

Lawrence Berkeley National Laboratory

LBL Publications

Title

“Quadrupling” the protein family space with global metagenomics

Permalink

<https://escholarship.org/uc/item/0c41x11t>

Journal

Nucleic Acids Research, 54(18)

ISSN

0305-1048

Authors

Aplakidou, Eleni
Baltoumas, Fotis A
Chasapi, Maria N
et al.

Publication Date

2026-09-22

DOI

10.1093/nar/gkag938

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

“Quadrupling” the protein family space with global metagenomics

Eleni Aplakidou^{1,2}, Fotis A. Baltoumas¹, Maria N. Chasapi¹, Efthimia Lamari¹, Iro N. Chasapi^{1,3}, Ilias Georgakopoulos-Soares^{4,5}, Evangelos Karatzas⁶, Ioannis Iliopoulos², Aydin Buluç^{7,8}, Robert D. Finn⁶, Antonio Pedro Camargo^{9,10}, Nikos C. Kyrpides⁹, Georgios A. Pavlopoulos^{1,11,*}

¹Institute for Fundamental Biomedical Research, BSRC “Alexander Fleming”, Vari 16672, Greece

²Division of Basic Sciences, University of Crete Medical School, Heraklion 71110, Greece

³Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens 16122, Greece

⁴Institute for Personalized Medicine, Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey, PA 17033, United States

⁵Division of Pharmacology and Toxicology, College of Pharmacy, The University of Texas at Austin, Dell Pediatric Research Institute, Austin, TX 78712, United States

⁶European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

⁷Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States

⁸Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, United States

⁹DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States

¹⁰Department of Biochemistry, Institute of Chemistry, University of São Paulo, São Paulo, SP 05508-900, Brazil

¹¹Division of Biological and Life Sciences, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Building 1B, Masdar City, Abu Dhabi, United Arab Emirates

*To whom correspondence should be addressed. Email: pavlopoulos@fleming.gr

Abstract

The known universe of protein families represents only a small fraction of nature’s molecular diversity. From 40.3 billion predicted open reading frames across 40 446 metagenomes, 9540 metatranscriptomes, and 539 million proteins from 167 415 reference genomes, we identified 608 258 previously uncharacterized putative protein families with ≥ 100 members and 6.5 million families with ≥ 25 members, none matching known Pfam domains or reference proteins. This effort doubles the known repertoire of large families and quadruples that of smaller families. Integration of AlphaFold2-based structural predictions with gene-neighborhood and taxonomic analyses enables the characterization of these previously unannotated proteins, revealing candidates for both novel and known biological functions in understudied microbial lineages and biomes. This expanded repertoire provides insights into microbial adaptation and broadens the molecular toolkit available for biotechnology, highlighting the power of global metagenomics to uncover hidden protein diversity.

Received: February 25, 2026. Revised: August 26, 2026. Accepted: August 27, 2026

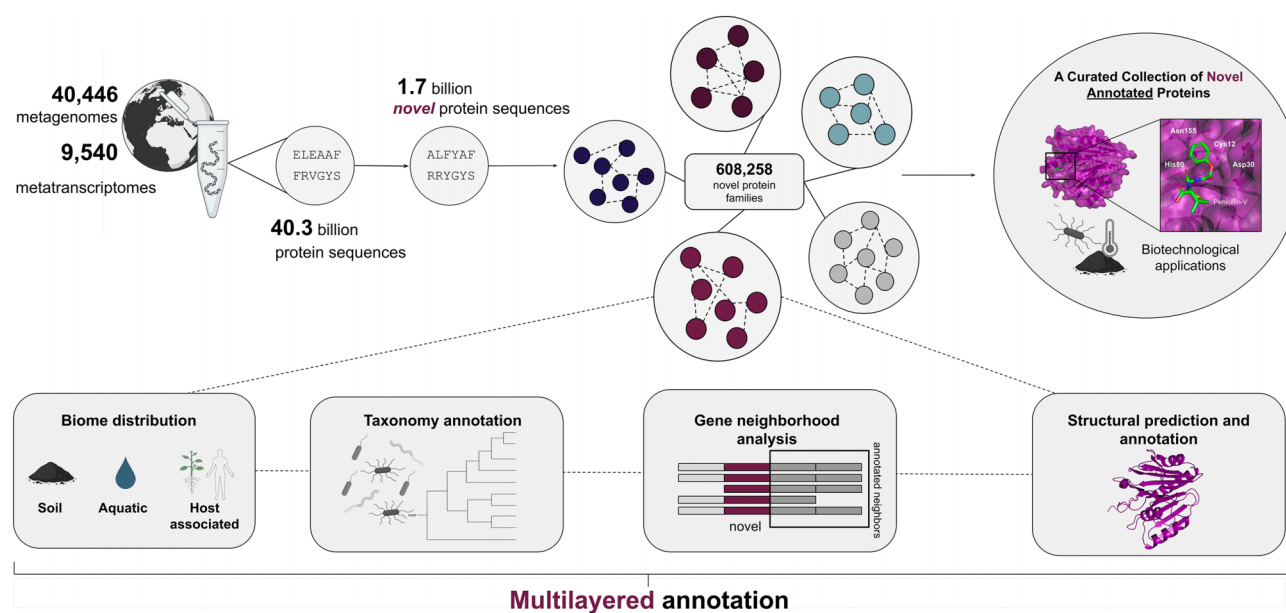
© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License

(<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other

permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Graphical abstract



Introduction

Proteins serve as the fundamental molecular machines of life, playing essential roles in metabolism, gene expression, cellular defense, and structural integrity. Gaining a comprehensive understanding of proteins is crucial for advancing knowledge in biology, evolution, and disease mechanisms. Although their central role in life, the immense diversity of microbial proteins remains largely uncharacterized, representing a vast reservoir of functional and ecological information yet to be explored. With the explosion of sequencing technologies, we now have access to vast amounts of metagenomic protein sequence data, housed in databases that are growing at an unprecedented rate [1–3]. Existing databases containing protein information, such as sequences, structures, and functional annotations, are confined to well-characterized protein families; others serve mainly as collections of unstructured datasets, and many focus on specific environments, resulting in large portions of protein diversity being overlooked (Supplementary Table S1). Consequently, despite this rapid expansion of data, a vast and largely uncharted metagenomic “protein dark matter” remains underexplored and underrepresented in current repositories. These novel proteins, often hidden within metagenomic datasets, are not merely rare variants; recent studies have shown that they include abundant families that are globally distributed and ecologically significant [4].

By clustering billions of metagenomic sequences and integrating structural prediction, gene-neighborhood context, and ecological metadata, it becomes possible to define both abundant and rare protein families, identify novel folds, and assign putative functions to previously uncharacterized sequences. This approach not only increases the coverage of microbial protein space but also bridges the gap between sequence discovery and potential functional insight, enabling targeted exploration of enzymes, metabolic pathways, and molecular scaffolds for biotechnology [5, 6].

Previously published studies, including FESNov [3], AGNOSTOS [7], and NMPFams v1 [8, 9], have specifically aimed to explore this uncharted protein space and systematically characterize novel proteins from metagenomic data. FESNov

leverages a large collection of assembled prokaryotic genomes (MAGs and SAGs) to identify novel gene families with strong evolutionary and functional signals, thereby focusing on genome-resolved uncultivated taxa while not addressing unassembled metagenomic sequences or non-prokaryotic components. AGNOSTOS is a computational workflow and conceptual framework developed to unify the known and unknown microbial coding sequence space by integrating large-scale genomic and metagenomic data from human-associated and marine ecosystems. The framework operates by inferring, validating, and refining homologous gene clusters and higher-order gene cluster communities. NMPFams v1 is a large-scale catalog of 106 198 novel metagenomic protein families, each comprising at least 100 related sequences and showing no detectable similarity to reference microbial genes or Pfam entries while also providing taxonomic, environmental, and predicted structural annotations.

In this study, we present a curated expansion of microbial protein families that builds upon and extends previous large-scale efforts in the field. Compared to these prior studies, our approach differs substantially in methodology, dataset composition, and scale. Specifically, we analyze raw metagenomic data across all domains of life, including viruses, and incorporate a significantly larger and more diverse protein dataset spanning multiple habitat types, together with metatranscriptomic data to enhance functional interpretation. Notably, the present study builds upon NMPFams v1 by substantially expanding the resource through the integration of a much larger and more diverse collection of metagenomic sequences. In addition to increasing its scale, this new version moves beyond sequence cataloging by incorporating biological context and putative functional annotations, thereby enhancing both the comprehensiveness and biological interpretability of the database.

Here, using a scalable, biologically validated computational pipeline applied to one of the most extensive microbial datasets assembled to date [5] (Fig. 1), we generate a large-scale catalog of novel protein families and predicted folds with potential functions. This framework enables systematic ex-

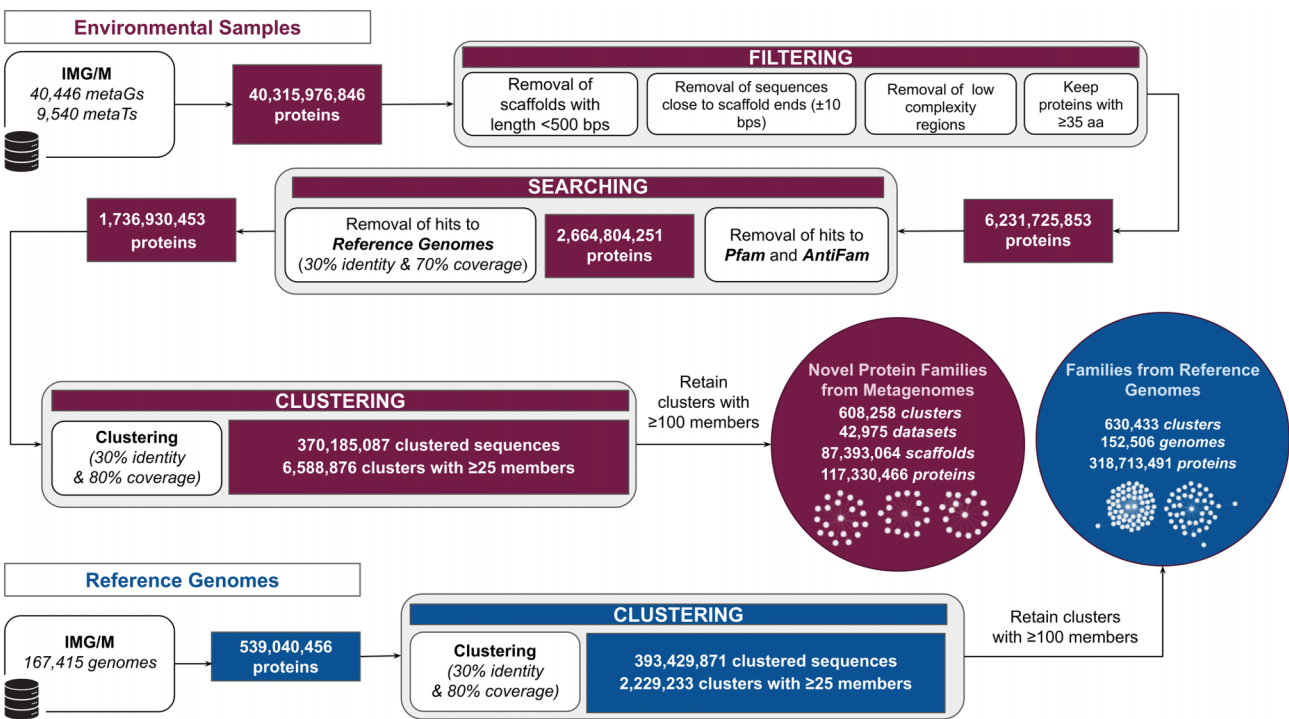


Figure 1. Pipeline illustrating the sequence filtering and clustering methodology for generating protein families from both reference genomes and metagenomes/metatranscriptomes.

ploration of the functional “dark matter” of metagenomes, uncovering new structural and biological diversity. It further supports applications ranging from enzyme discovery to synthetic biology, helping to transform metagenomic sequence space into biological insight.

Materials and methods

Data collection and filtering

All publicly accessible metagenomes, metatranscriptomes, and reference genomes were obtained from the Integrated Microbial Genomes & Microbiomes (IMG/M) database provided by the Joint Genome Institute (Fig. 1). Metagenomes and metatranscriptomes were retrieved in October 2023 [5], while reference genomes were collected in December 2024. A total of 522 135 784 protein sequences were extracted from 143 797 bacterial genomes, 761 975 sequences from 19 747 viral genomes, 6 651 575 sequences from 3128 archaeal genomes, and 9 491 122 sequences from 803 eukaryotic genomes, culminating in 539 040 456 sequences. All sequences were solely derived from isolated reference genomes, excluding metagenome-assembled genomes (MAGs), single-amplicon genomes (SAGs), cell enrichment samples, and single particle sorts. All data used in this analysis are publicly available and all domains of life have been included in the analysis. The study analyzed 49 986 microbiome datasets (40 446 metagenomes and 9 540 metatranscriptomes, totaling 10 851 187 196 scaffolds) that initially contained 40 315 976 846 open reading frames (ORFs) of varying lengths from assembled metagenomes. Protein predictions in IMG/M [5] are generated using Prodigal v2.6.3 [6] and GeneMarkS-2 [10]. To maintain a dataset composed primarily of reliable protein predictions, specific filtering criteria were employed. Given that genes situated toward the center of

a contig are more likely to be complete. At the same time, those closer to the ends may be truncated; sequences located within 10 nucleotide residues of scaffold ends were also removed to mitigate the inclusion of potentially truncated genes. Furthermore, only sequences with 35 or more residues from scaffolds exceeding 500 nucleotides in length were retained to ensure that mainly complete genes were represented. This is because shorter scaffolds increase the risk that genes are located near the ends, which could lead to truncation. Low-complexity regions were masked using the tantan application (version 26) [11]. After these filtering steps, any remaining sequences shorter than 35 amino acids were also discarded. The resulting final catalog comprised 6 231 725 853 sequences derived from 44 856 microbiome datasets and 2 893 181 669 scaffolds. We consider this to be the high-quality non-redundant microbiome protein sequence set (high quality/non redundant), which accounted for 15.46% of the starting dataset.

Identification of novel proteins: hits to Pfam and Antifam

To determine which of the filtered proteins possess known domains, the filtered metagenome dataset was compared against the Pfam-A (version 37) database of profile Hidden Markov Models (HMMs) (Fig. 1), which contains 21 979 families and 709 clans. Pfam hits were identified using the hmmsearch tool from the HMMER 3.3.2 package [12], applying the default trusted cutoff (–cut_tc). Using the trusted cutoff (TC) threshold provides the most stringent filtering available, ensuring that sequences excluded from our novelty search match Pfam families with the same high confidence as their founding members. While this conservative approach prioritizes high-confidence novelty, it naturally enriches for smaller, single-domain proteins and intentionally excludes multidomain proteins that might otherwise contain novel domains. Follow-

ing the removal of Pfam hits, the final dataset comprised 2 664 804 251 proteins, extracted from 44 796 datasets and 1 579 469 832 scaffolds.

The same methodology was applied to Antifam [7], a database specifically designed to identify and eliminate spurious or erroneous protein sequences arising from low-complexity regions or other repetitive patterns, improperly extended ORFs, or incorrectly translated non-protein-coding RNA genes (Fig. 1). While Pfam catalogs functional protein domains and families, Antifam focuses on detecting sequences or families that are likely artifacts (such as those arising from low-complexity regions, coiled-coil structures, or other repetitive patterns) that may falsely resemble functional domains. The Antifam filtering process yielded 398 073 hits, resulting in a final dataset of 2 664 406 178 proteins (43% of the HQ-NR-MB set).

Identification of novel proteins: hits to reference genomes

To identify novel proteins in the filtered microbiome protein dataset, additional comparisons were made against the most recent reference genomes collected in late 2024, which encompass 539 040 456 protein sequences. Two sequential methods were employed for the comparison (Fig. 1). Initially, the LAST algorithm [13] was utilized with parameters set at 30% similarity and 70% coverage bidirectional [14]. LAST is designed explicitly for large-scale genomic or protein sequence comparisons, utilizing a variation of spaced seeds to enhance sensitivity while minimizing memory usage. Due to its speed and ability to handle large datasets, LAST is well-suited for metagenomic data analysis but can miss hits with lower sequence identity. For this reason, the dataset that yielded no hits in the LAST comparison (1 957 621 091 proteins) was subsequently analyzed against the reference genomes using DIAMOND [15], which is more sensitive, albeit slower. The DIAMOND search was performed using the same search parameters. Requiring bidirectional coverage ensures that both the query and target are substantially aligned, maximizing biological relevance and eliminating false positives when matching metagenomic sequences against reference databases. This stringent requirement is critical given that metagenomic proteins are often truncated relative to full-length references; however, because meeting high coverage against larger reference proteins is difficult for longer candidates, this step naturally enriched our dataset for smaller proteins. Following these steps, the final dataset comprised 1 736 930 453 proteins, sourced from 44 721 microbiome datasets and 1 115 001 758 scaffolds.

Clustering and generation of protein families

In this analysis, MMseqs2 Linclust [16] was run in bidirectional mode with 30% sequence identity and 80% coverage, balancing sensitivity and computational load (Fig. 1). Clustering was performed on all sequences irrespective of taxonomy. A direct comparison of the previous clustering methodology (NMPFam v.1) is presented in [Supplementary File S1](#). Using these parameters, 370 185 087 proteins were clustered into 6 588 876 families, each containing at least 25 members. These families were derived from 44 198 datasets and spanned 253 359 325 scaffolds, representing 21.31% of the total 1 736 930 453 proteins analyzed. Proteins that did not meet the minimum clustering threshold were excluded from further analysis.

Mapping to metagenome-assembled genomes

In this study, all prokaryotic MAGs were sourced from the IMG/M database, while eukaryotic MAGs were excluded. According to the MIMAG standards [17], the bins were classified as high, medium, or low quality, with high-quality MAGs characterized by >90% completeness and <5% contamination, as determined with CheckM [18]. As of June 2024, the dataset encompassed 21 817 high-quality prokaryotic MAGs, providing a robust foundation for the analysis of microbial diversity [19]. For this comparison, DIAMOND was utilized with parameters set at 30% similarity and 70% coverage bidirectional. Rather than serving as a filter for novelty identification, given that many MAGs contain functionally unknown proteins, the comparison with MAGs was conducted primarily as a contextual analysis to evaluate sequence overlap and distribution across reconstructed genomes.

Profile generation

Novel protein families containing 100 or more members were gathered and subjected to multiple sequence alignment (MSA) generation using MAFFT [20]. After the initial alignment, trimming was applied to minimize gaps as follows: The “central” or “pivot” sequence of each MSA was identified by converting aligned sequences to numerical values, constructing an all-versus-all matrix, and computing the Hamming distance. Subsequently, the central sequence was used as a guide to identify gaps, isolate conserved regions, and ensure that poorly aligned or gapped areas were excluded from subsequent analysis. Finally, hhfilter [21] was employed to reduce sequence redundancy within each family MSA, with parameters set to an identity threshold of 95% and a coverage of 70%, thereby reducing bias by eliminating highly similar sequences. Only families with at least 16 effective sequences after filtering were retained for structural prediction, following the protocol established in the original version of NMPFamsDB. This cutoff was chosen as it was previously found to be the minimum number of sequences per MSA capable of producing the high-percentage of high-quality results [8].

Candidate families for structural model prediction

Structure prediction was performed using the ColabFold [22] implementation of AlphaFold2 [23] in *de novo* mode (no template input), following the protocol established in the original version of NMPFamsDB [8]. Five models were generated per run, with three recycle rounds computed per model. For each cluster, the model with the highest average predicted template-modeling (pTM) score and best average predicted local distance difference test (pLDDT) score was selected for downstream evaluation.

Structural homology search

Only high-quality ($pTM \geq 0.7$) and medium-quality ($0.5 \leq pTM < 0.7$) structures were retained for further analysis. To assess structural similarity with experimentally determined structures, these models were searched using Foldseek [24] against 5653 CATH [25] (v4.4, released October 2024) to detect matches to structural domains, and PDB assemblies [26] (March 2024 release), to account for multi-domain models with matches to multimeric complexes, as well as any PDB structures not already incorporated in CATH. Hits were collected for all family structural models

with an alignment TM-score >0.5. For the remaining family models, query and target lengths were considered. If the query length was shorter than the target, the query alignment score was assessed with a 0.5 threshold. Conversely, for cases where the target length was shorter, a family was considered a hit if the target alignment length was at least 0.5. For the remaining families, an additional search against AlphaFoldDB [23, 27] was performed to identify models that lacked experimental structure but had been previously predicted by AlphaFold. For each hit identified in AlphaFoldDB, further annotation was carried out using metadata retrieved from the corresponding UniProt entries to determine whether the hits were functionally annotated or merely predicted.

Superfamily generation

To assess how many families can be grouped into superfamilies based on structural information, Foldseek clustering was performed in bidirectional mode with a coverage threshold of 80% and a TM-score threshold of 0.5.

Taxonomic annotation

The taxonomic analysis in this study focused on families containing over 100 members, adhering to criteria established in previous studies [8, 9] to ensure statistical robustness. Protein families were defined as clusters meeting this threshold, and their associated scaffolded sequences were retrieved from the IMG/M database. Taxonomic classification of these sequences began with Kraken2 [28] (version 2.1.3), a highly accurate tool for taxonomic identification. To ensure comprehensive coverage across all domains of life, we utilized the full RefSeq database (~1 TB) rather than the standard Kraken2 database. To enhance the annotation of proteins and minimize false positives (given the limitations of taxonomic algorithms like Kraken2 in handling eukaryotic sequences [29]), a supplementary analysis using MMseqs2 taxonomy against UniRef100 [30] (version 13) for eukaryotic and unclassified scaffolds was performed. For scaffolds that remained unclassified after this step, the Whokaryote tool [31] (version 1.1.2) was used to differentiate eukaryotic from prokaryotic contigs based on domain-specific gene structural features. Scaffolds still unclassified following Whokaryote analysis were further analyzed with EukRep [32] (version 0.6.7), a tool for detecting eukaryotic sequences in metagenomic and metatranscriptomic datasets. Lastly, geNomad [33] (version 1.8.1) was used to detect additional plasmid and viral sequences among the remaining unclassified scaffolds.

Biome distribution

A protein family distribution across various ecosystems was performed to enhance our understanding of habitat-specific diversity within these protein families. To explore this distribution across different biomes, metadata from each environmental sample was extracted from the IMG/M database and organized using the GOLD classification system [34].

In this study, the three primary GOLD ecosystems (environmental, host-associated, and engineered) were further divided into 10 specific ecosystem types for direct comparison and enhanced detail, consistent with the NMPFamsDB study. These 10 ecosystem types include air, freshwater, marine, other aquatic environments, engineered, human, mammal host-associated, other host-associated environments, plants, and terrestrial. Network visualizations were created using Gephi

[35], with the layout generated by the Yifan-Hu algorithm [36], and the geographic distribution was computed from each dataset’s geolocation metadata in IMG.

Gene neighborhood

The analysis focuses on gene families with >100 members. The majority of novel proteins are located close to known genes, suggesting that Pfam annotations of the known genes can provide valuable insights into the functions of the novel proteins. Specifically, 81 908 529 out of 87 393 063 scaffolds contain both novel and known genes, highlighting potential interactions or shared functions. To conduct the gene neighborhood analysis, scaffolds associated with families comprising >100 members were collected. To maximize data recovery, gene neighborhood analysis was performed across entire scaffolds. This approach aligns with previous studies [8] and is justified by the relatively short length of the majority of scaffolds in our dataset. However, to prevent the inclusion of distant, functionally unrelated genes, a supplementary analysis has been performed by limiting the search window to 10 ORFs flanking each side of the protein of interest. These scaffolds were then examined against Pfam annotations. For each scaffold, the top five most frequent Pfam domains were identified. The Pfam domains were associated with their respective Pfam descriptions. In addition, these domains were subsequently matched to Gene Ontology (GO) terms using the InterPro2GO mapping from InterPro. To simplify and interpret the results, GO subsets (commonly referred to as GO slims), condensed versions of GO that contain curated subsets of terms, were used. The GO terms were further grouped into the three primary GO categories: molecular function, cellular component, and biological process, enabling distinct, targeted analysis within each category.

Results

Novel families from metagenomes versus families from reference genomes

We identified novel protein sequences, the ones with no hits to Pfam [37], Antifam [38], or reference genomes. These sequences were subsequently clustered into families using MMseqs2 [39]. For direct comparison with known protein families, the same methodology was applied to sequences derived from reference genomes, analyzing them across various cluster sizes (≥ 25 , ≥ 50 , ≥ 75 , and ≥ 100 members) (Table 1).

Clustering analysis of reference genomes grouped 393 429 871 of 539 040 456 proteins (72%) into 2 229 233 clusters of ≥ 25 members across 155 857 genomes. In contrast, metagenomic and metatranscriptomic data clustered 370 185 087 of 1 736 930 453 proteins (21.3%) into 6 588 876 clusters from 44 198 datasets and 253 359 325 scaffolds, representing a 296% increase in protein families and quadrupling the total for this size threshold. At ≥ 50 members, 358 091 618 proteins (66.4%) from reference genomes formed 1 200 486 clusters, while 221 646 880 proteins (12.8%) from metagenomes grouped into 2 160 632 clusters, reflecting a 62% increase in clustered proteins and a 180% rise in protein families. For ≥ 75 members, 335 371 741 reference proteins (62.2%) formed 824 621 clusters, versus 155 259 616 metagenomic proteins (8.9%) in 1 052 590 clusters, showing a 46% increase in clustered proteins and 128% more families. At ≥ 100 members, reference genomes

Table 1. Direct comparison of protein clusters produced from reference genomes versus novel protein clusters generated from metagenomic and metatranscriptomic data at varying cluster sizes (≥ 25 , ≥ 50 , ≥ 75 , and ≥ 100 members)

Reference genomes				
Proteins for clustering	539 040 456			
Family size	≥ 25 members	≥ 50 members	≥ 75 members	≥ 100 members
Families	2 229 233	1 200 486	824 621	630 433
Number of genomes	155 857	154 072	153 253	152 506
Proteins	393 429 871	358 091 618	335 371 741	318 713 491
% Of proteins	72.99%	66.43%	62.22%	59.13%
Environmental samples				
Proteins for clustering	1 736 930 453			
Family size	≥ 25 members	≥ 50 members	≥ 75 members	≥ 100 members
Families	6 588 876	2 160 632	1 052 590	608 258
Number of metagenomes	44 198	43 701	43 315	42 975
Scaffolds	253 359 325	157 845 393	113 497 493	87 393 064
Proteins	370 185 087	221 646 880	155 259 616	117 330 466
% Of proteins	21.31%	12.76%	8.94%	6.76%
Comparison with reference genomes				
% Of family increase	296%	180%	128%	96%
% Of protein increase	94%	62%	46%	37%

contained 318 713 491 proteins (59.1%) in 630 433 clusters, while metagenomes grouped 117 330 466 proteins (6.8%) into 608 258 clusters, doubling the protein family space.

As clustering thresholds increased, reference genomes showed a minimal reduction in the number of clustered proteins but a notable decrease in the number of clusters, indicating more consolidated protein families. In contrast, metagenomic data exhibited a 30% reduction in proteins within larger clusters and a 50% drop in cluster count, suggesting that metagenomic data tend to form smaller, more fragmented clusters due to greater sequencing fragmentation. Further family properties of protein clusters from reference genomes and metagenomes are shown in [Supplementary File S2](#) and [Supplementary Fig. S1](#).

Biome distribution

Metagenomic and metatranscriptomic datasets from IMG/M [5] included 49 986 samples with geographic coordinates (latitude and longitude) and GOLD [34] ecosystem classifications (environmental, engineered, host-associated). Most samples were environmental (67.7%), followed by host-associated (23%) and engineered (9.2%). Focusing on microbial families with at least 100 members and at least 5% of samples originating from a single ecosystem type, we identified 608 258 families across 42 975 datasets: 67.5% environmental, 22.6% host-associated, and 9.7% engineered. Geographic coordinates for environmental samples were used to plot their origin locations on a world map (Fig. 2A), underscoring the global reach and diversity of sampled environments.

Of these families, 385 791 (63.4%) are restricted to a single ecosystem: primarily environmental (57.5%), with smaller contributions from host-associated (4.9%) and engineered (0.7%). Among these ecosystems, families exclusive to a single ecosystem include 89 380 (14.6%) in terrestrial, 122 073 (20%) in freshwater, 109 286 (17.9%) in marine, 5701 (0.93%) in other aquatic environments, 4383 (0.7%) in engineered, 14 614 (2.4%) in plants, 3801 (0.62%) in mammals, 9538 (1.56%) in humans, and 2271 (0.37%) in other host subtypes (Fig. 2B). Biome-specific novel protein families highlight previously unrecognized functions and sequence traits that may be adapted to their particular ecological niches (see the “Gene neighborhood” section).

Among families spanning more than one ecosystem, 185 994 (30.5%) exist in two ecosystems. Notable overlaps include 96 049 families (16%) shared between terrestrial and plants, and 50 232 families (8.2%) shared between freshwater and marine environments. Additionally, 9539 families (1.5%) are found in terrestrial and freshwater ecosystems, and 9200 families (1.5%) span engineered and freshwater environments. Functional overlap between environments can uncover common biological roles among diverse microbial communities. An UpSet plot visualizing all possible overlaps in detail is shown in [Supplementary Fig. S2](#).

Figure 2C presents a network representation illustrating habitat-specific families and families shared across multiple environments. In this network, gray nodes represent novel protein families with 100 or more members, while colored nodes denote the different habitats.

Gene neighborhood

The analysis of gene neighborhoods within novel protein families, particularly those specific to different environments, provides valuable functional insights into microbial activities across various ecosystems. To gain a more accurate understanding of functions, this study focuses on biome-specific families, aiming to identify the most frequent functions and those that are relatively unique compared to the 50 most common functions in each environment. Gene neighborhood analysis was performed using two complementary approaches: considering the entire scaffold and analyzing a window of 10 ORFs from each side. The key findings presented below were consistently observed and are supported by both approaches. (See [Supplementary Figs S3–S8](#) and [Supplementary Tables S2–S7](#)).

Focusing on human-associated specific protein families, the majority of the analyzed samples originate from the digestive system. Gene neighborhood analysis reveals that the MatE [40] (multi antimicrobial extrusion) family, a member of the multidrug/oligosaccharidyl-lipid/polysaccharide (MOP) flip-pase superfamily, is significantly more prevalent in this environment. Metagenomic studies indicate that genes related to MatE contribute to antimicrobial resistance in gut microbiota. In addition, the unique enrichment of AraC-family transcriptional regulators as gene neighbors in human-associated

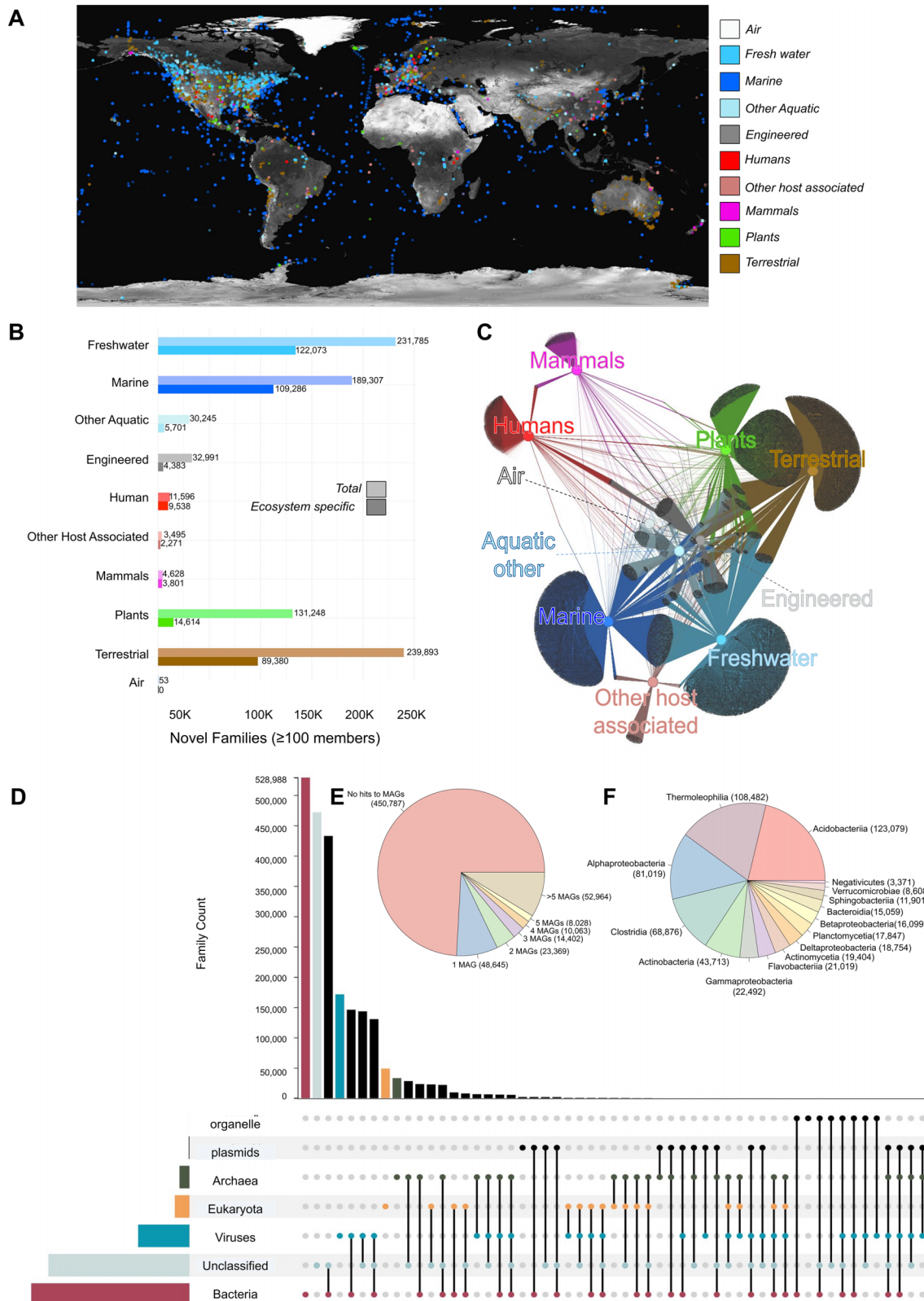


Figure 2. Ecosystem analysis of novel protein families with ≥ 100 members. **(A)** Visualization of sample geographical distribution. **(B)** Network representation of the protein clusters and their ecosystems. Ten ecosystem types were applied according to the GOLD ecosystem classification, represented by central, colored nodes (hubs), whereas the gray peripheral nodes represent the protein clusters. The edges represent the protein cluster–ecosystem associations. **(C)** The distribution of total versus ecosystem-type-specific novel protein families across the eight different ecosystem types. **(D)** Family taxonomy distribution across the three domains of life, viruses, plasmids, and organelles. **(E)** Distribution of protein hits against High-Quality Metagenome Assembled Genomes (HQ MAGs) within families. **(F)** The top 20 MAGs phyla that are represented across the families.

metagenomes implies that these novel protein families are functionally integrated into host-adaptive regulatory networks, potentially representing uncharacterized effectors of gut colonization and host–microbe interaction [41].

In plant-specific environments, microbes must adapt to unique ecological and physiological conditions, including interactions with plant hosts. The Plant mobile domain and PPR repeat are particularly relevant, as they play essential roles in RNA processing and gene regulation, often found in plant-associated microbes and organelles [42]. Additionally, the NB-ARC domain [43, 44], commonly associated with resistance genes, contributes to microbial adaptation and defense mechanisms in plant environments, highlighting the intricate microbial–plant interactions. Plant-associated novel protein families are also uniquely neighbored by retrotransposon-derived genes [45] (reverse transcriptase, RNase H, and gag-like structural proteins) reflecting the extraordinary proliferation of Ty1/Copia and Ty3/Gypsy LTR retrotransposons that can constitute up to 85% of plant genomes. These functions were identified as exclusively enriched in gene-neighbor protein families specific to biome-restricted plant families, providing interesting, valuable functional insights into previously uncharacterized protein families.

In freshwater ecosystems, hotspots for viral and phage activity, gene neighborhood analysis highlights the exclusive presence of viral-related proteins such as Caudovirus prohead serine proteases [46], phage portal protein [47], and Yqaj-like viral recombinases [48]. These proteins are crucial for viral replication, assembly, and horizontal gene transfer, emphasizing the central role of viruses in shaping microbial communities. Additionally, the microbial reliance on carbohydrate metabolism in freshwater environments is evident in the exclusive presence of proteins such as chitinases, glycosyltransferases, and transglycosylases [49]. These enzymes facilitate the degradation of plant matter, detritus, and chitinous material, supporting nutrient cycling and microbial adaptation in these nutrient-rich but variable habitats.

In terrestrial environments, gene neighborhood analysis identifies proteins with transaminase activity and carboxypeptidase regulatory-like domains as highly prevalent in soil ecosystems. These proteins are essential for amino acid metabolism, contributing to nitrogen cycling [50] by facilitating the synthesis, degradation, and transformation of amino acids and organic nitrogen compounds. Additionally, the significant presence of the tripartite tricarboxylate transporter (TTT) family receptor in these environments highlights its role in the uptake of organic acids from plant root exudates, decaying organic matter, and microbial metabolites. These organic acids serve as crucial energy sources and biosynthetic precursors, supporting microbial survival and activity in soil ecosystems.

Gene neighborhood analysis of novel protein families revealed a dual-layered functional architecture across biomes. While each biome harbored a distinct repertoire of unique protein families reflecting environment-specific adaptations and ecological pressures, a core set of functional categories, including membrane transport, DNA replication and repair, and transcriptional regulation, emerged consistently across all nine biomes examined. Crucially, however, these shared functions were not carried by the same protein families; instead, they were fulfilled by biome-specific gene families of independent evolutionary origin, as evidenced by the divergence in family composition despite functional convergence at the do-

main level. This pattern suggests that while the selective pressures governing fundamental biological processes are universal, the molecular solutions have arisen repeatedly and independently across distinct ecological contexts, a hallmark of convergent functional evolution.

These findings underscore that the functional landscape of the biosphere is far more conserved than its sequence landscape: biomes that are radically different in taxonomic composition and environmental conditions nonetheless converge on the same core functional repertoire, implying that ecological constraints shape function far more stringently than they shape sequence. Together, the combination of biome-unique families and convergently evolved shared functions positions gene neighborhood analysis as a powerful lens not only for characterizing environment-specific genomic innovations, but also for illuminating the universal functional grammar that underlies microbial life across the Earth's diverse ecosystems.

Taxonomy

Of the 87 393 064 scaffolds analyzed, 65 471 449 (74.92%) were successfully classified taxonomically, while 21 921 615 (25.08%) remained unclassified. Integrating multiple taxonomic algorithms in the analysis pipeline improved classification rates, reducing the proportion of unclassified scaffolds compared to the previous study [51]. Among the classified scaffolds, the majority were categorized as Bacteria (76.01%), followed by Viruses (14%), Eukaryota (7%), and Archaea (3%), with a small fraction classified as organelles and plasmids. The high classification rate achieved through multi-stage classification, utilizing a succession of taxonomic tools to resolve unassigned sequences, facilitates a deep exploration of diversity and the mapping of functions to specific taxonomic groups.

Using a threshold of at least 95% of samples derived from a single taxonomic domain, 115 377 families (18.9%) were taxonomically assigned to unique domains, while the majority encompassed scaffolds from multiple taxonomic groups. Specifically, 72 590 families (12%) were exclusively bacterial, 23 782 (4%) were solely eukaryotic, 10 547 (2%) were viral, and 3124 (1%) were archaeal. Most families displayed mixed taxonomic compositions, including combinations such as Bacteria and Unclassified (360 610 families, 59.2%), Bacteria and Viruses (105 338 families, 17.3%), and Viruses and Unclassified (30 355 families, 4.99%) (Fig. 2D).

Proteins from families that matched the 21 817 high-quality MAGs were used for taxonomic analysis. Since most MAGs are taxonomically classified, their taxonomy provides valuable insights into family taxonomy. Most families (74%) show no hits against MAGs, while the remaining families exhibit hits in descending order relative to the number of MAGs, indicating that more families have hits in fewer MAGs, while the number of families decreases as the number of MAGs with hits increases (Fig. 2E). Of the MAGs, 90% are bacterial, 6% remain unclassified, and 3% are archaeal. A taxonomic analysis of families based on MAGs identifies the top 20 bacterial phyla most commonly represented across these families (Fig. 2F).

Structural distribution

Focusing on protein families with >100 members, 3D structures were predicted using AlphaFold2/ColabFold [22, 27], following the workflow presented in Fig. 3A. Of the 608 258

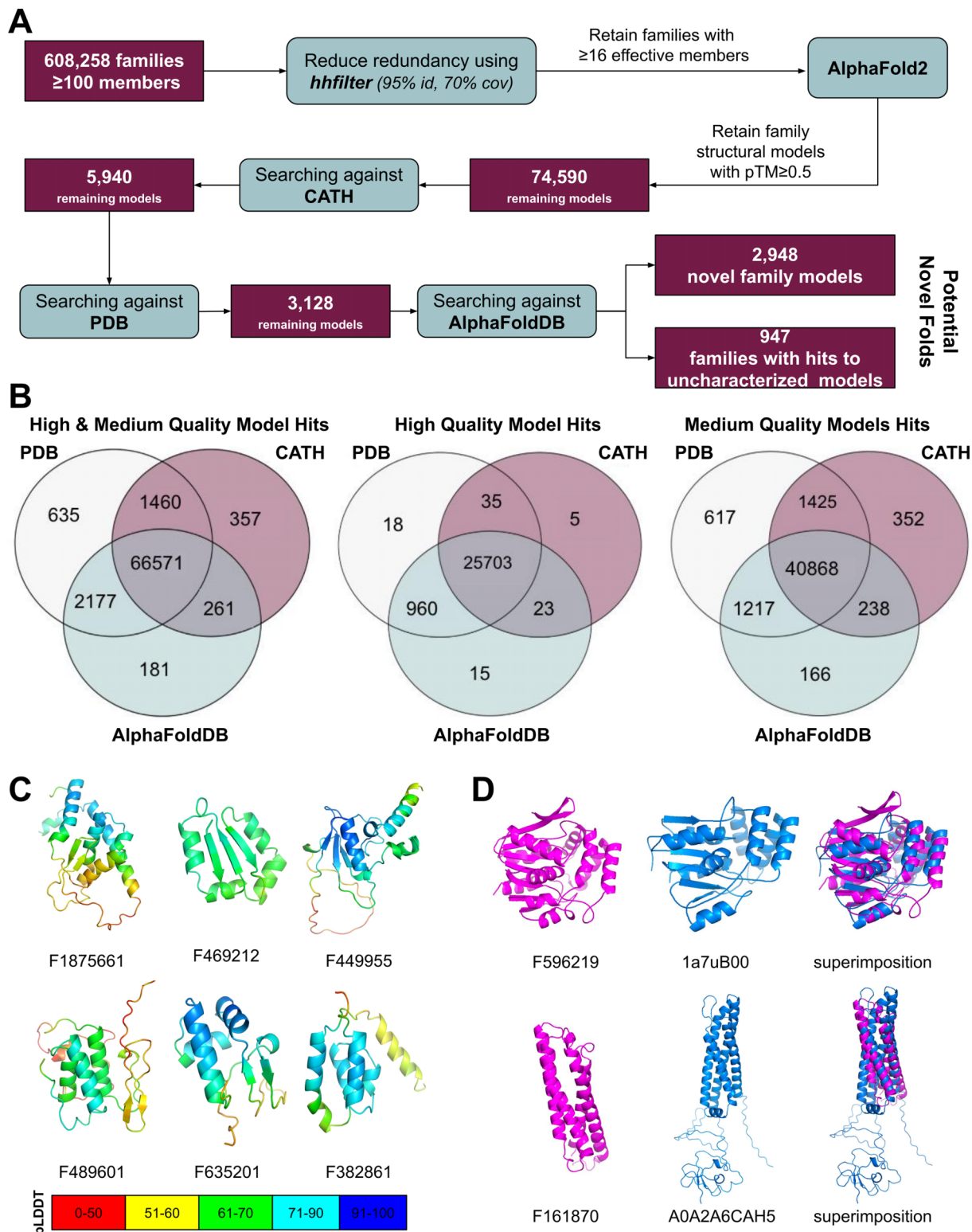


Figure 3. (A) Pipeline illustrating the preparation, structural prediction, and structural search using Foldseek against CATH, PDB, and AlphaFoldDB, highlighting the number of potential novel folds and unannotated structural hits. **(B)** Venn diagrams for high ($pTM \geq 0.7$), medium ($0.5 \leq pTM < 0.7$), and combined quality model structural hits against CATH, PDB, and AlphaFoldDB, highlighting the overlapping and unique hits for each database. **(C)** Example models with no significant hits to CATH, PDB, or AlphaFoldDB, considered as potential novel folds. pLDDT, per-residue confidence score. **(D)** Example models with hits to PDB and AlphaFoldDB domains.

families, representative structural models were successfully generated for 156 609. For each family, the model with the highest average pLDDT score was retained. Among these, 26 767 (17.09%) models had a pTM score >0.7, indicating high-quality models, while 47 823 (30.54%) had a pTM score between 0.5 and 0.7, indicating medium-quality models. This part of the analysis focuses on the 74 590 family models that fall into these two quality categories.

To provide insights into functional annotation and explore the novel structural space, these models were searched against well-known structural databases containing both experimentally validated and computationally predicted 3D models. Searches were conducted against CATH [25], PDB [26], and AlphaFoldDB [35] (Fig. 3B). The total number of hits for the 74 590 family models included 68 649 hits from CATH, 70 843 from PDB, and 69 190 from AlphaFoldDB. Specifically, searching against CATH yielded 68 649 family model hits, with 25 766 originating from high-quality models and 42 883 from medium-quality models. The remaining unmatched structures were then searched against the PDB, yielding 2812 additional hits (978 high-quality models and 1834 medium-quality models). Finally, a search against AlphaFoldDB identified 180 more hits (15 high-quality and 165 medium quality). An example of structural annotated folds is presented in Fig. 3D. Clustering with Foldseek produced 2781 singletons and 69 multi-member clusters, indicating most folds are unique. Further structural clustering of the whole dataset suggests these models represent genuinely novel partial or complete folds, rather than variations of known architectures.

These findings enhance functional annotation by leveraging structurally characterized proteins and uncover 2948 potential novel folds (Fig. 3C). Notably, 947 family models exhibit structural similarities to entries in AlphaFoldDB, yet their corresponding structures remain uncharacterized. This not only expands the pool of potential novel folds, as many remain unidentified, but also presents an opportunity to infer functional annotations for these previously uncharacterized structures.

Functional annotation through structural modeling

Structural annotations were performed based on structural homolog hits in CATH, PDB, and AlphaFoldDB. Out of the 26 767 high-quality structural models, 10 493 were derived from habitat-specific families, including 2668 from aquatic marine environments, 5290 from aquatic freshwater environments, 217 from other aquatic habitats, 22 from engineered environments, 8 from host-plants, 323 from mammals, 15 from humans, 4 from other host-associated environments, and 1946 from terrestrial environments. Out of 47 823 medium-quality structural models, 21 460 were derived from habitat-specific families. This includes 6088 from marine environments, 10 029 from freshwater environments, 369 from other aquatic environments, 82 from engineered environments, 131 from host plants, 318 from mammals, 48 from humans, 23 from other host-associated environments, and 4372 specific to terrestrial habitats.

For each family, the predicted fold was compared with the structural hits from these databases and combined with the gene neighborhood analysis. The F139409 family for example, which is specific to the human habitat, exhibits structural similarity to the winged helix DNA-binding do-

main (CATH: 1.10.10.10). Gene neighborhood analysis indicates that its neighboring genes predominantly encode helix–turn–helix motifs (PF01381), ERF superfamily domains (PF04404), and single-nucleotide strand DNA binding protein domains (PF00436), all of which are known for their roles in DNA binding. The helix–turn–helix domain is a well-characterized structural motif involved in DNA recognition, while the ERF superfamily includes single-strand annealing proteins. The presence of exclusively DNA-binding proteins in this gene neighborhood suggests a strong functional connection to DNA interactions (Supplementary Fig. S9A). Similarly, the F004558 family, which consists of proteins found exclusively in terrestrial environments, exhibits high structural similarity (as indicated by TM scores) to the CRAL-TRIO lipid-binding domain (CATH: 3.40.525.10). The CRAL-TRIO domain is primarily associated with lipid transport and the binding of small lipophilic molecules. Gene neighborhood analysis of the F004558 family reveals a strong correlation with membrane transport functions, with neighboring genes encoding ion channels (PF07885), sodium/calcium exchangers (PF01699), and transmembrane transport proteins (GO:0055085). Furthermore, this family is closely related to the SLOG superfamily (PF18306), which includes proteins that bind low-molecular-weight biomolecules. Notably, SLOG proteins and CRAL-TRIO domains share functional similarities as small-molecule-binding domains. This relationship extends to lipid transfer and membrane interactions, particularly in the context of membrane trafficking [36], further reinforcing their functional link in cellular transport mechanisms (Supplementary Fig. S9B).

Additionally, to focus on high-quality models, the 50 most frequent structural hits in each environment were selected, and the unique domains identified in each environment were reported (Supplementary Table S8). This analysis, along with the gene neighborhood analysis, does not pinpoint the uniquely structured domains for each habitat, but rather highlights the unique domains among the most frequently encountered ones. Interestingly, the most frequent and characteristic gene neighborhood functions presented in Supplementary Tables S2–S4 are closely related to the habitat-specific unique structural domains. For instance, the carboxypeptidase-like regulatory domain (CATH: 2.60.40.1120) is found in families specific to the terrestrial habitat, and this is further supported by neighborhood analysis, which identifies it as one of the most frequently associated functions (Pfam domain: PF13620). Another example is the STAS domain, which is strongly associated with regulation and signal transduction in bacterial systems [52]. This domain appears at the top of the list in terrestrial-specific families (CATH: 3.30.750.24). At the same time, the gene neighborhood analysis reveals that the top Pfam domains are related to bacterial regulatory proteins (Pfam: PF00196). Considering all of the above, structural homology not only gives insights into annotating novel protein families but also underscores the significance of gene neighborhood analysis, which plays a crucial role in characterizing novel or unannotated folds.

Discussion

The microbial world contains vast, largely unexplored protein diversity that underpins microbiome function and ecological interactions. By analyzing tens of billions of sequences from metagenomes, metatranscriptomes, and reference genomes,

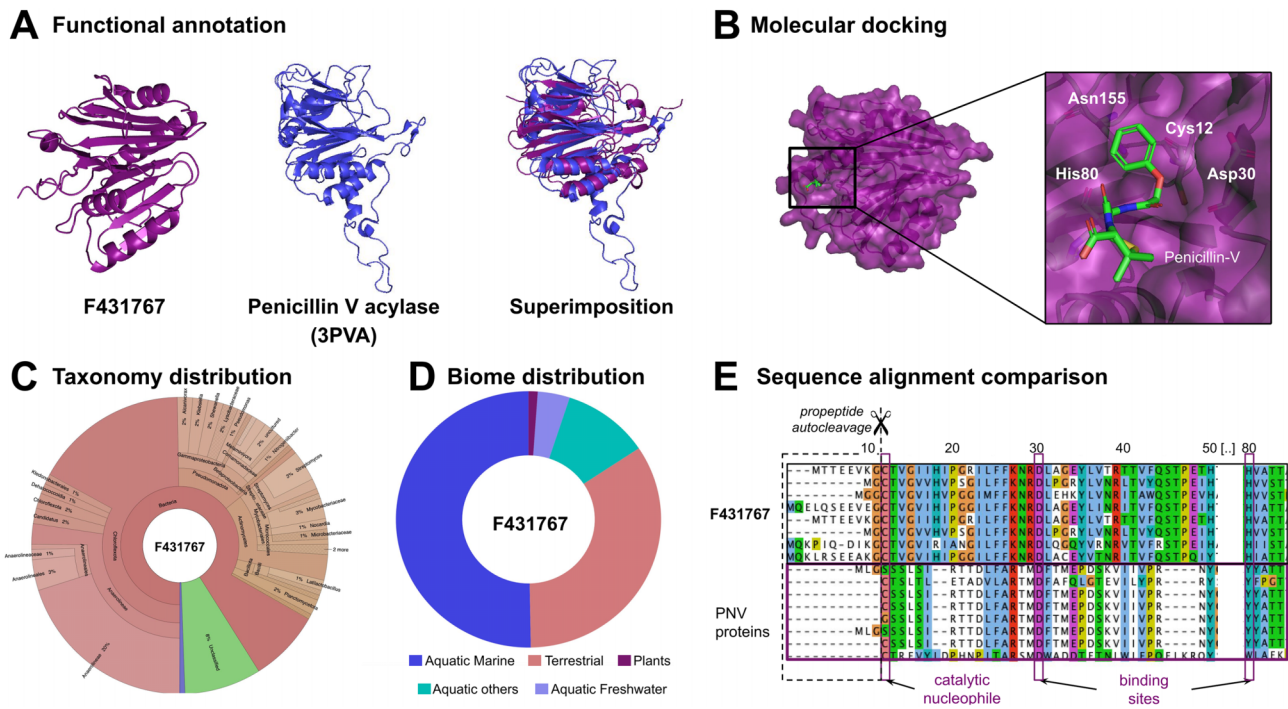


Figure 4. (A) Structural model of the novel protein family F431767 and its foldseek hit, penicillin V acylase (PDB code: 3PVA), including their structural superimposition. (B) Molecular docking by manual fitting of penicillin V in the F431767 structural model. Key residues conserved between penicillin V acylase and F431767 are highlighted. (C) KRONA visualization showing the taxonomic distribution of proteins belonging to the F431767 family. Members are predominantly found in Bacteria, including Pseudomonadota, Actinomycetota, and Chloroflexota. Chloroflexota includes aerobic thermophiles that utilize oxygen and thrive at high temperatures. (D) Pie chart illustrating the biome distribution of F431767 family members, with most sequences originating from aquatic marine and terrestrial environments, followed by aquatic freshwater, other aquatic environments, and plants. (E) Sequence alignment comparing penicillin V acylase PDB structures (2IWM, 2OQC, 2PVA, 2QUY, 2Z71, 3MJ1, 3PVA, 4WL2) with members of the F431767 family to identify conserved and divergent residues. Key residues such as Cys12, the catalytic N-terminal nucleophile, are highly conserved, while Asp30 is also conserved and contributes to the penicillin V binding site.

we substantially expand protein family space, identifying over 600 000 families with ≥ 100 members and 6.5 million families with ≥ 25 members, effectively quadrupling current coverage. Importantly, smaller and niche families capture rare proteins with potential functional and evolutionary significance. Our scalable framework integrates structural prediction with taxonomic, ecological, and genomic context, revealing biome-specific patterns of adaptation and uncovering both novel folds and known folds within previously uncharacterized sequence space. Beyond novel folds, we identify protein families that adopt established structural architectures but lack detectable sequence similarity to known domains.

An insightful example (protein family F431767) exemplifies this structure-guided functional inference (Fig. 4). Although no similarity to Pfam domains or isolated genomes is detected, members of this family exhibit clear structural similarity to penicillin V acylase (PVA) chain A (CATH: 3.60.60.10; PDB: 3PVA) (Fig. 4A). PVA catalyzes the hydrolysis of penicillin V to 6-aminopenicillanic acid (6-APA), a key intermediate in the semi-synthesis of penicillin products in industry [53, 54]. F431767 proteins are distributed primarily across marine and terrestrial environments and occur mainly within Pseudomonadota, Actinomycetota, and Chloroflexota (Fig. 4C and D). The presence of this family in Chloroflexota [55], an ecologically and metabolically versatile phylum, highlights its potential functional breadth and biotechnological relevance. Sequence comparisons with structurally characterized PVAs, guided by the penicillin V-bound structure of

2Z71, reveal conservation of catalytic features (Fig. 4E), including the N-terminal nucleophile Cys12 generated by autocatalytic self-cleavage, consistent with the N-terminal nucleophile (Ntn) hydrolase superfamily [56]. Additional conserved residues (Asp30 and Asn155) contribute to substrate binding, while substitution of Tyr80 with His80 in F431767 members suggests adaptive tuning, potentially enhancing electrostatic flexibility and stability in environmentally diverse or extreme habitats. Molecular docking confirms the penicillin V model potential binding (Fig. 4B).

Recent studies have leveraged metagenomic datasets to expand and characterize the global landscape of microbial protein diversity, focusing on taxonomic distribution, functional potential, and the discovery of novel sequence space (Supplementary Tables S9–S13). Comparisons of our dataset with MAGs from IMG/M and gcMeta [57] revealed only limited overlap (Supplementary Tables S14 and S15). Although SPIRE [58] substantially expanded the known protein space by organizing metagenomic proteins across taxa and biomes, it showed no overlap with the proteins identified in our study. These findings demonstrate that our identified proteins represent novel sequence space not only relative to reference genomes but also relative to previous metagenomic efforts aimed at characterizing uncultivated microbial communities.

The Global Microbial Gene Catalogue (GMGC) [59] organized metagenomic datasets into protein families to investigate the distribution of genes across diverse habitats. Comparison with GMGC revealed limited overlap with our dataset,

and none of the overlapping proteins received functional annotations. These findings highlight the significance of our study, which extends beyond the identification of novel protein families by providing a multilayered annotation framework to characterize their potential functions, taxonomic associations, and biological relevance.

Previous studies investigating novel protein family space, including FesNOV, NMPFam v1, and AGNOSTOS, employed distinct methodological approaches (Supplementary Table S16) and showed limited overlap with our dataset (Supplementary File S3 and Supplementary Table S17), highlighting the complementary nature of these efforts and the expansion of unexplored protein diversity.

Overall, this study shows that integrating global metagenomic clustering with structural and contextual annotation converts billions of uncharacterized sequences into a biome-resolved protein family landscape, providing new insights into microbial adaptation and revealing a broad reservoir of functional and biotechnological potential.

Limitations

All data used in this analysis were retrieved from IMG/M, where gene prediction is performed using established tools, including Prodigal v2.6.3 and GeneMarkS-2. Most eukaryotic gene prediction methods either rely on similarity to known proteins (e.g. MetaEuk) or require prior taxonomic classification to optimize *ab initio* models (e.g. AUGUSTUS). Consequently, prokaryotic gene callers are often the only practical option. However, this introduces a potential bias, as genes of eukaryotic origin may be missed or inaccurately predicted. Therefore, the quality and completeness of the predicted protein set are inherently dependent on the underlying gene-calling strategy. Protein sequence fragmentation is a major challenge in metagenomic studies. Despite the implementation of filtering strategies to reduce its impact, workflows for novel protein identification can still be biased toward short protein families. While the present computational framework provides a valuable resource for the discovery, annotation, and characterization of novel proteins, all functional and biological inferences remain computational predictions and should be considered hypotheses requiring experimental validation. While our 35-amino-acid threshold aligns with Pfam standards for functional domains, our strict novelty criteria naturally enriched the dataset for smaller sequences. Omitting coverage parameters during Pfam filtering intentionally excluded multidomain proteins to eliminate false positives, while high bidirectional coverage thresholds against larger reference proteins further constrained sequence length. Although these conservative filters may omit longer sequences containing isolated novel domains, they ensure that the retained candidates represent genuinely novel proteins, leaving multidomain protein discovery as a promising direction for future research.

Acknowledgements

This work was supported by computational resources provided by the ELIXIR-GR HYPATIA Cloud infrastructure.

Author contributions: Eleni Aplakidou: Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. Fotis A. Baltoumas: Data curation, Formal analysis, Methodology, Supervision, Validation, Soft-

ware, Writing – original draft, Writing – review & editing. Maria N. Chasapi: Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing. Efthimia Lamari: Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. Iro N. Chasapi: Formal analysis, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. Ilias Georgakopoulos-Soares: Methodology, Project administration, Resources, Software, Writing – original draft, Writing – review & editing. Evangelos Karatzas: Data curation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. Ioannis Iliopoulos: Data curation, Methodology, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing. Aydin Buluç: Data curation, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing. Robert D. Finn: Data curation, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing. Antonio Pedro Camargo: Data curation, Methodology, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing. Nikos C. Kyrpides: Data curation, Funding acquisition, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the ‘Third Call for H.F.R.I. Research Projects to support faculty members and researchers’ (23592-EMISSION); Hellenic Foundation for Research and Innovation (H.F.R.I.) under the ‘4th Call for H.F.R.I. Research Project to support Postdoctoral Researchers’ (28787-VIROMINE); Fondation Santé; ARISE program from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant agreement No 945405; Startup funds from the Penn State College of Medicine and by the Huck Innovative and Transformational Seed Fund (HITS) award from the Huck Institutes of the Life Sciences at Penn State University; Startup funds from the University of Texas at Austin; US Department of Energy Joint Genome Institute (<https://ror.org/04xm1d337>); US Department of Energy Office of Science user facilities, operated under contract no. DE-AC02-05CH11231; and Applied Mathematics program of the DOE Office of Advanced Scientific Computing Research (DE-AC02-05CH11231). E.A. received a short-term EMBO fellowship (12092) to collaborate closer to the NKI research team. Funding to pay the Open Access publication charges for this article was provided by the US Department of Energy Joint Genome Institute (<https://ror.org/04xm1d337>) and US Department of Energy Office of Science user facilities, operated under contract no. DE-AC02-05CH11231.

Data availability

Both novel families from metagenomes (minimum of 25 members) and families from reference genomes and medium- to high-quality predicted structures from novel families (at least 100 members) can be found at <https://zenodo.org/records/17225887>. The code is available in Zenodo at <https://zenodo.org/records/22253220>.

Code availability: Scripts and Unix commands used in this study are available at <https://github.com/PavlopoulosLab/NMPFamsDB2>.

References

- Baltoumas FA, Karatzas E, Paez-Espino D *et al.* Exploring microbial functional biodiversity at the protein family level—from metagenomic sequence reads to annotated protein clusters. *Front Bioinform* 2023;3:1157956. <https://doi.org/10.3389/fbinf.2023.1157956>
- Aplakidou E, Vergoulidis N, Chasapi M *et al.* Visualizing metagenomic and metatranscriptomic data: a comprehensive review. *Comput Struct Biotechnol J* 2024;23:2011–33. <https://doi.org/10.1016/j.csbj.2024.04.060>
- Rodríguez del Río Á, Giner-Lamia J, Cantalapiedra CP *et al.* Functional and evolutionary significance of unknown genes from uncultivated taxa. *Nature* 2024;626:377–84.
- Kim CY, Podlesny D, Schiller J *et al.* Planetary microbiome structure and generalist-driven gene flow across disparate habitats. *Cell* 2026;189:2073–91.e21. <https://doi.org/10.1016/j.cell.2025.12.051>
- Chen I-MA, Chu K, Palaniappan K *et al.* The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res* 2023;51:D723–32. <https://doi.org/10.1093/nar/gkac976>
- Hyatt D, Chen G-L, LoCascio PF *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119. <https://doi.org/10.1186/1471-2105-11-119>
- Vanni C, Schechter MS, Acinas SG *et al.* Unifying the known and unknown microbial coding sequence space. *eLife* 2022;11:e67667. <https://doi.org/10.7554/eLife.67667>
- Pavlopoulos GA, Baltoumas FA, Liu S *et al.* Unraveling the functional dark matter through global metagenomics. *Nature* 2023;622:594–602. <https://doi.org/10.1038/s41586-023-06583-7>
- Baltoumas FA, Karatzas E, Liu S *et al.* NMPFamsDB: a database of novel protein families from microbial metagenomes and metatranscriptomes. *Nucleic Acids Res* 2024;52:D502–12. <https://doi.org/10.1093/nar/gkad800>
- Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 2001;29:2607–18. <https://doi.org/10.1093/nar/29.12.2607>
- Frith MC. Gentle masking of low-complexity sequences improves homology search. *PLoS One* 2011;6:e28819. <https://doi.org/10.1371/journal.pone.0028819>
- Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol* 2011;7:e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>
- Frith MC, Wan R. Incorporating sequence quality data into alignment improves DNA read mapping. *Nucleic Acids Res* 2010;38:e100. <https://doi.org/10.1093/nar/gkq010>
- Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999;12:85–94. <https://doi.org/10.1093/protein/12.2.85>
- Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;12:59–60. <https://doi.org/10.1038/nmeth.3176>
- Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;35:1026–8. <https://doi.org/10.1038/nbt.3988>
- The Genome Standards Consortium, Bowers RM, Kyrpides NC, Stepanauskas R *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017;35:725–31. <https://doi.org/10.1038/nbt.3893>
- Parks DH, Imelfort M, Skennerton CT *et al.* CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25:1043–55. <https://doi.org/10.1101/gr.186072.114>
- Clum A, Huntemann M, Bushnell B *et al.* DOE JGI Metagenome workflow. *mSystems* 2021;6:e00804–20. <https://doi.org/10.1128/mSystems.00804-20>
- Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;30:772–80. <https://doi.org/10.1093/molbev/mst010>
- Steinegger M, Meier M, Mirdita M *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 2019;20:473. <https://doi.org/10.1186/s12859-019-3019-7>
- Mirdita M, Schütze K, Moriawaki Y *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* 2022;19:679–82. <https://doi.org/10.1038/s41592-022-01488-1>
- Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;596:583–9. <https://doi.org/10.1038/s41586-021-03819-2>
- van Kempen M, Kim SS, Tumescheit C *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* 2024;42:243–6.
- Knudsen M, Wiuf C. The CATH database. *Hum Genomics* 2010;4:207–12. <https://doi.org/10.1186/1479-7364-4-3-207>
- Berman HM. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–42. <https://doi.org/10.1093/nar/28.1.235>
- Varadi M, Bertoni D, Magana P *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* 2024;52:D368–75. <https://doi.org/10.1093/nar/gkad1011>
- Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;15:R46. <https://doi.org/10.1186/gb-2014-15-3-r46>
- Belliardo C, Koutsovoulos GD, Rancurel C *et al.* Improvement of eukaryotic protein predictions from soil metagenomes. *Sci Data* 2022;9:311. <https://doi.org/10.1038/s41597-022-01420-4>
- Mirdita M, Steinegger M, Breitwieser F *et al.* Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* 2021;37:3029–31. <https://doi.org/10.1093/bioinformatics/btab184>
- Pronk LJ, Medema MH. Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. *Microb Genom* 2022;8:mgen000823 b. <https://doi.org/10.1099/mgen.0.000823>
- West PT, Probst AJ, Grigoriev IV *et al.* Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res* 2018;28:569–80. <https://doi.org/10.1101/gr.228429.117>
- Fast and accurate identification of plasmids and viruses in sequencing data using geNomad. *Nat Biotechnol* 2023;42:1196–7. <https://doi.org/10.1038/s41587-023-01982-7>
- Mukherjee S, Stamatis D, Bertsch J *et al.* Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids Res* 2021;49:D723–33. <https://doi.org/10.1093/nar/gkaa983>
- Piovesan D, Monzon AM, Tosatto SCE. Intrinsic protein disorder and conditional folding in AlphaFoldDB. *Protein Sci* 2022;31:e4466. <https://doi.org/10.1002/pro.4466>
- Overduin M, Tran A, Eekels D *et al.* Transmembrane membrane readers form a novel class of proteins that include peripheral phosphoinositide recognition domains and viral spikes.

- Membranes* 2022;12:1161.
<https://doi.org/10.3390/membranes12111161>
37. Paysan-Lafosse T, Andreeva A, Blum M *et al.* The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res* 2025;53:D523–34. <https://doi.org/10.1093/nar/gkae997>
 38. Eberhardt RY, Haft DH, Punta M *et al.* AntiFam: a tool to help identify spurious ORFs in protein annotation. *Database* 2012;2012:bas003–. <https://doi.org/10.1093/database/bas003>
 39. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;35:1026–8. <https://doi.org/10.1038/nbt.3988>
 40. Baron SA, Diene SM, Rolain J-M. Human microbiomes and antibiotic resistance. *Hum Microb J* 2018;10:43–52. <https://doi.org/10.1016/j.humic.2018.08.005>
 41. Laville E, Perrier J, Bejar N *et al.* Investigating host microbiota relationships through functional metagenomics. *Front Microbiol* 2019;10:1286. <https://doi.org/10.3389/fmicb.2019.01286>
 42. Nicolau M, Picault N, Descombin J *et al.* The plant mobile domain proteins MAIN and MAIL1 interact with the phosphatase PP7L to regulate gene expression and silence transposable elements in *Arabidopsis thaliana*. *PLoS Genet* 2020;16:e1008324. <https://doi.org/10.1371/journal.pgen.1008324>
 43. Wang Y, Tan B-C. Pentatricopeptide repeat proteins in plants: cellular functions, action mechanisms, and potential applications. *Plant Commun* 2025;6:101203. <https://doi.org/10.1016/j.xplc.2024.101203>
 44. Van Ooijen G, Mayr G, Kasiem MMA *et al.* Structure–function analysis of the NB-ARC domain of plant disease resistance proteins. *J Exp Bot* 2008;59:1383–97.
 45. Orozco-Arias S, Isaza G, Guyot R. Retrotransposons in plant genomes: structure, identification, and classification through bioinformatics and machine learning. *Int J Mol Sci* 2019;20:3837. <https://doi.org/10.3390/ijms20153837>
 46. Fokine A, Rossmann MG. Common evolutionary origin of procapsid proteases, phage tail tubes, and tubes of bacterial type VI secretion systems. *Structure* 2016;24:1928–35. <https://doi.org/10.1016/j.str.2016.08.013>
 47. Gu Z, Wu K, Wang J. Structural morphing in the viral portal vertex of bacteriophage lambda. *J Virol* 2024;98:e0006824. <https://doi.org/10.1128/jvi.00068-24>
 48. Zhou Z, Tran PQ, Martin C *et al.* Unravelling viral ecology and evolution over 20 years in a freshwater lake. *Nat Microbiol* 2025;10:231–45. <https://doi.org/10.1038/s41564-024-01876-7>
 49. Rajput V, Samson R, Yadav R *et al.* Metagenomic mining of Indian river confluence reveal functional microbial community with lignocellulolytic potential. *3 Biotech* 2022;12:132. <https://doi.org/10.1007/s13205-022-03190-7>
 50. Mooshammer M, Wanek W, Hämmerle I *et al.* Adjustment of microbial nitrogen use efficiency to carbon:nitrogen imbalances regulates soil nitrogen cycling. *Nat Commun* 2014;5:3694. <https://doi.org/10.1038/ncomms4694>
 51. Pavlopoulos GA, Baltoumas FA, Liu S *et al.* Unraveling the functional dark matter through global metagenomics. *Nature* 2023;622:594–602. <https://doi.org/10.1038/s41586-023-06583-7>
 52. Moy BE, Seshu J. STAS domain only proteins in bacterial gene regulation. *Front Cell Infect Microbiol* 2021;11:679982. <https://doi.org/10.3389/fcimb.2021.679982>
 53. Torres R, Ramón F, de la Mata I *et al.* Enhanced production of penicillin V acylase from *Streptomyces lavendulae*. *Appl Microbiol Biotechnol* 1999;53:81–4. <https://doi.org/10.1007/s002530051618>
 54. Shewale JG, Sudhakaran VK. Penicillin V acylase: its potential in the production of 6-aminopenicillanic acid. *Enzyme Microb Technol* 1997;20:402–10. [https://doi.org/10.1016/S0141-0229\(96\)00176-7](https://doi.org/10.1016/S0141-0229(96)00176-7)
 55. Freches A, Fradinho JC. The biotechnological potential of the chloroflexota phylum. *Appl Environ Microbiol* 2024;90:e01756–23. <https://doi.org/10.1128/aem.01756-23>
 56. Oinonen C, Rouvinen J. Structural comparison of ntn-hydrolases. *Protein Sci* 2000;9:2329–37. <https://doi.org/10.1110/ps.9.12.2329>
 57. Sun Y, Chen Q, Fan G *et al.* gcMeta 2025: a global repository of metagenome-assembled genomes enabling cross-ecosystem microbial discovery and function research. *Nucleic Acids Res* 2026;54:D724–33. <https://doi.org/10.1093/nar/gkaf1115>
 58. Schmidt TSB, Fullam A, Ferretti P *et al.* SPIRE: a searchable, planetary-scale microbiome REsource. *Nucleic Acids Res* 2024;52:D777–83. <https://doi.org/10.1093/nar/gkad943>
 59. Coelho LP, Alves R, del Río ÁR *et al.* Towards the biogeography of prokaryotic genes. *Nature* 2022;601:252–6. <https://doi.org/10.1038/s41586-021-04233-4>