA1024631 |
| Vegetation of Alpine bogs metagenome | ERR4298344 | PRJEB39100 |
| Hot Spring metagenome | SRR15826900 | PRJNA761511 |
| Rimov Reservoir metagenome | ERR9631090 | PRJEB52406 |
| polyester (PS) contaminated Seawater metagenome | SRR28167096 | PRJNA1081682 |
| polyethylene (PE) contaminated Soil metagenome | SRR28167104 | PRJNA1081682 |
| Groundwater metagenome | SRR25490518 | PRJNA1001268 |
| Brine metagenome | SRR26901926 | PRJNA1042791 |
| Fecal metagenome of human | SRR13060942 | PRJNA678426 |
| Plastic contaminated soil metagenome (Varamin) | SRR23085642 | PRJNA924045 |
| Tannery Wastewater metagenome (Charmshahr) | SRR12059165 | PRJNA640737 |
| Ocean surface- Oyster farm metagenome | SRR29090698 | PRJNA1113386 |
| A previously published catalog of plastic-contaminated environmental metagenomes | – | - |

**Table 2**. Selected metagenomic samples representing diverse environments with enzybiotic-producing potential.

To identify genes encoding carbohydrate-active enzymes (CAZymes), a unified gene catalog was screened using run_dbCAN2, which integrates multiple domain detection algorithms (HMMER, DIAMOND, Hotpep) to assign CAZy families with high confidence.

Finally, MetaGeneMark was re-applied during the genome binning stage to annotate protein-coding genes in metagenome-assembled genomes (MAGs), ensuring consistency across gene-level and genome-level annotations.

The overall computational pipeline for processing, gene prediction, functional annotation, and catalog integration is summarized in Fig. 2.

## CAZyme analysis

To investigate the functional potential of microbial communities in carbohydrate metabolism and enzybiotic activity, carbohydrate-active enzyme (CAZyme) genes were identified across the metagenomic datasets using the standalone version of run_dbCAN2[54]. identified across the metagenomic datasets using the standalone version of run_dbCAN2. This tool integrates three complementary annotation algorithms, HMMER, DIAMOND, and Hotpep for the accurate detection of CAZyme domains based on the CAZy database classification criteria. The final set of MAGs was used as the input for this analysis[55]. The input for this analysis was a non-redundant gene catalog derived from all 15 metagenomic environments. Only CAZyme genes confirmed by at least two of the three tools (default run_dbCAN2 filtering criteria) were retained to enhance specificity.

Identified CAZyme genes were categorized into six major families:

- Glycoside Hydrolases (GH).
- Glycosyl Transferases (GT).
- Carbohydrate Esterases (CE).
- Polysaccharide Lyases (PL).
- Auxiliary Activities (AA).
- Carbohydrate-Binding Modules (CBM).

This classification enabled a comparative analysis of CAZyme distribution, relative abundance, and functional diversity across the different ecosystems. Special emphasis was placed on CAZyme families known to participate in antimicrobial mechanisms, such as cell wall degradation and biofilm disruption, in order to assess their potential roles in enzybiotic activity.

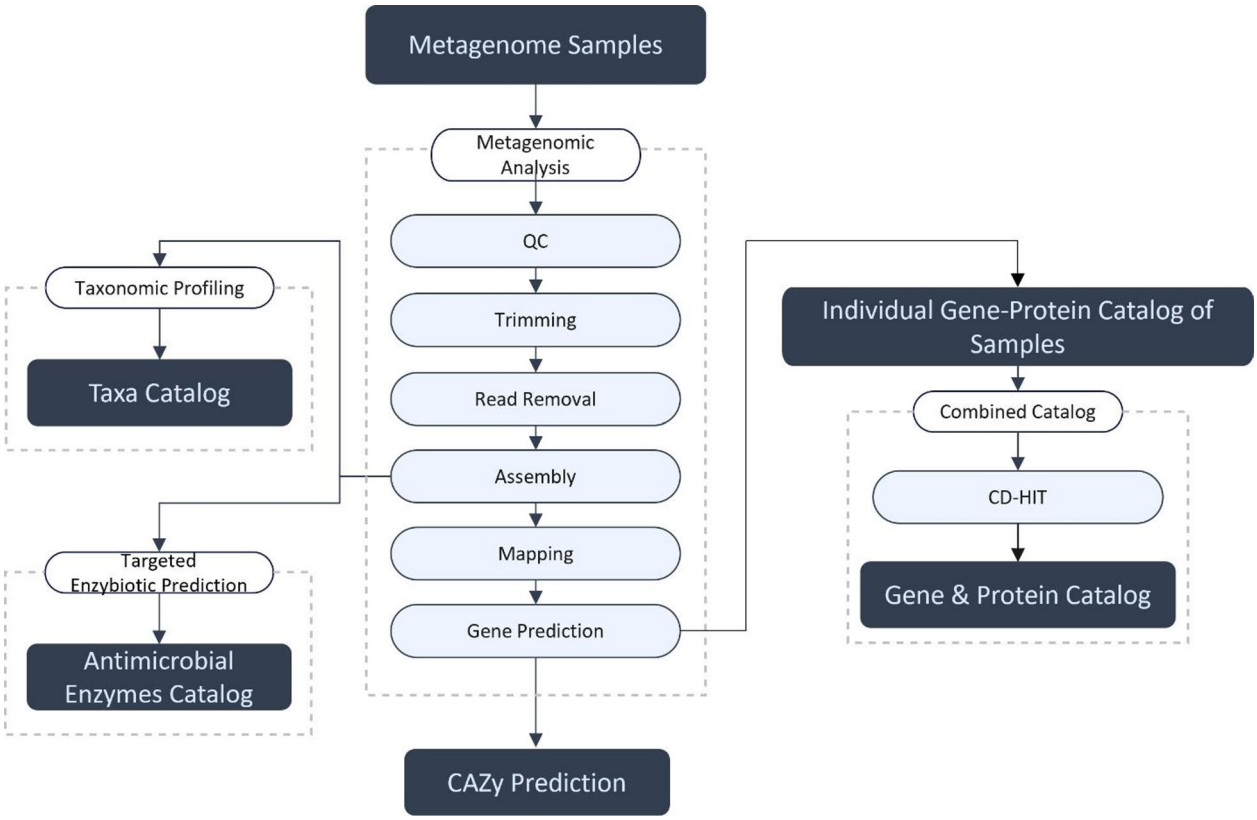


**Fig. 2**. Computational pipeline for gene prediction, functional annotation, and catalog integration of metagenomic samples.

## Taxonomic Profiling and Genome Reconstruction

To investigate the diversity and structure of microbial communities, raw metagenomic reads were assembled and used to reconstruct microbial genomes through binning. Genome binning was performed using MetaBAT2 [56], with the parameters --minContigLength 2000 and --minContigDepth 2 [37] to enhance bin quality. Redundant genome bins were subsequently dereplicated using dRep v2.2.3 with the following parameters: -p 8 -sa 0.95, ensuring a non-redundant set of metagenome-assembled genomes (MAGs). Genome quality was evaluated using CheckM v1.0.13 [57], and only MAGs with ≥ 65% completeness and ≤ 10% contamination were retained for downstream analyses.

## Statistical modeling and clustering

To examine how enzybiotic-producing bacterial genera correlate with environmental types, we analyzed taxonomic profiles obtained through NCBI BLAST and eggNOG-mapper. Specifically, we focused on the relative abundance of 14 enzybiotic-potential bacterial genera (*Bradyrhizobium*, *Streptomyces*, *Streptococcus*, *Enterococcus*, *Pseudomonas*, *Burkholderia*, *Lysobacter*, *Clostridium*, *Myxococcus*, *Phaeobacter*, *Rheinheimera*, *Marinomonas*, *Pseudoalteromonas*, *Staphylococcus)* across all 15 metagenomic environments[58,59]. The selection of the 14 enzybiotic-potential genera was based on (i) documented relevance to antibiotic resistance dissemination in clinical and environmental literature, (ii) evidence of involvement in enzyme-mediated antimicrobial interactions, and (iii) broad representation across the metagenomic datasets analyzed. These genera therefore serve as ecologically and functionally meaningful taxa for assessing the distribution of enzybiotic homologs across environments.

To enable meaningful comparison across samples, genus abundance data were standardized using Z-score normalization, calculated as:

$$X^{\wedge\prime} = (X - \mu)/\sigma \tag{1}$$

where X is the raw abundance, μ is the mean, and σ is the standard deviation for each genus across environments.

A hierarchical clustered heatmap was constructed using the 'pheatmap' package in R, allowing visualization of co-abundance patterns and environmental grouping based on microbial profiles.

To statistically test differences in bacterial abundance between identified environmental clusters, we employed a linear regression model of the form:

$$\mathrm{Y} = \mathrm{X}\beta \tag{2}$$

Here, Y denotes the Z-score–normalized abundance for a given bacterial genus, and X is a binary design matrix indicating cluster membership (0 or 1). The regression coefficient β quantifies the association between bacterial abundance and cluster assignment. A p-value threshold of $< 0.05$ was used to determine statistical significance.

To evaluate the ability of bacterial genera to classify environmental samples into distinct groups, we trained a logistic regression model using the parameters: C = 0.01, penalty = 'l2', and solver = 'liblinear'. This model provided insight into the predictive power of genus-level features for environmental classification.

In parallel, we applied a Gaussian Mixture Model (GMM) to assess the underlying structure of the data, assuming that observations were drawn from a mixture of two Gaussian distributions. The likelihood function of the GMM is expressed as:

$$P(X|\theta) = \sum_{k=1}^{K} \pi_k . \mathcal{N}(X|\mu_k, \Sigma_k) \tag{3}$$

Where $\pi_k$ is the mixing coefficient for the $k$-th Gaussian component, constrained such that.

$\sum \pi_k = 1$. The function $\mathcal{N}(X|\mu_k, \Sigma_k)$ denotes the multivariate normal distribution with mean $\mu_k$ and covariance matrix $\Sigma_k$. For this study K = 2 was assumed.

Together, these statistical and machine learning models enabled the identification of environmental clusters and their microbial signatures, particularly those enriched in enzybiotic-producing taxa.

To explore ecological patterns associated with enzybiotic-producing microbial communities across diverse environments, statistical modeling and clustering analyses were performed using taxonomic abundance profiles. Specifically, the relative abundance of 14 bacterial genera (*Bradyrhizobium*, *Streptomyces*, *Streptococcus*, *Enterococcus*, *Pseudomonas*, *Burkholderia*, *Lysobacter*, *Clostridium*, *Myxococcus*, *Phaeobacter*, *Rheinheimera*, *Marinomonas*, *Pseudoalteromonas*, *Brevibacterium)* known for antimicrobial potential or antibiotic resistance was calculated for each of the 15 metagenomic samples.

To enable comparison across environments, abundance values were standardized using Z-score normalization, defined as:

$$Z = \frac{X - \mu}{\sigma} \tag{4}$$

where $X$ is the raw abundance of a given genus, $\mu$ is the mean abundance across all environments, and $\sigma$ is the corresponding standard deviation.

A hierarchical clustering heatmap was constructed using the pheatmap package in R, with Euclidean distance and complete linkage method. This analysis revealed two distinct ecological clusters of environments, based on the co-abundance patterns of antimicrobial-associated genera.

To statistically test differences in genus abundance between the two identified clusters, a linear regression model was applied for each genus as follows:

$$Y = \beta_0 + \beta_1 \cdot C + \epsilon \tag{5}$$

where $Y$ is the standardized abundance of the genus, $C$ is a binary variable representing cluster membership (Cluster 1 = 0, Cluster 2 = 1), $\beta_0$ is the intercept, $\beta_1$ represents the effect of cluster, and $\epsilon$ is the error term. A p-value < 0.05 for $\beta_1$ was considered statistically significant, indicating a differential abundance between clusters.

In addition to hypothesis testing, a logistic regression classifier was developed to evaluate whether genus-level features could predict environmental cluster membership. The model was trained using L2-regularized logistic regression with the following settings:

- Regularization strength: $C = 0.01$
- Penalty: L2.
- Solver: liblinear.

Performance metrics, including accuracy, precision, recall, and F1-score, were calculated via cross-validation. The model achieved perfect classification (accuracy = 1.00), highlighting the strong discriminatory power of taxonomic profiles.

To assess the latent structure of the dataset, we also implemented a Gaussian Mixture Model (GMM) using the expectation-maximization algorithm. The probability density function for a GMM with $K = 2$ components is given by:

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x \mid \mu_k, \Sigma_k) \tag{6}$$

where $\pi_k$ is the mixing coefficient for component $k$, constrained such that $\sum \pi_k = 1$; $\mu_k$ and $\Sigma_k$ are the mean vector and covariance matrix of the $k$-th Gaussian component; and $\mathcal{N}(x \mid \mu_k, \Sigma_k)$ denotes the multivariate normal distribution.

The GMM revealed two well-separated clusters, which was consistent with the hierarchical clustering and logistic regression results. Together, these statistical models highlight clear ecological distinctions among environments based on the distribution of enzybiotic-associated bacterial genera. This provides a framework for identifying microbial signatures linked to enzybiotic richness and facilitates the targeted screening of ecosystems with high biotechnological potential.

### Antimicrobial enzyme identification

To identify potential enzybiotic genes within the assembled contigs, we employed MeTarEnz (Metagenomic Targeted Enzyme Miner)[60,61] a targeted screening tool designed for high-throughput analysis of metagenomic datasets. MeTarEnz (with the following command:./metarenz cs -i input.fa -t 20 -db database.fa -o outpot_name) enables the use of user-defined reference databases and customizable bit-score thresholds to enhance specificity.

A curated database of 19 known enzybiotic enzyme sequences was used as the reference, and candidate genes were identified using a minimum bit-score cutoff of 100. Putative enzybiotic sequences were further analyzed using NCBI's Conserved Domain Database (CDD)[62] to verify functional domains. To assess structural similarity and predict protein conformations, we utilized AlphaFold3 for 3D structure prediction and TM-align for structural alignment and comparison[39,63].

## Results and discussion

### Gene and protein catalog statistics

To the best of our knowledge, this study presents the first comprehensive gene and protein targeted enzybiotic catalog (MiGPC) focused on enzybiotic-active environments, constructed through an integrated metagenomic workflow. The analysis incorporated 15 distinct environmental datasets, comprising approximately 100 million unique genes and proteins. The bioinformatics pipeline included steps such as quality control, read trimming, assembly, and contig binning. Applying this workflow, we identified approximately 100 million unique genes and proteins across the enzybiotic-active environments. This integrated strategy provides a robust foundation for further investigations into the functional and taxonomic diversity of these genes.

Raw metagenomic reads totaling approximately 1.2 terabytes were collected from environments such as ocean water, plastic-contaminated soil, and groundwater, and assembled using MEGAHIT. This process resulted in 55,284,777 contigs (length > 300 bp and without plastic catalog sample), with the longest contig measuring 600,358 bp in length. The compiled catalog includes both non-redundant genes and protein sequences, derived from the predicted coding regions. The final output represents the first integrated gene and protein targeted enzybiotic catalog dedicated to enzybiotic-active ecosystems. Table 3 summarizes the key characteristics of each individual gene catalog used in this study.

| Metagenome sample | Num of sequences | binned contigs | non redundant genes | min length (bp) | max length (bp) | average length (bp) | Acc Number |
|---|---|---|---|---|---|---|---|
| Charmshahr | 4,241,532 | 57,589 | 4,353,769 | 300 | 162,584 | 627 | SRR12059165 |
| Seawater | 7,287,549 | 442 | 5,555,423 | 200 | 47,088 | 495.9 | SRR6441339 |
| PET_Cow Gut | 1,144,946 | 12,488 | 1,632,568 | 300 | 212,414 | 591.1 | SRR28167100 |
| Varamin | 7,370,307 | 399,494 | 1,305,332 | 300 | 600,358 | 1411 | SRR23085642 |
| River Estuary | 1,939,501 | 123,339 | 3,046,857 | 300 | 84,957 | 1,083.1 | SRR26623185 |
| Vegetation of Alpine bogs | 3,020,755 | 3,851 | 3,556,119 | 300 | 56,946 | 522.9 | ERR4298344 |
| Rimov Reservoir | 7,179,926 | 24,083 | 8,801,145 | 300 | 225,203 | 537.1 | ERR9631090 |
| PS Seawater | 596,279 | 7,750 | 922,926 | 300 | 261,036 | 651.1 | SRR28167096 |
| PE Soil | 2,193,359 | 4,503 | 2,991,396 | 300 | 89,523 | 536.3 | SRR28167104 |
| Hot Spring | 854,026 | 8,425 | 980,659 | 300 | 547,154 | 596.4 | SRR15826900 |
| Ocean | 5,046,411 | 9,186 | 6,184,015 | 300 | 94,601 | 520.1 | SRR29090698 |
| Groundwater | 1,758,937 | 6,499 | 2,301,386 | 300 | 224,109 | 554.9 | SRR25490518 |
| Fecal | 2,633,054 | 17,236 | 3,300,294 | 300 | 414,855 | 546.1 | SRR13060942 |
| Brine | 10,018,195 | 4182 | 1,1842,302 | 300 | 173,742 | 517.5 | SRR26901926 |
| Plastic Catalog | – | – | 53,318,076 | – | – | – | – |

**Table 3**. Summary of key characteristics of individual enzybiotic gene catalogs.

## Functional annotation of genes

KEGG orthologs were assigned using KofamKOALA, resulting in the identification of ~3 million genes in the MiGPC. In parallel, eggNOG-mapper annotated ~35 million genes based on Clusters of Orthologous Groups (COG) functional categories. These two annotation pipelines were applied independently to the five individual datasets as well as the integrated MiGPC catalog. Figure. 3 A illustrates the number of genes annotated by eggNOG and KEGG across the 15 samples, while Figure. 3B compares their relative proportions in the MiGPC.

Overall, eggNOG consistently mapped more genes than KEGG across all environments, a finding aligned with previous studies such as the chicken gut gene catalog[32],[30]. The highest proportion of unannotated genes was observed in seawater and estuarine samples (~80%), followed by groundwater (~60%). In contrast, the fecal metagenome exhibited the lowest proportion of unmapped genes (~8%), highlighting variation in database coverage across environments.

## CAZyme profiling

The Carbohydrate-Active enZyme (CAZy) database served as a critical reference for profiling genes involved in carbohydrate metabolism. Despite the central role of enzybiotics in hydrolyzing the peptidoglycan layer of bacterial cell walls, the distribution of enzybiotic-associated CAZymes remains underexplored. To address this gap, we performed CAZyme annotation on the MiGPC using dbCAN2, which identified 4,993,963 CAZyme-encoding genes. Figures. 3 C and D show the distribution and relative abundance of CAZy families across all 15 metagenomic datasets.

Among the six major CAZy families Glycosyltransferases (GT), Glycoside Hydrolases (GH), Carbohydrate Esterases (CE), Auxiliary Activities (AA), Carbohydrate-Binding Modules (CBM), and Polysaccharide Lyases (PL) the GT and GH families were the most dominant, particularly at the integrated MiGPC level (Figure. 3C). These enzymes contribute to enzybiotic function by targeting and degrading structural polysaccharides in bacterial cell walls.

- GTs were the most abundant group overall.
- PLs were the least represented across samples.
- CE and AA ranked second and third in abundance, respectively.

Environmental trends were also observed:

- CE enzymes were notably enriched in Rimov and Estuary samples.
- AA enzymes showed the highest relative abundance in Rimov and Ocean datasets.
- Both CE and AA genes were least represented in the Groundwater dataset.

The relative abundances of all six CAZy families in MiGPC (Fig. 3C) underscore their potential roles in enzybiotic mechanisms and antimicrobial enzyme activity (Fig. 3D).

## Taxonomic distribution of MiEGPC

Given the limited number of experimentally validated microorganisms encoding antimicrobial enzymes in existing literature, this study aimed to identify microbial strains associated with enzybiotic genes within our curated dataset. To achieve this, we employed two complementary approaches: (1) sequence similarity search using NCBI BLAST, and (2) orthologous gene mapping via eggNOG-mapper. Taxonomic profiling revealed diverse microbial communities across the 15 metagenomic environments (Table 4), with the dominant phyla including *Pseudomonadota*, *Bacillota*, and *Actinomycetota*,

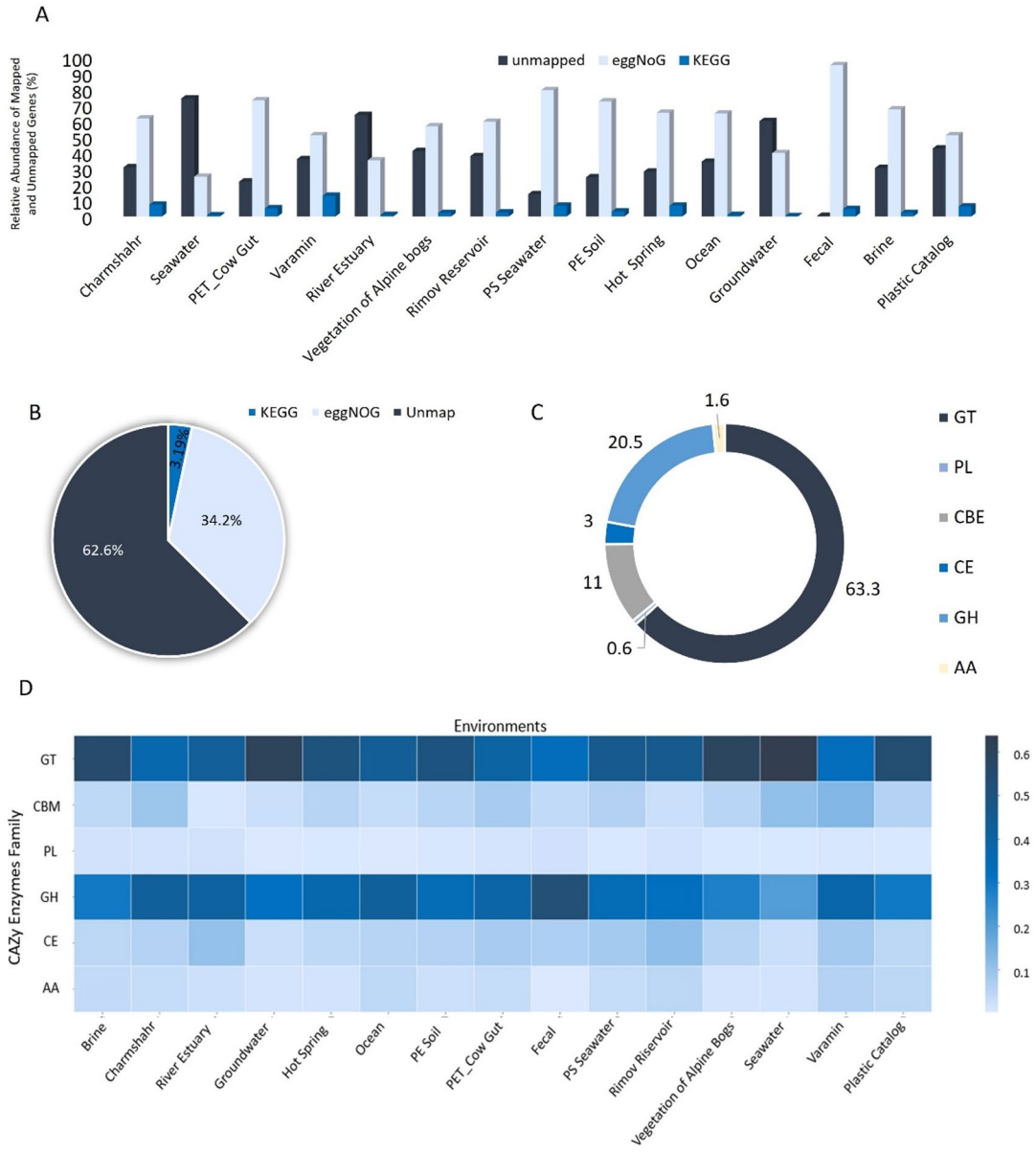


**Fig. 3**. Annotation statistics and CAZyme distribution across enzybiotic-active metagenomic samples. (**A**) Number of genes annotated using eggNOG and KEGG in individual gene catalogs and in the integrated MiGPC catalog. (**B**) Proportion of genes in the MiGPC assigned to KEGG (via KofamKOALA), eggNOG, or remaining unannotated. (**C**) Donut chart showing the relative abundance of six major CAZyme families (GH, GT, CE, AA, CBM, PL) identified in the MiGPC. (**D**) Heatmap illustrating the variation in CAZyme family abundance across the 15 metagenomic samples. Columns represent CAZyme groups, with a color gradient from blue to green indicating relative abundance levels. Several enzymes from the six major CAZyme families have been associated with antimicrobial functions. These CAZymes contribute to the degradation of structural polysaccharides in bacterial cell walls, compromising microbial integrity and enhancing enzybiotic activity. The identification of antimicrobial CAZymes is particularly relevant in enzybiotic-active environments, where microbial communities are subject to intense competition and selective pressure, often leading to the evolution of unique enzymatic strategies for survival. The high abundance of enzybiotic-associated CAZymes in these metagenomes highlights their potential as a source of novel antimicrobial agents, with promising applications in biotechnology and medicine. By functionally mapping CAZyme families to known antimicrobial roles, researchers have gained valuable insights into how these enzymes mediate microbial antagonism and ecological defense mechanisms, particularly in contaminated or competitive ecosystems.

To directly support enzybiotic discovery and ecological interpretation, we mapped predicted enzybiotic genes and enzybiotic-encoding MAGs onto taxonomic profiles constructed from BLAST and eggNOG-based assignments. Taxonomic profiling of the 15 metagenomic environments revealed recurrent phyla (notably *Pseudomonadota*, *Bacillota*, *Actinomycetota*, and *Bacteroidota*). However, rather than emphasizing broad

| Species name | Genus | Phylum |
|---|---|---|
| *Phaeobacter inhibens DSM 17,395* | *Phaeobacter* | *Pseudomonadota* |
| *Bradyrhizobium sp* | *Bradyrhizobium* | *Pseudomonadota* |
| *Enterococcus faecalis* | *Enterococcus* | *Bacillota* |
| *Lysobacter antibioticus* | *Lysobacter* | *Pseudomonadota* |
| *Clostridium drakei* | *Clostridium* | *Bacillota* |
| *Streptomyces scabiei* | *Streptomyces* | *Actinomycetota* |
| *Burkholderia sp* | *Burkholderia* | *Pseudomonadota* |
| *Myxococcus xanthus* | *Myxococcus* | *Myxococcota* |
| *Pseudoalteromonas rubra* | *Pseudoalteromonas* | *Pseudomonadota* |
| *Pseudoalteromonas tunicata* | *Pseudoalteromonas* | *Pseudomonadota* |
| *Rheinheimera sp.* | *Rheinheimera* | *Pseudomonadota* |
| *Marinomonas mediterranea* | *Marinomonas* | *Pseudomonadota* |
| *Pseudomonas aeruginosa* | *Pseudomonas* | *Pseudomonadota* |
| *Staphylococcus pettenkoferi* | *Staphylococcus* | *Bacillota* |
| *Streptococcus pneumoniae* | *Streptococcus* | *Bacillota* |
| *Streptococcus agalactiae* | *Streptococcus* | *Bacillota* |

**Table 4**. The main species that identified as Enzybiotic-encoding bacteria.

phylum-level composition alone, we focused on taxon–enzyme associations that illuminate enzybiotic reservoirs. The specific enzybiotic-related genera examined in our study are explicitly annotated within their corresponding phyla (Fig. 4A).

Using MeTarEnz screening of assembled contigs against our curated enzybiotic reference set, we identified 19,756 putative antimicrobial enzyme sequences. Mapping these sequences to taxonomic bins showed that a significant portion of the enzybiotics were associated with *Pseudomonadota* and *Bacillota* lineages phyla known to harbor genera such as *Pseudomonas* and *Bacillus*, which are widely reported for antimicrobial enzyme production.

Environmental trends also emerged: plastic-contaminated sites showed high enzybiotic gene diversity, often with phage-derived lytic enzymes, suggesting adaptation to competitive niches. In contrast, aquatic and vegetation-associated samples displayed fewer enzybiotic genes but occasionally with dominant gene copies, indicating potentially different ecological roles or reduced competitive pressure.

To prioritize bioprospecting targets, we recommend focusing on MAGs and taxa that co-localize with high-confidence enzybiotic genes, especially those from polluted environments, where ecological stress may favor the evolution of competitive antimicrobial strategies.

## Environmental clustering and regression models

To explore ecological patterns among environments harboring antimicrobial gene carriers, we analyzed the relative abundance of 14 enzybiotic-potential bacterial genera (*Bradyrhizobium*, *Streptomyces*, *Streptococcus*, *Enterococcus*, *Pseudomonas*, *Burkholderia*, *Lysobacter*, *Clostridium*, *Myxococcus*, *Phaeobacter*, *Rheinheimera*, *Marinomonas*, *Pseudoalteromonas*, *Staphylococcus)* across 15 metagenomic sample. The seawater metagenome sample was excluded from downstream classification analyses to avoid introducing bias into ecological clustering because it showed incomplete genus-level representation and lacked sufficient species count resolution for quantitative comparison across the 15 environments. Unsupervised clustering based on the co-abundance of three key genera *Bradyrhizobium*, *Streptomyces*, and *Streptococcus* revealed two distinct ecological clusters (Fig. 4B). Linear regression analysis further demonstrated that *Streptomyces* abundance significantly differed between the two clusters ($p < 0.05$; Table 5), indicating that ecological conditions selectively influence specific antimicrobial-producing taxa.

To assess the classification potential of these bacterial genera, we employed logistic regression and the Gaussian Mixture Model (GMM). The logistic regression model achieved perfect classification (*accuracy* = 1.00), with corresponding precision, recall, and F1-score values of 1.00 for both clusters, confirming robust model performance. The optimal parameters included $C = 0.01$, L2 regularization, and the 'liblinear' solver. The estimated coefficients were:

$$\beta_1\ (Bradyrhizobium):\ 0.0171$$

$$\beta_2\ (Streptomyces):\ 0.0480$$

$$\beta_3\ (Streptococcus):\ -0.0160$$

Intercept: 0.02399.

These values reflect the relative contribution of each genus in differentiating between clusters.

The GMM analysis corroborated these findings by identifying two distinct Gaussian components. The cluster means and covariance matrices exhibited minimal overlap, providing further evidence of a clear separation

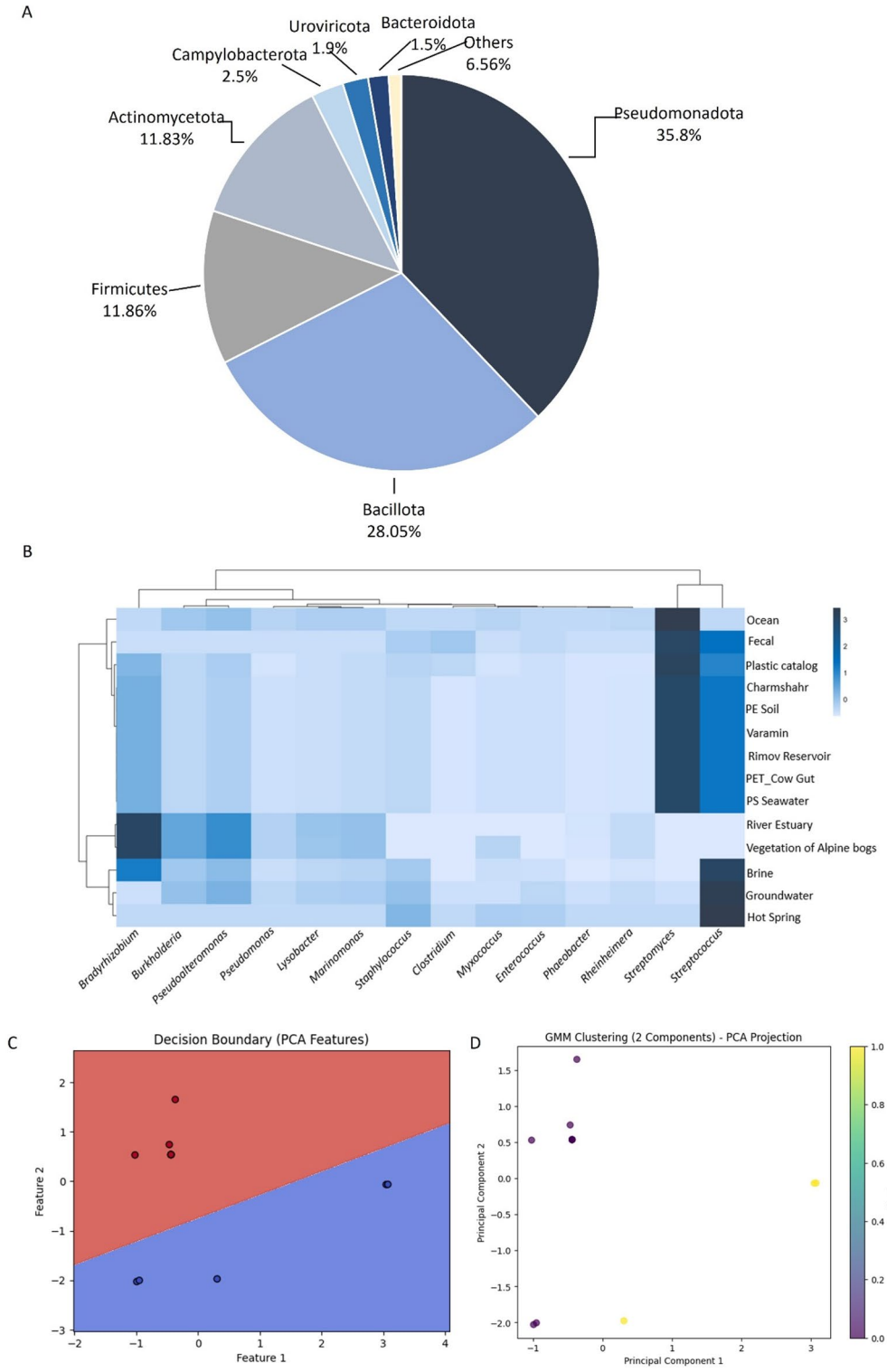


**Fig. 4**. Phylum-level taxonomic distribution and clustering of metagenomic environments based on the co-abundance of enzybiotic-potential bacterial genera. (**A**) Pie chart showing the relative abundance of bacterial phyla across the integrated MiGPC dataset. Dominant phyla include *Pseudomonadota*, *Bacillota*, and *Actinomycetota*. (**B**) Hierarchical clustering heatmap of 15 metagenomic environments based on the abundance of 14 enzybiotic-potential bacterial genera. Two major ecological clusters are evident, with *Bradyrhizobium*, *Streptomyces*, and Streptococcus contributing most strongly to the separation. (**C**) Decision boundary plot from logistic regression using PCA-transformed features. The red and blue regions represent the two predicted environmental clusters, indicating clear separation. (**D**) Gaussian Mixture Model (GMM) clustering of samples projected onto PCA space. Points are colored based on cluster membership probability, confirming the presence of two distinct clusters.

| | Bradyrhizobium | Streptomyces | Streptococcus |
|---|---|---|---|
| Class | −1.128* (0.581) | 3.594*** (0.078) | −0.620 (0.733) |
| Constant | 1.303** (0.466) | −0.526*** (0.063) | 1.735** (0.588) |
| Observations | 14 | 14 | 14 |
| $R^2$ | 0.239 | 0.994 | 0.056 |
| Adjusted $R^2$ | 0.176 | 0.994 | −0.022 |
| Residual Std. Error (df = 12) | 1.041 | 0.140 | 1.315 |
| F Statistic (df = 1; 12) | 3.775* | 2,108.082*** | 0.714 |

**Table 5**. Summary of linear regression models evaluating the association between environmental cluster (predictor) and the relative abundance of three enzybiotic-potential bacterial genera (Bradyrhizobium, Streptomyces, Streptococcus). Significance levels: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

(Fig. 4 C and D). Figure. 4 illustrates the taxonomic composition and clustering results derived from the co-abundance analyses.

The classification revealed two major ecological groups:

Cluster 1 included rivers, vegetation, brine, groundwater, and hot springs ecosystems with minimal anthropogenic disturbance and dominated by natural ecological and geochemical processes. While brine and hot springs are extreme (salinity, temperature, mineral load), rivers and vegetation zones are moderate and shaped by seasonal hydrology and soil variability. Despite differences, they all support structured, low-pollution microbial communities.

Cluster 2 encompassed ocean, fecal-contaminated sites, reference areas, Varamin, Rimov Reservoir, and environments contaminated with plastic pollutants (e.g., PET, PS, PE). These systems experience substantial human impact through organic and synthetic waste, including microplastics and fecal matter, which reshape microbial composition and increase ecological risk. The resulting microbiomes are typically stress-adapted and diverse.

These findings underline the ecological divergence of the two clusters and emphasize the discriminatory power of *Bradyrhizobium*, *Streptomyces*, and *Streptococcus* as key indicators of antimicrobial activity in different environmental contexts. While certain genera appeared predominantly in specific environments within our dataset, this pattern should not be interpreted as evidence that their dominance is solely driven by antimicrobial activity. Multiple ecological and physiological factors may contribute to the prevalence of a given taxon, including sporulation ability, metabolic flexibility, environmental resilience, and adaptation to nutrient- or stress-rich conditions. The enzybiotic profiles we report therefore reflect a combination of antimicrobial potential and broader ecological strategies, rather than a direct one to one relationship between taxonomic abundance and antimicrobial enzyme production.

The binary variable "Class" represents cluster membership (0 or 1), derived from unsupervised co-abundance analysis (Fig. 4B).

Each row reports the estimated regression coefficient (with standard error), intercept, $R^2$, adjusted $R^2$, residual standard error, and F-statistic. A $p$-value of $< 0.05$ indicated the statistical significance of the regression model. The standard errors are presented in parentheses below the coefficient estimates.

While both metagenomics and genomics offer valuable insights, we chose to focus on metagenomic data because of its ability to capture the genetic diversity of microbial communities in their natural, uncultured states. Unlike traditional genomic approaches that rely on reference genomes or cultured organisms, metagenomics allows for a more comprehensive understanding of microbial function and interaction in environmental contexts. This is particularly crucial for studying antimicrobial resistance and the functional potential of microbiomes, which may include unculturable species or novel genes that would otherwise remain undetected in conventional culture-based methods. By leveraging metagenomic data, we explored the full spectrum of microbial capabilities in diverse environments, avoiding strong cultivation bias and providing a broader ecological and functional perspective.

## Identification and classification of enzybiotics

To build a comprehensive antimicrobial enzyme catalog, contigs were screened against a curated enzybiotic database using the MeTarEnz tool[64]. Sequences were filtered based on high bit-scores (≥100) and low E-values. This approach identified a total of 19,756 putative antimicrobial enzyme sequences from the metagenomic datasets. The identified enzymes were classified into two major functional groups: hydrolases and oxidoreductases. Conserved domain analysis using NCBI-CDD, and 3D structure prediction via Alphafold and TM-align, confirmed their functional and structural similarities to known antimicrobial enzymes.

A phylogenetic tree constructed from 19 representative enzymes including 12 hydrolases and 7 oxidoreductases revealed distinct clades and functional divergence (Fig. 5A). Among hydrolases, enzymes such as M23 metallopeptidases (WP_191821694.1) and N-acetylmuramoyl-L-alanine amidases (AHZ45666.1) showed highly conserved catalytic residues, highlighting their essential roles in peptidoglycan degradation. In contrast, amidases and proteases from uncultured bacteria displayed greater divergence, suggesting adaptation to varied bacterial targets.

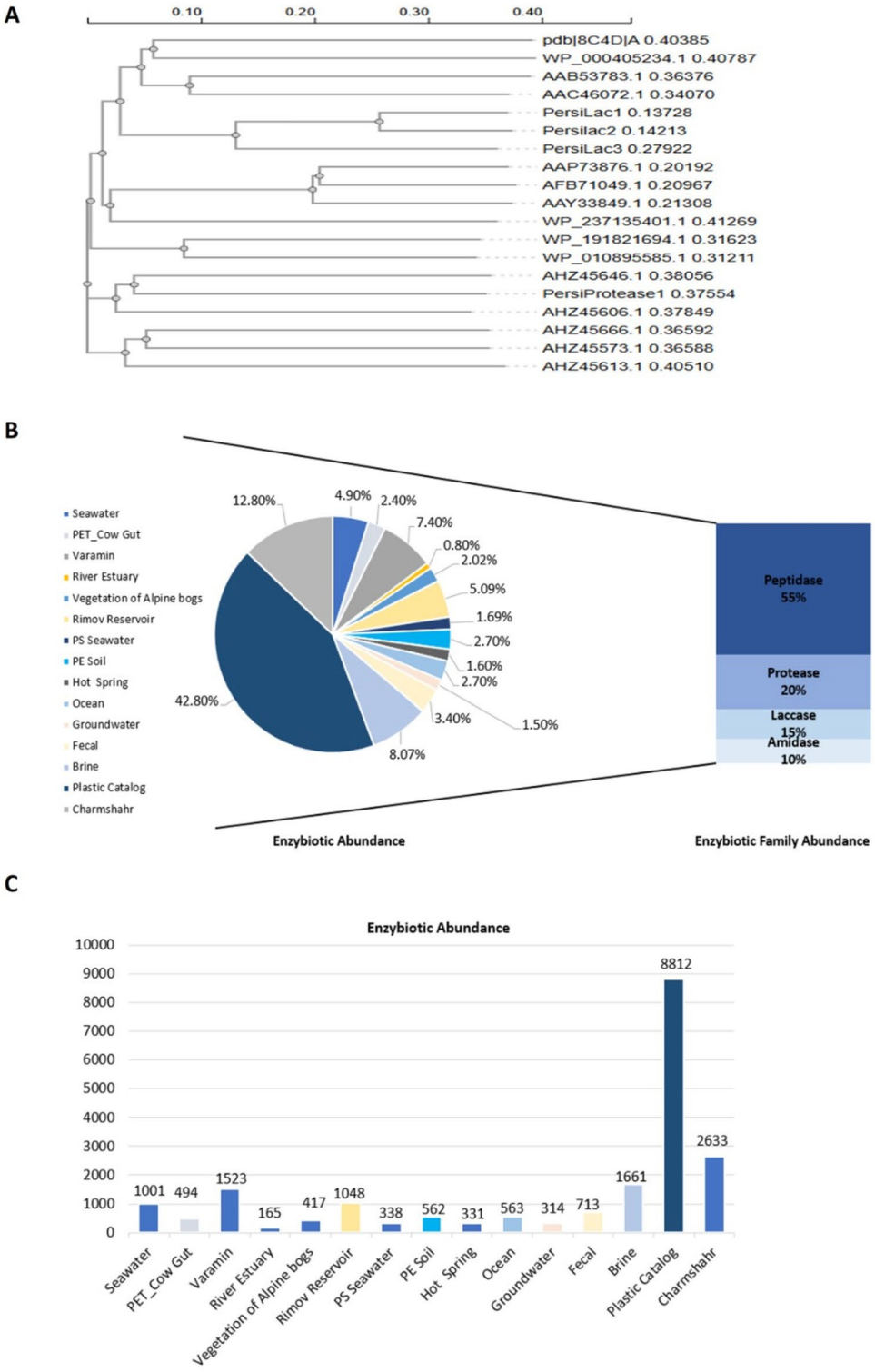


**Fig. 5**. Phylogenetic and ecological distribution of enzybiotic enzymes across environmental samples. (**A**) Phylogenetic tree of 19 enzybiotic enzymes, including hydrolases and oxidoreductases, illustrating functional divergence and evolutionary clustering. (**B**) Pie chart showing the relative abundance (%) of enzybiotic genes across 15 metagenomic environments. Percentages represent the proportion of total identified enzybiotic genes attributed to each ecosystem. The inset panel summarizes the overall distribution of major enzyme families, including peptidases (55%), proteases (20%), laccases (15%), and amidases (10%). (**C**) Bar plot representing the total number of enzybiotic genes detected in each ecosystem (absolute counts).

Oxidoreductases such as FAD-dependent oxidases (WP_237135401.1) and marinocine (AAY33849.1) formed separate branches, reflecting functional specialization in ROS-mediated antimicrobial mechanisms. Enzymes like PersiLac1, PersiLac2, and PersiLac3 possessed conserved redox domains associated with oxidative stress pathways.

Interestingly, zoocin A (AAC46072.1) and N-acetylmuramoyl-L-alanine amidase (AHZ45666.1) appeared as phylogenetic outliers, suggesting possible evolutionary transitions that may have broadened their functional repertoires.

Enzyme distribution analysis across environments revealed distinct ecosystem-specific patterns (Fig. 5B and C). To provide a comprehensive assessment of enzybiotic distribution, we report both relative and absolute abundance metrics. Figure 5B summarizes the relative abundance (%) of enzybiotic genes across the 15 metagenomic environments, highlighting proportional differences in ecosystem-level representation. Complementarily, Fig. 5C presents the absolute (raw) counts of enzybiotic genes detected in each ecosystem, enabling direct identification of environments harboring the highest total enzybiotic gene numbers (e.g., Plastic Catalog and Charmshahr samples showed the highest absolute counts). In addition, Supplementary Table S1 reports total predicted genes, absolute enzybiotic gene counts, and normalized enzybiotic density (per 10,000 predicted genes) to account for differences in gene catalog size across environments and enable standardized cross-ecosystem comparison.

At the individual gene level, AHZ45606.1 was the most abundant sequence across seven environments, including brine, Charmshahr, PE-soil, PET-cow-gut, Rimov Reservoir, Varamin, and alpine vegetation. In contrast, AAC46072.1 predominated in plastic-polluted environments such as seawater, groundwater, ocean, and river estuaries. Hot spring samples were uniquely characterized by AHZ45646.1, while WP_000405234.1 was the most prevalent in fecal samples. Certain enzymes, including AHZ45606.1, AHZ45613.1, AAC46072.1, and PersiProtease1, were consistently found across all environments. Moreover, PersiLac1 and PersiLac2 displayed a strong correlation, suggesting shared ecological roles.

Analysis of ecological niches highlighted the strong presence of enzybiotic genes in plastic-polluted environments, supporting the hypothesis that microbial communities utilize enzybiotics to cope with environmental stress.

In uncontaminated aquatic environments, enzybiotic genes were more evenly distributed, with a single gene often dominating. In contrast, polluted terrestrial environments exhibited higher genetic diversity, indicating the selective pressure imposed by contamination. Notably, in polluted aquatic systems, one enzybiotic gene could account for over 70% of the total gene composition, pointing to the dominance of specific antimicrobial strategies under stress.

Conversely, in uncontaminated terrestrial environments, gene distribution was more uniform, suggesting lower selective pressure and reduced microbial competition. These findings underscore the ecological significance of enzybiotics and their potential biotechnological applications in targeting antimicrobial resistance and environmental resilience.

In our analysis, we considered the potential for both convergent and divergent evolution of antimicrobial enzymes in different ecosystems. The phylogenetic tree of the 19 representative enzybiotics revealed distinct evolutionary clusters for hydrolases, peptidases, oxidoreductases, and lytic enzymes, with conserved catalytic domains in each family. However, branch length variation suggests functional divergence, particularly within families such as zoocin A homologs, which display longer branches, indicating adaptation to different bacterial cell wall structures of the host. This divergence was more pronounced between ecosystems, such as hot springs and terrestrial soils, whereas aquatic ecosystems showed lower divergence, particularly among similar environments, such as seawater and estuaries.

Furthermore, we observed that some taxa, previously reported to have few antimicrobial enzymes, now show an unexpectedly high count of enzybiotics, particularly in contaminated aquatic environments. This increase could be attributed to ecological pressures that favor the survival of stress-tolerant and oxidative stress-associated genera, which were enriched in our datasets. Notably, *Actinomycetota*, particularly *Streptomyces*-like lineages, were the most enriched in many ecosystems, whereas *Pseudomonadota* dominated oxidative and stress-related enzymes in contaminated habitats. This highlights the dynamic and complex relationship between microbial taxa and their functional potential in various environments.

These findings suggest that microbial communities, particularly those in diverse and less-explored environments, harbor a significant diversity of antimicrobial enzymes, some of which are novel and potentially applicable in various fields. The observed functional divergence and taxa expansion provide new insights into the evolution of antimicrobial mechanisms and underscore the need to explore a wider range of ecological niches to fully understand the antimicrobial capabilities of microbial communities.

In this study, we emphasize the analytical foundations of the MiGPC framework, which extend beyond catalog construction to establish a comprehensive computational pipeline for enzybiotic discovery. By integrating hybrid sequence curation, multilayer functional and taxonomic annotation, genome-resolved analyses, ecological modeling, and structure-based validation, the workflow presented here provides a unified approach for identifying, contextualizing, and characterizing enzybiotic enzymes across diverse environments. This integrated analytical perspective highlights how the MiGPC not only aggregates sequence information but also enables the discovery of ecological patterns, functional signatures, and genome-level associations that advance our understanding of the environmental distribution and evolutionary potential of enzybiotics.

## Conclusions

This study presents the first integrated gene and genome targeted enzybiotic catalog the MiGPC dedicated to environments enriched with antimicrobial enzymes and bacteria. By analyzing metagenomic data from a wide range of ecological niches, this catalog offers a comprehensive and high-resolution resource for advancing

research on antimicrobial resistance (AMR). Our findings revealed an extensive repertoire of enzybiotic genes and proteins, including over 100 million unique sequences, many of which are novel and currently unannotated in the database. This highlights a largely untapped pool of functional diversity, with significant potential for future antimicrobial discovery. Taxonomic profiling and CAZyme analysis further illuminate the enzymatic capacity and microbial ecology of these environments. Through unsupervised clustering and statistical modeling, we demonstrated that environmental factors play a key role in shaping the distribution of antimicrobial agents, separating samples into distinct ecological groups. Notably, polluted environments especially those affected by plastics harbor unique enzybiotic profiles, suggesting that microbial communities adapt to anthropogenic stress through enzybiotic innovation. The MiGPC serves as a foundational reference for functional metagenomics, drug discovery, and microbial ecology, enabling the identification of promising candidates for next-generation antimicrobial agents. Moving forward, this catalog can facilitate multi-omics investigations, the characterization of unannotated genes, and the development of novel biotechnological and therapeutic strategies to combat multidrug-resistant pathogens. This catalog is designed to be extensible and can be periodically updated with metagenomic data from newly explored environments, enabling the release of future versions that reflect a broader ecological and functional diversity.

## Data availability

The metagenomic sequencing data used in this study are publicly available in the NCBI Sequence Read Archive (SRA) under the following accession numbers: SRR6441339, SRR28167100, SRR26623185, ERR4298344, SRR15826900, ERR9631090, SRR28167096, SRR28167104, SRR25490518, SRR26901926, SRR13060942, SRR29090698, SRR23085642, and SRR12059165. Additionally, the final integrated gene and protein catalogs, as well as the catalog of predicted antimicrobial enzymes (enzybiotics) generated in this study, are available in a publicly accessible GitHub repository (https://github.com/kkavousi/MiGPC-Catalog).

Received: 1 September 2025; Accepted: 10 March 2026
Published online: 26 March 2026

## References

1. Murray, C. J. et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **399**, 629–655 (2022).
2. Laxminarayan, R. et al. Antibiotic resistance-the need for global solutions. *Lancet Infect. Dis.* **13**, 1057–1098 (2013).
3. Pantiora, P. D., Georgakis, N. D., Premetis, G. E. & Labrou, N. E. Metagenomic analysis of hot spring soil for mining a novel thermostable enzybiotic. *Appl. Microbiol. Biotechnol.* https://doi.org/10.1007/s00253-023-12979-2 (2024).
4. Georgakis, N., Premetis, G. E., Pantiora, P., Varotsou, C. & Bodourian, C. S. The impact of metagenomic analysis on the discovery of novel endolysins. *Appl. Microbiol. Biotechnol.* https://doi.org/10.1007/s00253-025-13513-2 (2025).
5. Thallinger, B., Prasetyo, E. N., Nyanhongo, G. S. & Guebitz, G. M. Antimicrobial enzymes: An emerging strategy to fight microbes and microbial biofilms. *Biotechnol. J.* **8**, 97–109 (2013).
6. Nagashima, Y. et al. Isolation and cDNA cloning of an antibacterial L-amino acid oxidase from the skin mucus of the great sculpin Myoxocephalus polyacanthocephalus. *Comp. Biochem. Physiol. - B Biochem. Mol. Biol.* **154**, 55–61 (2009).
7. Villa, T. G., Feijoo-Siota, L., Rama, J. L. R., Sánchez-Pérez, A. & de Miguel-Bouzas, T. Chapter 43 - Enzybiotics: Application in Food Packaging. in (ed. Barros-Velázquez, J. B. T.-A. F. P. (Second E.) 681–703Academic Press, (2025). https://doi.org/10.1016/B978-0-323-90747-7.00044-2
8. Derollez, E. & Lesterlin, C. Design, potential and limitations of conjugation- based antibacterial strategies. 1–11 (2024). https://doi.org/10.1111/1751-7915.70050
9. Rahimi-Midani, A., Lee, S. W. & Choi, T. J. Potential Solutions Using Bacteriophages against Antimicrobial Resistant Bacteria. *Antibiot (Basel Switzerland)* **10**, (2021).
10. Zhang, X., Wang, S. & Wang, F. Microbial proteases and their applications. 1–24 (2023). https://doi.org/10.3389/fmicb.2023.1236368
11. Zhou, C., Wang, Q., Jiang, J. & Gao, L. Nanozybiotics: Nanozyme-Based Antibacterials against Bacterial Resistance. *Antibiot (Basel Switzerland)* **11**, (2022).
12. Villa, T. G. & Veiga-Crespo, P. Enzybiotics: Antibiotic enzymes as drugs and therapeutics. *Enzybiotics: Antibiotic Enzymes as Drugs Ther.* https://doi.org/10.1002/9780470570548 (2009).
13. Sampaio, L. M. P. et al. Laccase immobilization on bacterial nanocellulose membranes: Antimicrobial, kinetic and stability properties. *Carbohydr. Polym.* **145**, 1–12 (2016).
14. Ozdemir, H., Hacibeyoglu, H. I. & Uslu, H. Purification of lactoperoxidase from creek-water buffalo milk and investigation of kinetic and antibacterial properties. *Prep Biochem. Biotechnol.* **32**, 143–155 (2002).
15. Ullah, A. et al. Structural insights into selectivity and cofactor binding in snake venom l-amino acid oxidases. *Biochem. Biophys. Res. Commun.* **421**, 124–128 (2012).
16. Kasai, K. et al. Novel l-amino acid oxidase with antibacterial activity against methicillin-resistant Staphylococcus aureus isolated from epidermal mucus of the flounder Platichthys stellatus. *FEBS J.* **277**, 453–465 (2010).
17. Taraszkiewicz, A., Fila, G., Grinholc, M. & Nakonieczna, J. Inno ati e trategies to ercome io lm esistance. (2013).
18. Nakamura, H., Fang, J. & Maeda, H. Protective role of D-amino acid oxidase against Staphylococcus aureus infection. *Infect. Immun.* **80**, 1546–1553 (2012).
19. Andreo-Vidal, A., Sanchez-Amat, A. & Campillo-Brocal, J. C. The pseudoalteromonas luteoviolacea L-amino acid oxidase with antimicrobial activity is a flavoenzyme. *Mar Drugs* **16**, (2018).
20. Oram, J. D. & Reiter, B. The inhibition of streptococci by lactoperoxidase, thiocyanate and hydrogen peroxide. The oxidation of thiocyanate and the nature of the inhibitory compound. *Biochem. J.* **100**, 382–388 (1966).
21. Mittal, P., Pk, V. P., Dhakan, D. B., Kumar, S. & Sharma, V. K. Metagenome of a polluted river reveals a reservoir of metabolic and antibiotic resistance genes. 1–12 (2019).
22. Ahmad, T., Singh, R. S., Gupta, G., Sharma, A. & Kaur, B. Metagenomics in the search for industrial enzymes. *Biomass Biofuels Biochemicals: Adv. Enzyme Technol.* https://doi.org/10.1016/B978-0-444-64114-4.00015-7 (2019).
23. Penders, J. et al. The human microbiome as a reservoir of antimicrobial resistance. *Front. Microbiol.* **4**, 87 (2013).
24. Kelly, S. A., Rodgers, A. M., O'Brien, S. C., Donnelly, R. F. & Gilmore, B. F. Gut Check Time: Antibiotic Delivery Strategies to Reduce Antimicrobial Resistance. *Trends Biotechnol.* **38**, 447–462 (2020).
25. Shang, Z., Chan, S. Y., Song, Q., Li, P. & Huang, W. The Strategies of Pathogen-Oriented Therapy on Circumventing Antimicrobial Resistance. *Res. (Washington, D.C.)* 2016201 (2020). (2020).
26. Li, Z. et al. Regionalization and Shaping Factors for Microbiomes and Core Resistomes in Atmospheric Particulate Matters. *mSystems* 7, e0069822 (2022).

27. Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
28. Han, Y., Zhang, C., Zhao, Z., Peng, Y. & Liao, J. A comprehensive genomic catalog from global cold seeps. 1–10 (2023). https://doi.org/10.1038/s41597-023-02521-4
29. Han, Y., Zhang, C., Zhao, Z., Peng, Y. & Liao J. A comprehensive catalog with 100 million genes and 3, 000 metagenome-assembled genomes from global cold seep sediments. (2023).
30. Zhu, B. & Gao, G. F. Metagenome-assembled genomes and gene catalog from the chicken gut microbiome aid in deciphering antibiotic resistomes. 1–9 (2021). https://doi.org/10.1038/s42003-021-02827-2
31. Chen, C. et al. Expanded catalog of microbial genes and metagenome-assembled genomes from the pig gut microbiome. *Nat Commun* 1–13 https://doi.org/10.1038/s41467-021-21295-0
32. Jahanshahi, D. A., Ariaeenejad, S. & Kavousi, K. A metagenomic catalog for exploring the plastizymes landscape covering taxa, genes, and proteins. *Sci. Rep.* **13**, 1–14 (2023).
33. Iqbal, H. A., Craig, J. W. & Brady, S. F. Antibacterial enzymes from the functional screening of metagenomic libraries hosted in Ralstonia metallidurans. *FEMS Microbiol. Lett.* **354**, 19–26 (2014).
34. Premetis, G. E., Stathi, A., Papageorgiou, A. C. & Labrou, N. E. Structural and functional features of a broad-spectrum prophage-encoded enzybiotic from Enterococcus faecium. *Sci. Rep.* **13**, 1–15 (2023).
35. Afoshin, A. et al. Structural and Functional Characterization of β–lytic Protease from Lysobacter capsici VKM B–2533 T bacteriolytic enzymes. The structural and functional studies of lytic tively developing at present [ 22–27 ]. However, a Blp, the structure has no. (2022).
36. Abedanzadeh, S., Ariaeenejad, S. & Karimi, B. International Journal of Biological Macromolecules Revolutionizing protein hydrolysis in wastewater: Innovative immobilization of metagenome-derived protease in periodic mesoporous organosilica with imidazolium framework. *Int. J. Biol. Macromol.* **278**, 134966 (2024).
37. Ariaeenejad, S. et al. Enzymatically triggered delignification through a novel stable laccase: A mixed in-silico /in-vitro exploration of a complex environmental microbiota. *Int. J. Biol. Macromol.* **211**, 328–341 (2022).
38. Motamedi, E. et al. Efficient removal of various textile dyes from wastewater by novel thermo-halotolerant laccase. *Bioresour Technol* **337**, (2021).
39. Ariaeenejad, S., Kavousi, K., Zeinalabedini, M., Afshar Jahanshahi, D. & Beh-Afarin, S. R. Enhanced bioremediation of saline azo dye effluents using PersiLac3: A thermo-halotolerant laccase from a tannery metagenome. *Results Eng.* **24**, 103172 (2024).
40. Mai-Prochnow, A. et al. Biofilm development and cell death in the marine bacterium Pseudoalteromonas tunicata. *Appl. Environ. Microbiol.* **70**, 3232–3238 (2004).
41. Yu, M. et al. Purification and characterization of antibacterial compounds of Pseudoalteromonas flavipulchra JG1. *Microbiology* **158**, 835–842 (2012).
42. Lucas-Elío, P., Gómez, D., Solano, F. & Sanchez-Amat, A. The antimicrobial activity of marinocine, synthesized by Marinomonas mediterranea, is due to hydrogen peroxide generated by its lysine oxidase activity. *J. Bacteriol.* **188**, 2493–2501 (2006).
43. Spencer, J., Murphy, L. M., Conners, R., Sessions, R. B. & Gamblin, S. J. Crystal structure of the LasA virulence factor from Pseudomonas aeruginosa: substrate specificity and mechanism of M23 metallopeptidases. *J. Mol. Biol.* **396**, 908–923 (2010).
44. Thumm, G. & Götz, F. Studies on prolysostaphin processing and characterization of the lysostaphin immunity factor (Lif) of Staphylococcus simulans biovar staphylolyticus. *Mol. Microbiol.* **23**, 1251–1265 (1997).
45. García, P., García, J. L., García, E. & López, R. Nucleotide sequence and expression of the pneumococcal autolysin gene from its own promoter in Escherichia coli. *Gene* **43**, 265–272 (1986).
46. Simmonds, R. S., Pearson, L., Kennedy, R. C. & Tagg, J. R. Mode of action of a lysostaphin-like bacteriolytic agent produced by Streptococcus zooepidemicus 4881. *Appl. Environ. Microbiol.* **62**, 4536–4541 (1996).
47. Ariaeenejad, S., Kavousi, K., Han, J. L., Ding, X. Z. & Salekdeh, G. H. Efficiency of an alkaline, thermostable, detergent compatible, and organic solvent tolerant lipase with hydrolytic potential in biotreatment of wastewater. *Sci. Total Environ.* **866**, 161066 (2023).
48. Gu, S. & Fang, L. X. X. Using SOAPaligner for Short Reads Alignment. *Curr Protoc Bioinformatics*. (2013). 10.1002/0471250953.bi1111s44.
49. Li, D. et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
50. Li, H. et al. The Sequence Alignment / Map format and SAMtools. 25, 2078–2079 (2009).
51. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. **38**, 1–15 (2010).
52. Goudarzi, R., Afshar, D., Kavousi, A. & Ariaeenejad, S. Discovery and engineering of bifunctional enzymes for lignocellulose degradation: Metagenomic and computational approaches. *Biotechnol. Rep.* **48**, e00926 (2025).
53. Jahanshahi, D. A., Tiyoula, F. N., Kavousi, K., Beh-Afarin, S. R. & Ariaeenejad, S. Sustainable Solutions through Metagenomic Laccases: Industrial and Environmental Perspectives. *Environ. Technol. Innov.* **104568** https://doi.org/10.1016/j.eti.2025.104568 (2025).
54. Zhang, H. et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. **46**, 95–101 (2018).
55. Afshar, D., Reza, M., Barzani, R. & Bahram, M. Ecotoxicology and Environmental Safety Metagenomic exploration and computational prediction of novel enzymes for polyethylene terephthalate degradation. *Ecotoxicol. Environ. Saf.* **289**, 117640 (2025).
56. Kang, D. D. et al. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* https://doi.org/10.7717/peerj.7359 (2019).
57. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* https://doi.org/10.1101/gr.186072.114 (2015).
58. López-Leal, G., Cornejo-Granados, F., Hurtado-Ramírez, J. M., Mendoza-Vargas, A. & Ochoa-Leyva, A. Functional and taxonomic classification of a greenhouse water drain metagenome 06 Biological Sciences 0604 Genetics. *Stand. Genomic Sci.* **13**, 1–9 (2018).
59. Goussarov, G. et al. Accurate prediction of metagenome-assembled genome completeness by MAGISTA, a random forest model built on alignment-free intra-bin statistics. *Environ. Microbiomes*. **17**, 1–13 (2022).
60. Shahraki, M. F., Atanaki, F. F., Ariaeenejad, S., Ghaffari, M. R., Norouzi-Beirami, M. H., Maleki, M. & Kavousi, K. computational learning paradigm to targeted discovery of biocatalysts from metagenomic data: A case study of lipase identification. *Biotechnol. Bioeng.* **119**, (2022).
61. Ariaeenejad, S. et al. Highly Efficient Computationally Derived Novel Metagenome α-Amylase With Robust Stability Under Extreme Denaturing Conditions. *Front. Microbiol.* **12**, 1–14 (2021).
62. Id, S. A., Mousivand, M. & Dezfouli, P. M. A computational method for prediction of xylanase enzymes activity in strains of Bacillus subtilis based on pseudo amino acid composition features. 1–16 (2018).
63. Ariaeenejad, S. & Motamedi, E. Improved saccharification of rice straw by removing phenolic compounds using a stable immobilized metagenome-derived laccase on sodium alginate-based hydrogel. *Biochem. Eng. J.* **198**, 109021 (2023).
64. Shahraki, M. F. et al. K. K. A computational learning paradigm to targeted discovery of biocatalysts from metagenomic data: A case study of lipase identification. *Biotechnol. Bioeng.* **119**, 1115–1128 (2022).
65. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

## Acknowledgements
This research was carried out by the support of the Institute of Biochemistry and Biophysics (IBB), University of Tehran and Department of Systems and Synthetic Biology, Agricultural Biotechnology Research Institute of Iran (ABRII).

## Author contributions
**Shohreh Ariaeenejad: ** Conceptualization, Experimental Methodology, Writing - original draft, Resources, Project administration, Writing - review & editing.**Kaveh Kavousi: ** Conceptualization, Methodology, computational and machine learning approaches and bioinformatics data analysis, Project administration, Writing - review & editing.**Donya Afshar Jahanshahi: ** Contributed to the computational and machine learning approaches and bioinformatics data analysis, Writing - original draft.**Mohammad Reza Zabihi: ** Contributed to the computational and machine learning approaches, Writing - original draft.**Arad Ariaeenejad: ** Contributed to the computational and machine learning approaches and bioinformatics data analysis, Writing - original draft.**Arman Hasannejad; ** Contributed to the computational and machine learning approaches and bioinformatics data analysis, Writing - original draft.**Mohammad Reza Ghaffari: ** Conceptualization, Experimental Methodology, Writing - original draft, Resources, Project administration, Writing - review & editing.

## Funding
This research was supported by grants from Agricultural Biotechnology Research Institute of Iran (ABRII) and Center for International Scientific Studies & Collaborations (CISSC) (funding reference number 4020052).

## Declarations

## Competing interests
The authors declare no competing interests.

## Statistical analysis
Statistical analyses and visualizations were conducted in the Python 3.11.0 and PyCharm 2022.2.4 environment, utilizing the Matplotlib and Seaborn packages. Additionally, the diversity of taxa and taxonomic phylogenetic tree were visualized using Interactive Tree Of Life (iTOL) v5 [65], respectively. Further information can be obtained from the corresponding author upon request.

## Additional information
**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-44250-9.

**Correspondence** and requests for materials should be addressed to S.A. or K.K.