EcoFoldDB: Protein structure-guided functional profiling of ecologically relevant microbial traits at the metagenome scale

Timothy M. Ghaly[1,2]*, Vaheesan Rajabal[1,2], Dylan Russell[1,2], Elena Colombi[1,2], Sasha G. Tetu[1,2]*

[1]School of Natural Sciences, Macquarie University, Australia

[2]ARC Centre of Excellence in Synthetic Biology, Australia

*Corresponding Authors:

Timothy M. Ghaly: timothy.ghaly@mq.edu.au

Sasha G. Tetu: sasha.tetu@mq.edu.au

Running title: Structure-based functional profiling of metagenomes

Keywords: Protein folding; protein function; protein homology detection; structural bioinformatics; metagenomic annotations; biogeochemical cycling; plant growth-promotion

Abstract

Microbial communities are fundamental to planetary health and ecosystem processes. High-throughput metagenomic sequencing has provided unprecedented insights into the structure and function of these communities. However, functionally profiling metagenomes remains constrained due to the limited sensitivity of existing sequence homology-based methods to annotate evolutionarily divergent genes. Protein structure, more conserved than sequence, and intrinsically tied to molecular function, offers a solution. Capitalising on recent breakthroughs in structural bioinformatics, we present EcoFoldDB, a database of protein structures curated for ecologically relevant microbial traits, and its companion pipeline, *EcoFoldDB-annotate*, which leverages Foldseek with the ProstT5 protein language model for rapid structural homology searching directly from sequence data. *EcoFoldDB-annotate* outperforms state-of-the-art sequence-based methods in annotating metagenomic proteins, in terms of sensitivity and precision. To demonstrate its utility and scalability, we performed structure-guided functional profiling of 32 million proteins encoded by 8,000 high-quality metagenome-assembled genomes from the global soil microbiome. *EcoFoldDB-annotate* could resolve the phylogenetic partitioning of important nitrogen cycling pathways, from taxonomically restricted nitrifiers to more widespread denitrifiers, as well as identifying novel, uncultivated bacterial taxa enriched in plant growth-promoting traits. We anticipate that EcoFoldDB will enable researchers to extract ecological insights from environmental genomes and metagenomes, and accelerate discoveries in microbial ecology.

Introduction

Microbial community functioning shapes Earth system health, influencing global carbon dynamics, driving essential nutrient cycles and ecosystem services, and underpinning the biodiversity of entire biomes (Wardle et al., 2004; Gillings and Paulsen, 2014; Steffen et al., 2015; Cavicchioli et al., 2019). Deciphering the complexity of microbial communities, including their composition, interactions, and functions, is critical for understanding ecosystem processes and their responses to environmental change. High-throughput shotgun metagenomic sequencing has revolutionised the field of microbial ecology, allowing a culture-independent view of the structure and function of microbiomes (Huttenhower et al., 2012; Sunagawa et al., 2015; Bahram et al., 2018; Bahram et al., 2022), and facilitating unprecedented insights into their ecology and evolution at scales ranging from individual strains to whole communities (Rohwer et al., 2025; Zhou et al., 2025).

Functional annotation of microbial genes represents a cornerstone of metagenomic analysis. Nevertheless, this task represents a formidable challenge due to the extensive size of metagenomic datasets, and the massive functional diversity encoded within metagenomes (Pavlopoulos et al., 2023). Several bioinformatic tools exist to enable large-scale functional profiling, e.g., InterProScan (Jones et al., 2014), eggNOG-mapper (Cantalapiedra et al., 2021), KofamScan (Aramaki et al., 2019). However, these tools prioritise exhaustive functional identification over ecological interpretation. Researchers thus often need to manually screen annotation results to gain meaningful insights into the ecological functions of microbial communities. This has led to the development of ecology-focused tools for functional profiling, such as METABOLIC (Zhou et al., 2022) and DRAM (Shaffer et al., 2020), which curate annotations into ecologically coherent functional categories, enabling user-friendly interpretation and analysis. However, existing functional annotation tools rely

on sequence homology-based methods (e.g., MMSeqs2 (Steinegger and Söding, 2017), DIAMOND (Buchfink et al., 2021), or HMMER (Finn et al., 2011)). Consequently, annotation of metagenomes remains constrained since they are overwhelmingly composed of genes from uncultivated microbial taxa, and are enriched with proteins lacking detectable sequence homology to reference databases (Vanni et al., 2022; Pavlopoulos et al., 2023).

The structure of proteins, which are inherently linked to their molecular function, allows homology discovery beyond the limits of sequence similarity detection, as structure evolves much slower than sequence (Illergård et al., 2009). However, until recent breakthroughs in structural bioinformatics, the use of protein structure for large-scale functional profiling has not been feasible. Notably, AlphaFold2 revolutionised protein structure predictions, achieving near-experimental accuracy from sequence data alone (Jumper et al., 2021). Yet, identifying similar protein structures at scale remained a significant challenge. Foldseek addressed this limitation by enabling protein structural homology searching at speeds comparable to sequence searching (van Kempen et al., 2024). Foldseek achieves this by mapping the 3D coordinates for each residue in a protein structure to one of the 20 states in their 3D interaction (3Di) alphabet. Thus, Foldseek effectively represents the structure of a protein as a 20-letter alphabet, enabling the use of fast sequence searching algorithms that were originally developed for the 20-letter amino acid alphabet (Farrar, 2007; Zhao et al., 2013; Steinegger and Söding, 2017). However, despite these advances, the computational bottleneck of AlphaFold2 structure predictions still hindered metagenomic applications, even with performance optimisations like ColabFold (Mirdita et al., 2022). To that end, the ProstT5 protein language model was developed, which translates between protein amino acid and Foldseek's 3Di sequences, without the need for structure predictions (Heinzinger et al., 2024). ProstT5 was built from an existing metagenomic protein language model (ProtT5) (Elnaggar et al., 2021), and further trained on 17 million non-redundant and

diverse protein sequences, paired with their corresponding 3Di sequences from the AlphaFold Protein Structure Database (Varadi et al., 2023). Consequently, ProstT5 enables ~4,000× faster structural homology searching, compared to AlphaFold2-ColabFold, without sensitivity loss (Heinzinger et al., 2024).

To take advantage of these recent breakthroughs for applications in microbial ecology, we created EcoFoldDB, a curated Foldseek database of protein structures spanning microbial traits of ecological relevance. The included traits cover biogeochemical cycling of carbon, nitrogen, sulphur, phosphorus, iron, and trace atmospheric gases, as well as key plant-microbe interactions, and osmotic stress tolerance mechanisms. Complementing this resource, we introduce *EcoFoldDB-annotate*, an annotation pipeline that leverages the speed and accuracy of ProstT5's 3Di sequence translation, and Foldseek's structural homology search capabilities. This annotation framework enables structure-guided functional profiling, scalable to metagenomic datasets.

Methods

Database construction

We included eight microbial functional categories, and 35 sub-categories, within EcoFoldDB that span several key ecological processes (Supplementary Table S1). Pathways and associated genes were sourced from published literature or existing functional databases. These consisted of *(1) Trace gas oxidation*: atmospheric hydrogen oxidation (Søndergaard et al., 2016; Bay et al., 2021), carbon monoxide oxidation (Bay et al., 2021), and methane oxidation (Bay et al., 2021); *(2) Carbon cycling*: carbon fixation (Caspi et al., 2013; Ruiz-Fernández et al., 2020; Garritano et al., 2022), one-carbon (C1) metabolism (Caspi et al.,

2013), complex carbohydrate degradation (Zheng et al., 2023), fatty acid degradation (Caspi et al., 2013), and polyphenol and aromatic hydrocarbon cycling (McGivern et al., 2024b; McGivern et al., 2024a); *(3) Nitrogen cycling*: nitrogen fixation, nitrification, denitrification, dissimilatory nitrate reduction to ammonium (DNRA), nitrate assimilation, anaerobic ammonium oxidation (anammox), and organic nitrogen mineralisation (Caspi et al., 2013; Tu et al., 2018; Ghaly et al., 2025); *(4) Sulphur cycling*: sulphate reduction, sulphite reduction, sulphide oxidation, sulphur disproportionation, sulphite oxidation, thiosulphate oxidation, thiosulphate disproportionation, sulphur reduction, and organic sulphur mineralisation (Morra and Dick, 1991; Eichhorn et al., 1997; Soutourina et al., 2001; Jamshad et al., 2007; Ellis, 2011; Caspi et al., 2013; Zhou et al., 2022); *(5) Phosphorus cycling*: organic phosphorus mineralisation and inorganic phosphorus solubilisation (Siles et al., 2022; Zeng et al., 2022; Ghaly et al., 2025); *(6) Iron cycling*: siderophore biosynthesis (Caspi et al., 2013); *(7) Plant-microbe interactions*: ethylene suppression via ACC deaminase activity (Glick et al., 2007), indole-3-acetic acid production (Tang et al., 2023), gibberellin production (Bottini et al., 2004), abscisic acid production (Khalil et al., 2024), trans-Zeatin production (Frébortová and Frébort, 2021), jasmonic acid production (Li et al., 2005; Khalil et al., 2024), brassinolide production (Bajguz et al., 2020), GABA production (Kinnersley and Turano, 2000; Ueno, 2000; Ma et al., 2012), putrescine and spermidine production (Dunn and Becerra-Rivera, 2023), and legume root nodulation (Rossen et al., 1985; Prell and Poole, 2006); and (8) *Osmotic stress tolerance*: osmoprotectant biosynthesis (Jebbar et al., 1992; Chen and Beattie, 2008; Ruhal et al., 2013; Yang et al., 2024). Included genes and pathways, and their functional categorisations are available as Supplementary Table S1.

Proteins with these functional annotations were selected from the manually curated databases, Swiss-Prot (Boeckmann et al., 2003) and MetaCyc (Caspi et al., 2013). The corresponding predicted structures for these protein sequences were then retrieved from the

AlphaFold Protein Structure Database (AFDB) (Jumper et al., 2021; Varadi et al., 2023), with protein fragments discarded. Redundancy was removed from this set of protein structures using Foldseek *cluster* (Barrio-Hernandez et al., 2023; van Kempen et al., 2024) with a minimum TM-score threshold of 0.9 and a bidirectional coverage of 0.9 [parameters: foldseek cluster --alignment-type 1 --tmscore-threshold 0.9 -c 0.9 --cov-mode 0 --cluster-reassign 1]. This final set of target EcoFoldDB structures were combined with structures of non-target proteins (i.e., those not included in our curated set of functions) from AFDB-Swiss-Prot. This was done to reduce false positives, by allowing the distinction between structurally homologous proteins with divergent functions.

EcoFoldDB-annotate pipeline and performance benchmarking

To enable metagenome-scale annotations using EcoFoldDB, we created *EcoFoldDB-annotate*, which employs Foldseek (van Kempen et al., 2024) with the ProstT5 protein language model (Heinzinger et al., 2024). Together, Foldseek with ProstT5, allow protein structural alignments from amino acid sequences without the need for structure predictions (Heinzinger et al., 2024). *EcoFoldDB-annotate* begins by filtering protein sequences exceeding 4,000 amino acids in length, archiving excluded sequences in a dedicated output file for user inspection. Proteins larger than this drastically increase the runtime of ProstT5 inference, and are likely too large to represent any protein in EcoFoldDB, with its largest entry being 2,378 amino acids. *EcoFoldDB-annotate* then uses ProstT5 to translate the amino acid alphabet to Foldseek's 3D interaction (3Di) structural alphabet, allowing structural alignment of input proteins against EcoFoldDB with Foldseek. Finally, the pipeline assigns functionally categorised annotations based on EcoFoldDB ontology, using the best-hit

approach with default query and target coverage cut-offs of 80%, and an e-value threshold of 1e-10.

To test its functional annotation capacity, we benchmarked *EcoFoldDB-annotate* against two widely used state-of-the-art annotation tools, eggNOG-mapper v2.1.13 (Cantalapiedra et al., 2021) and KofamScan v1.3.0 (Aramaki et al., 2019). To create a benchmarking dataset with 'known' positives, we used the ESM Metagenomic Atlas 30 (Lin et al., 2023), a non-redundant database of 37 million high-confidence (pTM > 0.7 and pLDDT > 0.7) predicted protein structures from metagenomic assemblies (derived from the MGnify database (Richardson et al., 2022)). The ESM Atlas 30 was chosen as it includes diverse proteins found in metagenomes, with protein entries sharing <30% sequence similarity with each other, and includes protein 'dark matter' sequences from genomes of uncultivated taxa. To identify a set of ground truth proteins, we first generated a subset of ESM Atlas proteins exhibiting high structural similarity to those encoding target functions, as determined by Foldseek. We selected only hits with both a query and target TM-score > 0.8, since this threshold indicates a near certain (~100%) posterior probability of sharing the same fold topology (Xu and Zhang, 2010). Next, the functions of these structural homologs were predicted using CLEAN v1.0.1 (Yu et al., 2023), which uses a contrastive machine learning model for high-quality enzymatic function predictions. Thus, in the absence of a validated reference dataset, these high-confidence matches (i.e., from both structural homology and CLEAN enzyme predictions) were used as a proxy for ground truth functional annotations. This resulted in 36,639 proteins to serve as 'known' positives. To include known negatives in the benchmarking dataset, we used all Swiss-Prot entries with non-target functions based on their assigned EC (Enzyme Commission) numbers. We limited this to Swiss-Prot entries (as of July 2025) not in AFDB-Swiss-Prot (derived from UniProt April 2021 data) as the latter are included in EcoFoldDB. This resulted in 4,106 proteins to act as known negatives.

The corresponding amino acid sequences of the combined positive and negative benchmarking dataset were then used as input for *EcoFoldDB-annotate*, eggNOG-mapper, and KofamScan. *EcoFoldDB-annotate* was run with default settings; eggNOG-mapper sequence searches were performed using DIAMOND (Buchfink et al., 2021) against eggNOG v5.0.2 orthology data (Huerta-Cepas et al., 2018); and KofamScan was run using its model-specific score thresholds (Aramaki et al., 2019).

EcoFoldDB annotations of microbial genomes from the global soil microbiome

To demonstrate its utility and scalability, we employed *EcoFoldDB-annotate* on >8,000 high-quality metagenome-assembled genomes (MAGs) from the global soil microbiome. Genome sequences were obtained from the Soil Microbial Dark Matter MAG (Ma et al., 2023) and Microflora Danica long-read MAG (Sereika et al., 2024) catalogues, collectively covering diverse soil habitats sampled across the globe. Genome sequences were filtered to retain only high-quality (HQ) MAGs based on MIMAG standards (Bowers et al., 2017). That is, genomes that are >90% complete, with <5% contamination, have the 23S, 16S, and 5S rRNA genes, and at least 18 different tRNA genes. Genome completion and contamination was determined for all MAGs using CheckM2 (Chklovski et al., 2023) with default settings. Genomes were screened for rRNA genes using barrnap (https://github.com/tseemann/barrnap) and bakta v1.8.1 (Schwengers et al., 2021) (which uses Infernal (Nawrocki and Eddy, 2013) against the Rfam database (Ontiveros-Palacios et al., 2025)), and tRNA genes were screened using bakta (which employs tRNAscan-SE 2.0 (Chan et al., 2021)) with default settings. MAGs were dereplicated at approximately species-level (95% average nucleotide identity, ANI, and 50% aligned fraction, AF) with dRep v3.5.0 (Olm et al., 2017). dRep was run using only skani v0.2.2 (Shaw and Yu, 2023) for ANI and

AF calculations, and CheckM2 results were provided to dRep for genome quality information [command: dRep dereplicate --SkipMash --S_algorithm skani -sa 0.95 -nc 0.5 --genomeInfo checkm2_info.csv -p 48].

Proteins were predicted from the HQ soil MAGs using Prodigal v2.6.3 (Hyatt et al., 2010) [parameters: -p meta], which was parallelised across 48 CPU cores using GNU Parallel (Tange, 2018). This resulted in the prediction of 31,890,584 proteins. These proteins were functionally annotated against EcoFoldDB using *EcoFoldDB-annotate* with default settings. Genome sequences and predicted protein sequences of these HQ MAGs have been made available via https://doi.org/10.5281/zenodo.15054612, and EcoFoldDB annotations are available as Supplementary Table S2.

Taxonomy of the HQ soil MAGs was assigned with GTDB-Tk v2.4.0 (Chaumeil et al., 2019, 2022) using GTDB release 220 (Parks et al., 2018; Parks et al., 2020; Parks et al., 2021; Rinke et al., 2021) [parameters: classify_wf --skip_ani_screen]. A whole-genome phylogeny was generated using PhyloPhlAn 3.0 (Asnicar et al., 2020), employing DIAMOND (Buchfink et al., 2021) for marker detection, MAFFT (Katoh and Standley, 2013) for sequence alignment, trimAl (Capella-Gutiérrez et al., 2009) for alignment trimming, and FastTree2 (Price et al., 2010), with the WAG substitution model (Whelan and Goldman, 2001), for phylogenetic inference [PhyloPhlAn parameters: --diversity high --fast]. The resulting phylogenetic tree was rooted with minimal ancestor deviation (MAD) rooting (Tria et al., 2017) using MADroot (Bryant and Charleston, 2018) (https://github.com/davidjamesbryant/MADroot), and visualised using the *ggtree* (Yu et al., 2017) and *ggtreeExtra* (Xu et al., 2021) R packages.

Results and Discussion

Here, we present EcoFoldDB, a database of protein structures that cover microbial functions of ecological relevance, including those involved in: (1) trace gas oxidation, (2) carbon cycling, (3) nitrogen cycling, (4) sulphur cycling, (5) phosphorus cycling, (6) iron cycling, (7) plant-microbe interactions, and (8) osmotic stress tolerance (Fig. 1). These eight functional categories are organised into 35 sub-categories, and encompass 166 different functional activities/pathways (Fig. 1; Supplementary Table S1). Together, EcoFoldDB comprises 842 protein structures linked to 637 genes (Supplementary Table S1). These functions cover key biogeochemical cycles important for ecosystem health, biodiversity, and climate change (i.e., carbon sequestration and greenhouse gas emissions), as well as microbial traits relevant to adaptive responses to environmental stress, and plant growth promotion (Fig. 1). Future versions of EcoFoldDB can be readily expanded to incorporate new functions, particularly as novel microbial traits influencing ecosystem dynamics are uncovered.

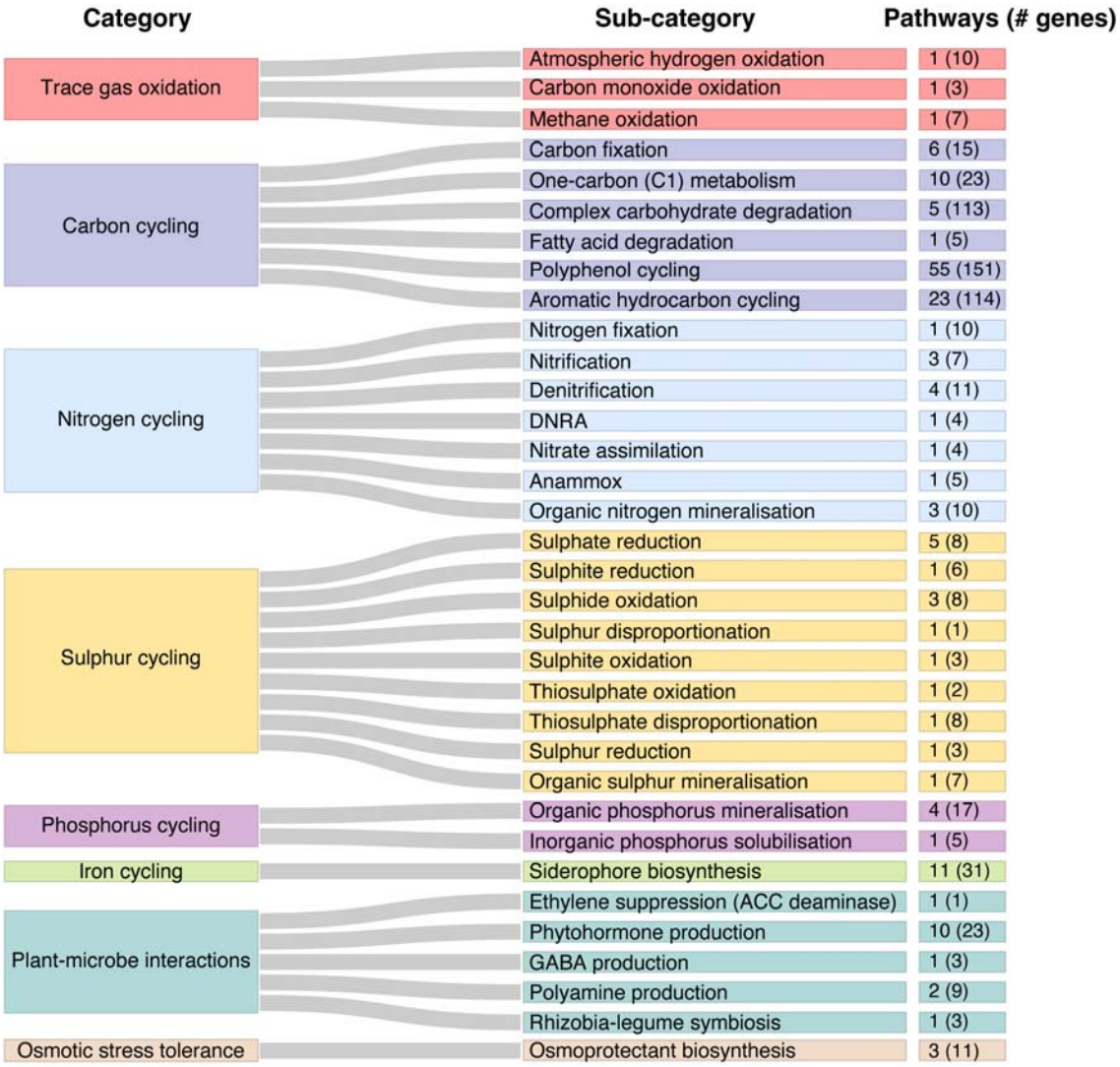

Fig. 1. Ecologically relevant functions covered by EcoFoldDB. The EcoFoldDB ontology is comprised of 8 functional categories (left), which are divided into 34 sub-categories (middle). These cover 166 functional pathways/activities and 637 gene families (right). These gene families are represented by 842 protein structures in EcoFoldDB. For a full list of EcoFoldDB entries and their functional categorisations, see Supplementary Table S1.

'EcoFoldDB-annotate' performs highly sensitive and accurate structure-guided functional profiling

To enable scalable, structure-guided functional annotation using EcoFoldDB, we developed *EcoFoldDB-annotate* (Fig. 2a). This pipeline leverages Foldseek (van Kempen et al., 2024)

and the ProstT5 protein language model (Heinzinger et al., 2024) to directly translate amino acid sequences into 3Di structural representations, enabling homology searches against EcoFoldDB without the need for computationally intensive structure predictions. ProstT5 achieves this translation ~4,000-fold faster than AlphaFold2-ColabFold without losing sensitivity (Mirdita et al., 2022), allowing structural homology detection scalable to metagenomic datasets.

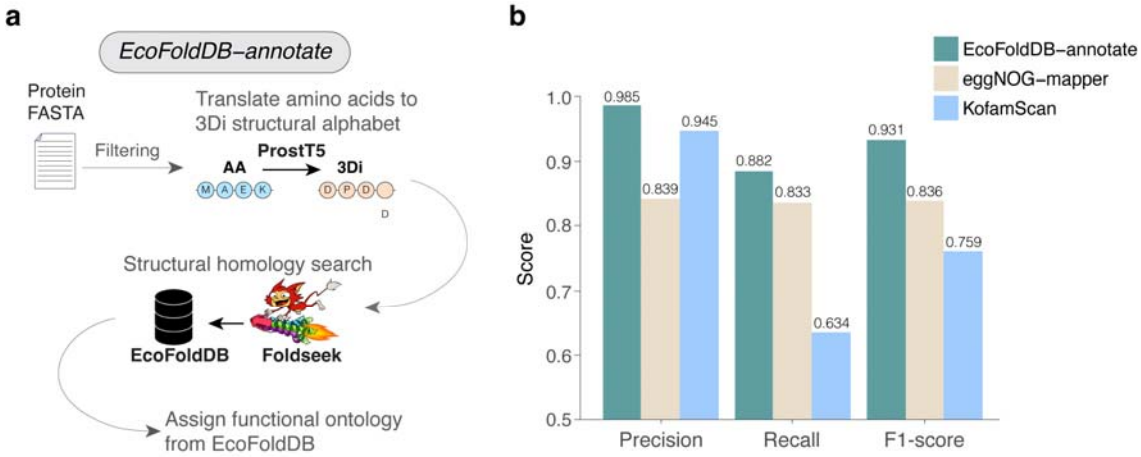

Fig. 2. *EcoFoldDB-annotate* pipeline and performance benchmarking. a, *EcoFoldDB-annotate* first filters protein sequences to remove those larger than 4,000 amino acids in length, then uses the ProstT5 protein language model to translate amino acid sequences to Foldseek's 3D interaction (3Di) structural alphabet for structural homology searching against EcoFoldDB, and finally assigns hierarchically categorised functional annotations based on EcoFoldDB ontology. **b**, The performance of *EcoFoldDB-annotate* was benchmarked against eggNOG-mapper and KofamScan. The following benchmarking metrics were used: recall/sensitivity (proportion of all possible positives correctly identified), precision (proportion of predicted positives that were true), and F1-score (the harmonic mean of recall and precision).

EcoFoldDB-annotate outperforms sequence-based annotations for proteins derived from metagenomic assemblies (Fig. 2b). *EcoFoldDB-annotate* achieves higher sensitivity, with a recall of 0.882 compared to 0.833 (eggNOG-mapper) and 0.634 (KofamScan), while maintaining the highest precision (0.985) and F1-score (0.931) among the three tools (Fig. 2b). This combination of high sensitivity and precision establishes *EcoFoldDB-annotate* as an effective solution for annotating metagenomes, which are frequently dominated by

sequences lacking detectable amino acid homology to reference databases (Vanni et al., 2022; Pavlopoulos et al., 2023). Its performance, scalability, and hierarchical classification of ecologically relevant functions (Fig. 1) positions *EcoFoldDB-annotate* as an invaluable resource for environmental metagenomic studies.

EcoFoldDB-annotate is scalable to millions of proteins from soil metagenomes

To demonstrate the utility of *EcoFoldDB-annotate*, we used it to perform structure-based functional annotations of 32 million protein sequences encoded by 8,000 high-quality metagenome-assembled genomes (MAGs), derived from diverse soil habitats across the globe. Soil microbes harbor a diverse range of metabolic capabilities, and influence global carbon dynamics and biogeochemical cycles that shape ecosystem health and productivity (Fierer, 2017). Soil environments generally harbour greater phylogenetic diversity and genetic novelty compared to other environments (Thompson et al., 2017; Brewer et al., 2019; Holland-Moritz et al., 2021), making sequenced-based functional annotations particularly difficult (Holland-Moritz et al., 2021). Leveraging the recent breakthroughs in protein structure predictions (Jumper et al., 2021; Mirdita et al., 2022) can improve functional annotations of proteins that lie beyond the limits of sequence homology detection with known proteins (Durairaj et al., 2023). To that end, we employed *EcoFoldDB-annotate* to functionally annotate MAGs from the global soil microbiome. We processed 32 million proteins in batches of 1 million sequences, taking ~5.3 hours per batch using a server with 4 Nvidia Tesla Volta V100 GPUs and 48 CPU cores (see Supplementary Table S2 for all EcoFoldDB annotations). This demonstrates the scalability of *EcoFoldDB-annotate* to functionally profile large metagenomic datasets.

EcoFoldDB reveals the taxonomic distribution of nitrogen cycling functions in the global soil microbiome

As a use case example, we used the EcoFoldDB annotations of the high-quality MAGs to reveal the phylogenetic distribution of encoded nitrogen cycling pathways in soils. Our findings reveal distinct taxonomic partitioning, with some pathways, particularly nitrification, being taxonomically restricted, while others, like dissimilatory nitrate reduction (DNRA) and denitrification, were broadly distributed across taxa (Fig. 3). For instance, ammonia oxidation, the initial step in nitrification, was restricted to just three prokaryotic families. Archaea dominated this functional guild (80% of ammonia oxidisers), with all archaeal representatives belonging to the families Nitrosopumilaceae and Nitrososphaeraceae. This agrees with previous studies showing archaea to be the predominant drivers of ammonia oxidation on the planet (Leininger et al., 2006; Wuchter et al., 2006). The remaining 20% belong to the bacterial families Nitrospiraceae (phylum Nitrospirota) and Nitrosomonadaceae (phylum Pseudomonadota). Similarly, nitrite oxidation (nitrification) was phylogenetically constrained, occurring predominantly within Nitrospirota (Fig. 3), a pattern consistent with prior reports identifying this phylum as the most widespread nitrite oxidisers (Leung et al., 2022). In contrast, pathways like DNRA and denitrification (particularly nitrate, nitrite, and nitrous oxide reduction) were distributed across diverse bacterial and archaeal taxa, underscoring their functional versatility across lineages and environmental niches.

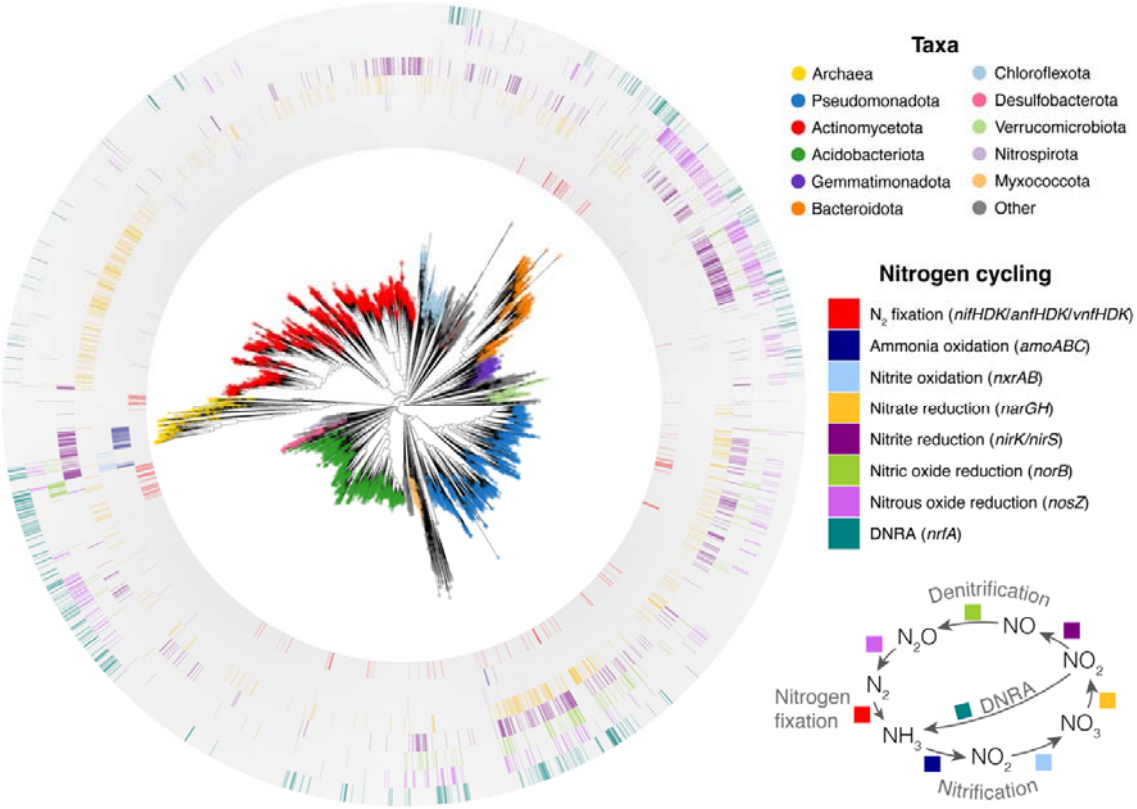

Fig. 3. Nitrogen cycling functions identified by *EcoFoldDB-annotate* among high-quality MAGs from the global soil microbiome. A whole genome phylogeny of high-quality bacterial and archaeal MAGs, derived from soil metagenomes across the globe. Coloured tip points on the tree indicate GTDB phyla, except for archaeal genomes, which have all been assigned a single domain-level colour (yellow tip points). Coloured rings indicate the genomes identified to have the genetic potential for nitrogen cycling functions. As a minimum, the following genes were required for N_2 fixation: (*nifH* and *nifD* and *nifK*) or (*anfH* and *anfD* and *anfK*) or (*vnfH* and *vnfD* and *vnfK*); ammonia oxidation: *amoA*, *amoB* and *amoC*; nitrite oxidation: *nxrA* and *nxrB*; nitrate reduction: *narG* and *narH*; nitrite reduction: *nirK* or *nirS*; nitric oxide reduction: *norB*; nitrous oxide reduction: *nosZ*; and DNRA (dissimilatory nitrite reduction to ammonia): *nrfA*.

The taxonomic and metabolic insights generated here, from phylogenetically limited nitrifying taxa to more widespread denitrifiers, demonstrate *EcoFoldDB-annotate*'s capacity to disentangle complex microbial roles in biogeochemical cycling. This scalable and sensitive annotation framework thus allows efficient linking of microbial community structure to ecosystem-scale functions in large datasets.

EcoFoldDB identifies microbial genomes encoding plant-microbe interaction traits

As a second use case example, we sought to identify potential plant growth-promoting members from the global soil microbiome. We used the EcoFoldDB annotations to identify soil MAGs encoding plant-microbe interaction traits. We found that many genomes encode multiple plant-microbe interaction traits, with 15.9% encoding two or more, and 2.5% encoding three or more (Fig. 4). A few of these traits were widespread, e.g., putrescine and spermidine production, which were present in most bacterial phyla. This is not surprising since biosynthesis of these polyamines is also prevalent in non-plant-associated bacteria given their diverse roles in bacterial physiology and growth, such as biofilm formation, motility, and siderophore production (Michael, 2018). The genetic capacity to produce the phytohormone auxin (indole-3-acetic acid, IAA) was also widespread, with 7.2% of MAGs containing at least one of the known IAA biosynthesis pathways (Fig. 4).

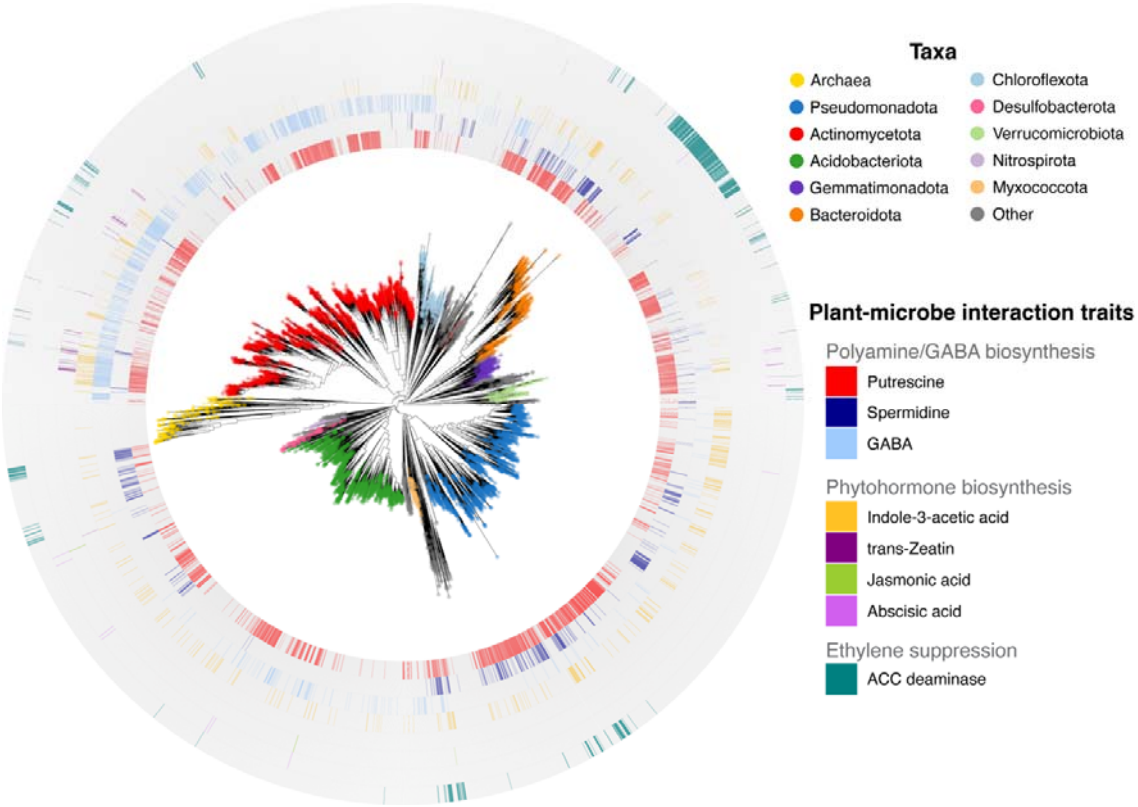

Fig. 4. Plant-microbe interaction functions identified by *EcoFoldDB-annotate* among high-quality MAGs from the global soil microbiome. A whole genome phylogeny of high-

quality bacterial and archaeal MAGs, derived from soil metagenomes across the globe. Coloured tip points on the tree indicate GTDB phyla, except for archaeal genomes, which have all been assigned a single domain-level colour (yellow tip points). Coloured rings indicate the genomes identified to have the genetic potential for plant-microbe interaction traits. As a minimum, the following genes were required for putrescine production: *speC* or (*speA* and *speB*) or (*speA* and *aguA* and *aguB*); spermidine production: (*speD*/*speH* and *speE*) or (*cansdh* and *nspC*); GABA (gamma-aminobutyric acid) production: *gadA*/*gadB* and *gadC*; indole-3-acetic acid (IAA) production via the indole-3-acetamide (IAM) pathway: *iaaM* or *iaaH*; IAA production via the indole-3-pyruvic acid (IPA) and tryptamine (TAM) pathways: *aldH*; IAA production via the indole-3-acetonitrile (IAN) pathway: *nit*; trans-Zeatin production: *log*; jasmonic acid production: *ehhadh* and *acox*; abscisic acid production: *aao*; brassinolide production: *br6ox* (not shown as no genomes were identified to carry this gene, Supplementary Table S2); gibberellin production: *ko* and *kao* (not shown as no genomes were identified to carry both genes, Supplementary Table S2); and ACC deaminase: *acdS*.

Other plant-microbe interaction traits, however, had strong taxonomic signals (Fig. 4). For example, 95% of trans-Zeatin producers and 76.2% of GABA producers belonged to the phylum Actinomycetota, largely from the order Mycobacteriales. ACC deaminase activity showed a strong Bacteroidota signature. Indeed, over half (55.2%) of all Bacteroidota MAGs encoded ACC deaminase, known to induce plant tolerance to abiotic stress (Glick et al., 2007). While ACC deaminase activity has been experimentally confirmed in Bacteroidota isolates (Maimaiti et al., 2007; Gandhi Pragash et al., 2009; Nadeem et al., 2009; Marques et al., 2010), research on this trait has historically focused on Pseudomonadota (Blaha et al., 2006; Nascimento et al., 2014). Thus, to our knowledge, the pronounced enrichment of ACC deaminase activity within Bacteroidota observed here has not been previously reported.

To identify possible novel plant growth promoting strains, we examined the subset of genomes encoding the greatest number of identified plant-microbe interaction traits, with 22 species-clustered MAGs being positive for four different traits (Fig. 5). Of these, 14 currently have no cultured species representative, with 10 lacking a cultured genus-level representative. The enrichment of encoded plant-microbe interaction traits, suggest that they likely have plant-growth promoting effects. Indeed, most of the species that do have a cultured representative have already been shown to comprise plant growth-promoting members (e.g.,

Pseudomonas putida (Hall et al., 1996; Ahemad and Khan, 2012; Costa-Gutierrez et al., 2020), *Pseudomonas sichuanensis* (a.k.a. *P. oryziphila*) (Yang et al., 2021), *Pseudomonas* sp024809425 (Izquierdo, 2017), *Pseudomonas furukawaii* (Rafique et al., 2022), *Rahnella variigena* (Raj et al., 2023)). The remaining species, which are also likely to comprise plant growth-promoting members, represent ideal candidates for targeted isolation and culturing. Such isolation efforts can be guided by genomic information from the MAG sequences, e.g., via optimisation of culture medium and conditions, including substrate utilisation, oxygen requirements, resistance to antibiotics, etc. (Pope et al., 2011; Karnachuk et al., 2021; Liu et al., 2022). This demonstrates how EcoFoldDB can help bridge ecological discovery and real-world application, translating metagenomic insights into actionable strategies for sustainable management.

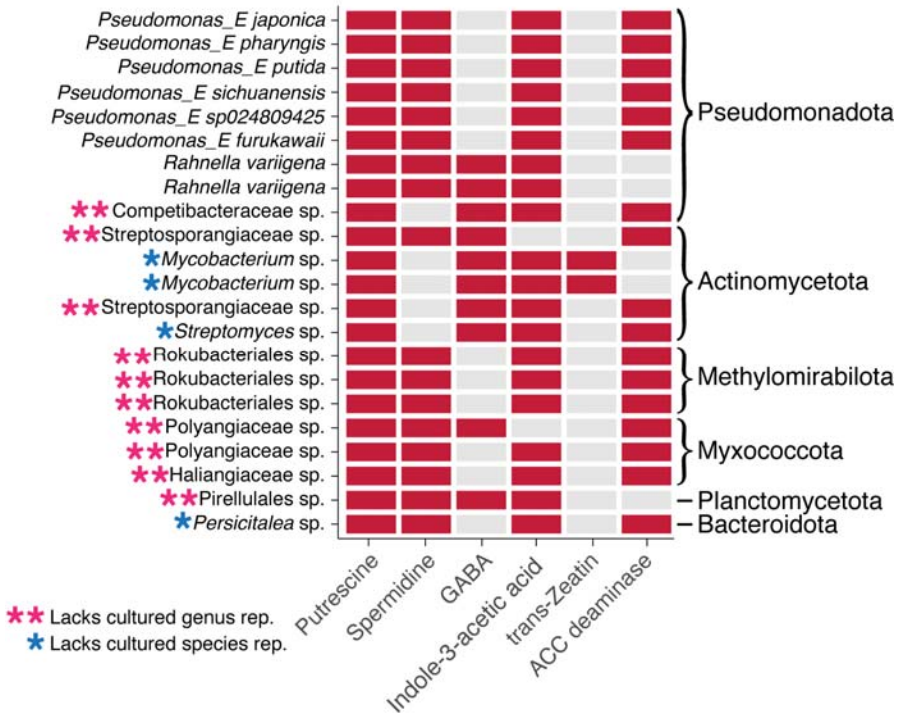

Fig. 5. Identification of potential plant growth-promoting bacteria as candidates for targeted culturing. Soil MAGs encoding the greatest number of plant-microbe interaction traits (x-axis), as determined by *EcoFoldDB-annotate* are shown, labelled with their lowest named GTDB taxonomic rank (y-axis), and GTDB phylum (right). Each MAG represents a distinct 95% average nucleotide identity species cluster. The minimum gene sets for a genome to be considered positive (red rectangles) for each trait are as listed in Fig. 4. Strains

lacking a cultured species representative are denoted with a single blue asterisk, and those also lacking a cultured genus representative are shown with double pink asterisks. Figure data, including the genome identifiers for these 22 MAGs, are available as Supplementary Table S3.

Conclusion

Here, we introduce EcoFoldDB, a database of protein structures linked to ecologically relevant functions, and its companion pipeline, *EcoFoldDB-annotate*, for structure-guided functional profiling. *EcoFoldDB-annotate* achieves highly sensitive and accurate annotations, scalable to metagenomic datasets (millions of proteins). We showcase the use of *EcoFoldDB-annotate* on a global soil microbiome MAG catalogue, revealing key insights into the taxonomic distribution of important nitrogen cycling pathways, and plant-microbe interaction traits, advancing our understanding of important ecosystem processes. We anticipate that EcoFoldDB will serve as a valuable resource for microbial ecology studies, enabling researchers to extract ecological insights from environmental genomes and metagenomes, and accelerating discoveries in microbiome science.

Availability of data and material

EcoFoldDB and *EcoFoldDB-annotate* pipeline are available via https://github.com/timghaly/EcoFoldDB. Genomic and protein sequences for the high-quality soil MAGs are available via https://doi.org/10.5281/zenodo.15054612. EcoFoldDB annotations of these genomes are available as Supplementary Table S2.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by a Macquarie University Research Fellowship (T.M.G) and the ARC Centre of Excellence in Biology (S.G.T).

Authors' contributions

T.M.G was involved in the project conception and design, conducted the research, and wrote the manuscript draft. V.R., D.R and E.C were involved in pipeline testing and analysis. S.G.T was involved in project design and management. All authors contributed to the final editing of the manuscript.

Acknowledgements

We would like to thank the developers of all the software used in this study, particularly Foldseek and ProstT5, without which, this work would not have been possible. TMG would also like to thank Mary, Saoirse, and Maebh Ghaly for their loving support.

References

Ahemad, M., and Khan, M.S. (2012) Evaluation of plant-growth-promoting activities of rhizobacterium *Pseudomonas putida* under herbicide stress. *Annals of microbiology* **62**: 1531-1540.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2019) KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**: 2251-2252.

Asnicar, F., Thomas, A.M., Beghini, F., Mengoni, C., Manara, S., Manghi, P. et al. (2020) Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nature Communications* **11**: 2500.

Bahram, M., Espenberg, M., Pärn, J., Lehtovirta-Morley, L., Anslan, S., Kasak, K. et al. (2022) Structure and function of the soil microbiome underlying N_2O emissions from global wetlands. *Nature Communications* **13**: 1430.

Bahram, M., Hildebrand, F., Forslund, S.K., Anderson, J.L., Soudzilovskaia, N.A., Bodegom, P.M. et al. (2018) Structure and function of the global topsoil microbiome. *Nature* **560**: 233-237.

Bajguz, A., Chmur, M., and Gruszka, D. (2020) Comprehensive overview of the brassinosteroid biosynthesis pathways: Substrates, products, inhibitors, and connections. *Frontiers in Plant Science* **11**: 1034.

Barrio-Hernandez, I., Yeo, J., Jänes, J., Mirdita, M., Gilchrist, C.L.M., Wein, T. et al. (2023) Clustering predicted structures at the scale of the known protein universe. *Nature* **622**: 637-645.

Bay, S.K., Dong, X., Bradley, J.A., Leung, P.M., Grinter, R., Jirapanjawat, T. et al. (2021) Trace gas oxidizers are widespread and active members of soil microbial communities. *Nature Microbiology* **6**: 246-256.

Blaha, D., Prigent-Combaret, C., Mirza, M.S., and Moenne-Loccoz, Y. (2006) Phylogeny of the 1-aminocyclopropane-1-carboxylic acid deaminase-encoding gene *acdS* in phytobeneficial and pathogenic Proteobacteria and relation with strain biogeography. *FEMS Microbiology Ecology* **56**: 455-470.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* **31**: 365-370.

Bottini, R., Cassán, F., and Piccoli, P. (2004) Gibberellin production by bacteria and its involvement in plant growth promotion and yield increase. *Applied Microbiology and Biotechnology* **65**: 497-503.

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K. et al. (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* **35**: 725-731.

Brewer, T.E., Aronson, E.L., Arogyaswamy, K., Billings, S.A., Botthoff, J.K., Campbell, A.N. et al. (2019) Ecological and genomic attributes of novel bacterial taxa that thrive in subsurface soil horizons. *mBio* **10**: 10.1128/mbio.01318-01319.

Bryant, D., and Charleston, M. (2018) MAD roots for large trees. *arXiv*: 1811.03174.

Buchfink, B., Reuter, K., and Drost, H.-G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**: 366-368.

Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021) eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution* **38**: 5825-5829.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972-1973.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A. et al. (2013) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research* **42**: D459-D471.

Cavicchioli, R., Ripple, W.J., Timmis, K.N., Azam, F., Bakken, L.R., Baylis, M. et al. (2019) Scientists' warning to humanity: microorganisms and climate change. *Nature Reviews Microbiology* **17**: 569-586.

Chan, Patricia P., Lin, Brian Y., Mak, Allysia J., and Lowe, Todd M. (2021) tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research* **49**: 9077-9096.

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2019) GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**: 1925-1927.

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2022) GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**: 5315-5316.

Chen, C., and Beattie, G.A. (2008) *Pseudomonas syringae* BetT Is a low-affinity choline transporter that is responsible for superior osmoprotection by choline over glycine betaine. *Journal of Bacteriology* **190**: 2717-2725.

Chklovski, A., Parks, D.H., Woodcroft, B.J., and Tyson, G.W. (2023) CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods* **20**: 1203-1212.

Costa-Gutierrez, S.B., Lami, M.J., Santo, M.C.C.-D., Zenoff, A.M., Vincent, P.A., Molina-Henares, M.A. et al. (2020) Plant growth promotion by *Pseudomonas putida* KT2440 under saline stress: role of *eptA*. *Applied microbiology and biotechnology* **104**: 4577-4592.

Dunn, M.F., and Becerra-Rivera, V.A. (2023) The biosynthesis and functions of polyamines in the interaction of plant growth-promoting rhizobacteria with plants. *Plants* **12**: 2671.

Durairaj, J., Waterhouse, A.M., Mets, T., Brodiazhenko, T., Abdullah, M., Studer, G. et al. (2023) Uncovering new families and folds in the natural protein universe. *Nature* **622**: 646-653.

Eichhorn, E., van der Ploeg, J.R., Kertesz, M.A., and Leisinger, T. (1997) Characterization of α-ketoglutarate-dependent taurine dioxygenase from *Escherichia coli*. *Journal of Biological Chemistry* **272**: 23031-23036.

Ellis, H.R. (2011) Mechanism for sulfur acquisition by the alkanesulfonate monooxygenase system. *Bioorganic Chemistry* **39**: 178-184.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L. et al. (2021) Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* **44**: 7112-7127.

Farrar, M. (2007) Striped Smith–Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics* **23**: 156-161.

Fierer, N. (2017) Embracing the unknown: disentangling the complexities of the soil microbiome. *Nature Reviews Microbiology* **15**: 579-590.

Finn, R.D., Clements, J., and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic acids research* **39**: W29-W37.

Frébortová, J., and Frébort, I. (2021) Biochemical and structural aspects of cytokinin biosynthesis and degradation in bacteria. *Microorganisms* **9**: 1314.

Gandhi Pragash, M., Narayanan, K.B., Naik, P.R., and Sakthivel, N. (2009) Characterization of *Chryseobacterium aquaticum* strain PUPC1 producing a novel antifungal protease from rice rhizosphere soil. *Journal of Microbiology and Biotechnology* **19**: 99-107.

Garritano, A.N., Song, W., and Thomas, T. (2022) Carbon fixation pathways across the bacterial and archaeal tree of life. *PNAS Nexus* **1**: pgac226.

Ghaly, T.M., Shah, B.S., Coleman, N.V., Elbourne, L.D.H., Le Roux, J.J., Gillings, M.R. et al. (2025) Agriculture alters protein evolution of nitrogen cycling genes in soil bacteria at a global scale. *bioRxiv*: 2025.2005.2007.652758.

Gillings, M.R., and Paulsen, I.T. (2014) Microbiology of the Anthropocene. *Anthropocene* **5**: 1-8.

Glick, B.R., Todorovic, B., Czarny, J., Cheng, Z., Duan, J., and McConkey, B. (2007) Promotion of plant growth by bacterial ACC deaminase. *Critical Reviews in Plant Sciences* **26**: 227-242.

Hall, J.A., Peirson, D., Ghosh, S., and Glick, B. (1996) Root elongation in various agronomic crops by the plant growth promoting rhizobacterium *Pseudomonas putida* GR12–2. *Israel Journal of Plant Sciences* **44**: 37-42.

Heinzinger, M., Weissenow, K., Sanchez, Joaquin G., Henkel, A., Mirdita, M., Steinegger, M., and Rost, B. (2024) Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics* **6**: lqae150.

Holland-Moritz, H., Vanni, C., Fernandez-Guerra, A., Bissett, A., and Fierer, N. (2021) An ecological perspective on microbial genes of unknown function in soil. *bioRxiv*: 2021.2012.2002.470747.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H. et al. (2018) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**: D309-D314.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T. et al. (2012) Structure, function and diversity of the healthy human microbiome. *Nature* **486**: 207-214.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.

Illergård, K., Ardell, D.H., and Elofsson, A. (2009) Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics* **77**: 499-508.

Izquierdo, J. (2017) Plant growth-promoting bacteria from the rhizosphere of the beachgrass Ammophila breviligulata. In: USDOE Joint Genome Institute (JGI), Berkeley, CA (United States).

Jamshad, M., Murrell, J.C., and Fülöp, V. (2007) Purification and crystallization of the hydroxylase component of the methanesulfonate monooxygenase from *Methylosulfonomonas methylovora* strain M2. *Protein Expression and Purification* **52**: 472-477.

Jebbar, M., Talibart, R., Gloux, K., Bernard, T., and Blanco, C. (1992) Osmoprotection of *Escherichia coli* by ectoine: uptake and accumulation characteristics. *Journal of Bacteriology* **174**: 5027-5035.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C. et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236-1240.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O. et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**: 583-589.

Karnachuk, O.V., Lukina, A.P., Kadnikov, V.V., Sherbakova, V.A., Beletsky, A.V., Mardanov, A.V., and Ravin, N.V. (2021) Targeted isolation based on metagenome-assembled genomes reveals a phylogenetically distinct group of thermophilic spirochetes from deep biosphere. *Environmental Microbiology* **23**: 3585-3598.

Katoh, K., and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**: 772-780.

Khalil, A., Bramucci, A.R., Focardi, A., Le Reun, N., Willams, N.L.R., Kuzhiumparambil, U. et al. (2024) Widespread production of plant growth-promoting hormones among marine bacteria and their impacts on the growth of a marine diatom. *Microbiome* **12**: 205.

Kinnersley, A.M., and Turano, F.J. (2000) Gamma aminobutyric acid (GABA) and plant responses to stress. *Critical Reviews in Plant Sciences* **19**: 479-509.

Leininger, S., Urich, T., Schloter, M., Schwark, L., Qi, J., Nicol, G.W. et al. (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* **442**: 806-809.

Leung, P.M., Daebeler, A., Chiri, E., Hanchapola, I., Gillett, D.L., Schittenhelm, R.B. et al. (2022) A nitrite-oxidising bacterium constitutively consumes atmospheric hydrogen. *The ISME Journal* **16**: 2213-2219.

Li, C., Schilmiller, A.L., Liu, G., Lee, G.I., Jayanty, S., Sageman, C. et al. (2005) Role of β-oxidation in jasmonate biosynthesis and systemic wound signaling in tomato. *The Plant Cell* **17**: 971-986.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W. et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**: 1123-1130.

Liu, S., Moon, C.D., Zheng, N., Huws, S., Zhao, S., and Wang, J. (2022) Opportunities and challenges of using metagenomic data to bring uncultured microbes into cultivation. *Microbiome* **10**: 76.

Ma, B., Lu, C., Wang, Y., Yu, J., Zhao, K., Xue, R. et al. (2023) A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. *Nature Communications* **14**: 7318.

Ma, D., Lu, P., Yan, C., Fan, C., Yin, P., Wang, J., and Shi, Y. (2012) Structure and mechanism of a glutamate–GABA antiporter. *Nature* **483**: 632-636.

Maimaiti, J., Zhang, Y., Yang, J., Cen, Y.P., Layzell, D.B., Peoples, M., and Dong, Z. (2007) Isolation and characterization of hydrogen-oxidizing bacteria induced following exposure of soil to hydrogen gas and their impact on plant growth. *Environmental Microbiology* **9**: 435-444.

Marques, A.P., Pires, C., Moreira, H., Rangel, A.O., and Castro, P.M. (2010) Assessment of the plant growth promotion abilities of six bacterial isolates using *Zea mays* as indicator plant. *Soil Biology and Biochemistry* **42**: 1229-1235.

McGivern, B.B., Woyda, R., Flynn, R.M., and Wrighton, K.C. (2024a) CAMPER: curated annotations for profiling microbial polyphenol metabolic potential. *bioRxiv*: 2023.2009.2024.559193.

McGivern, B.B., Cronin, D.R., Ellenbogen, J.B., Borton, M.A., Knutson, E.L., Freire-Zapata, V. et al. (2024b) Microbial polyphenol metabolism is part of the thawing permafrost carbon cycle. *Nature Microbiology* **9**: 1454-1466.

Michael, A.J. (2018) Polyamine function in archaea and bacteria. *Journal of Biological Chemistry* **293**: 18693-18701.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022) ColabFold: making protein folding accessible to all. *Nature Methods* **19**: 679-682.

Morra, M.J., and Dick, W.A. (1991) Mechanisms of H_2S production from cysteine and cystine by microorganisms isolated from soil by selective enrichment. *Applied and Environmental Microbiology* **57**: 1413-1417.

Nadeem, S.M., Zahir, Z.A., Naveed, M., and Arshad, M. (2009) Rhizobacteria containing ACC-deaminase confer salt tolerance in maize grown on salt-affected fields. *Canadian Journal of Microbiology* **55**: 1302-1309.

Nascimento, F.X., Rossi, M.J., Soares, C.R.F.S., McConkey, B.J., and Glick, B.R. (2014) New insights into 1-aminocyclopropane-1-carboxylate (ACC) deaminase phylogeny, evolution and ecological significance. *PLoS One* **9**: e99168.

Nawrocki, E.P., and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**: 2933-2935.

Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017) dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal* **11**: 2864-2868.

Ontiveros-Palacios, N., Cooke, E., Nawrocki, E.P., Triebel, S., Marz, M., Rivas, E. et al. (2025) Rfam 15: RNA families database in 2025. *Nucleic acids research* **53**: D258-D267.

Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., and Hugenholtz, P. (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology* **38**: 1079-1086.

Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., and Hugenholtz, P. (2021) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research* **50**: D785-D794.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* **36**: 996-1004.

Pavlopoulos, G.A., Baltoumas, F.A., Liu, S., Selvitopi, O., Camargo, A.P., Nayfach, S. et al. (2023) Unraveling the functional dark matter through global metagenomics. *Nature* **622**: 594-602.

Pope, P., Smith, W., Denman, S., Tringe, S., Barry, K., Hugenholtz, P. et al. (2011) Isolation of Succinivibrionaceae implicated in low methane emissions from Tammar wallabies. *Science* **333**: 646-648.

Prell, J., and Poole, P. (2006) Metabolic changes of rhizobia in legume nodules. *Trends in Microbiology* **14**: 161-168.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**: e9490.

Rafique, E., Mumtaz, M.Z., Ullah, I., Rehman, A., Qureshi, K.A., Kamran, M. et al. (2022) Potential of mineral-solubilizing bacteria for physiology and growth promotion of *Chenopodium quinoa* Willd. *Frontiers in Plant Science* **13**: 1004833.

Raj, Y., Kumar, A., Kumari, S., Kumar, R., and Kumar, R. (2023) Comparative genomics and physiological investigations supported multifaceted plant growth-promoting activities in two *Hypericum perforatum* L.-associated plant growth-promoting rhizobacteria for microbe-assisted cultivation. *Microbiology Spectrum* **11**: e00607-00623.

Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, Maxwell L., Burdett, T. et al. (2022) MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research* **51**: D753-D759.

Rinke, C., Chuvochina, M., Mussig, A.J., Chaumeil, P.-A., Davín, A.A., Waite, D.W. et al. (2021) A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nature Microbiology* **6**: 946-959.

Rohwer, R.R., Kirkpatrick, M., Garcia, S.L., Kellom, M., McMahon, K.D., and Baker, B.J. (2025) Two decades of bacterial ecology and evolution in a freshwater lake. *Nature Microbiology* **10**: 246-257.

Rossen, L., Shearman, C., Johnston, A., and Downie, J. (1985) The nodD gene of Rhizobium leguminosarum is autoregulatory and in the presence of plant exudate induces the nodA, B, C genes. *The EMBO Journal* **4**: 3369-3373.

Ruhal, R., Kataria, R., and Choudhury, B. (2013) Trends in bacterial trehalose metabolism and significant nodes of metabolic pathway in the direction of trehalose accumulation. *Microbial Biotechnology* **6**: 493-502.

Ruiz-Fernández, P., Ramírez-Flandes, S., Rodríguez-León, E., and Ulloa, O. (2020) Autotrophic carbon fixation pathways along the redox gradient in oxygen-depleted oceanic waters. *Environmental Microbiology Reports* **12**: 334-341.

Schwengers, O., Jelonek, L., Dieckmann, M.A., Beyvers, S., Blom, J., and Goesmann, A. (2021) Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics* **7**.

Sereika, M., Mussig, A.J., Jiang, C., Knudsen, K.S., Jensen, T.B.N., Petriglieri, F. et al. (2024) Recovery of highly contiguous genomes from complex terrestrial habitats reveals over 15,000 novel prokaryotic species and expands characterization of soil and sediment microbial communities. *bioRxiv*: 2024.2012.2019.629313.

Shaffer, M., Borton, M.A., McGivern, B.B., Zayed, A.A., La Rosa, Sabina L., Solden, L.M. et al. (2020) DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research* **48**: 8883-8900.

Shaw, J., and Yu, Y.W. (2023) Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nature Methods* **20**: 1661-1665.

Siles, J.A., Starke, R., Martinovic, T., Parente Fernandes, M.L., Orgiazzi, A., and Bastida, F. (2022) Distribution of phosphorus cycling genes across land uses and microbial taxonomic groups based on metagenome and genome mining. *Soil Biology and Biochemistry* **174**: 108826.

Søndergaard, D., Pedersen, C.N.S., and Greening, C. (2016) HydDB: A web tool for hydrogenase classification and analysis. *Scientific Reports* **6**: 34212.

Soutourina, J., Blanquet, S., and Plateau, P. (2001) Role of d-cysteine desulfhydrase in the adaptation of *Escherichia coli* to d-cysteine. *Journal of Biological Chemistry* **276**: 40864-40872.

Steffen, W., Richardson, K., Rockström, J., Cornell, S.E., Fetzer, I., Bennett, E.M. et al. (2015) Planetary boundaries: Guiding human development on a changing planet. *Science* **347**: 1259855.

Steinegger, M., and Söding, J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* **35**: 1026-1028.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G. et al. (2015) Structure and function of the global ocean microbiome. *Science* **348**: 1261359.

Tang, J., Li, Y., Zhang, L., Mu, J., Jiang, Y., Fu, H. et al. (2023) Biosynthetic pathways and functions of indole-3-acetic acid in microorganisms. *Microorganisms* **11**: 2077.

Tange, O. (2018) GNU Parallel 2018, Mar 2018, ISBN 9781387509881, DOI https://doi.org/10.5281/zenodo.1146014.

Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Locey, K.J. et al. (2017) A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**: 457-463.

Tria, F.D.K., Landan, G., and Dagan, T. (2017) Phylogenetic rooting using minimal ancestor deviation. *Nature Ecology and Evolution* **1**: 193.

Tu, Q., Lin, L., Cheng, L., Deng, Y., and He, Z. (2018) NCycDB: a curated integrative database for fast and accurate metagenomic profiling of nitrogen cycling genes. *Bioinformatics* **35**: 1040-1048.

Ueno, H. (2000) Enzymatic and structural aspects on glutamate decarboxylase. *Journal of Molecular Catalysis B: Enzymatic* **10**: 67-79.

van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M. et al. (2024) Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* **42**: 243-246.

Vanni, C., Schechter, M.S., Acinas, S.G., Barberán, A., Buttigieg, P.L., Casamayor, E.O. et al. (2022) Unifying the known and unknown microbial coding sequence space. *eLife* **11**: e67667.

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M. et al. (2023) AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research* **52**: D368-D375.

Wardle, D.A., Bardgett, R.D., Klironomos, J.N., Setälä, H., van der Putten, W.H., and Wall, D.H. (2004) Ecological linkages between aboveground and belowground biota. *Science* **304**: 1629-1633.

Whelan, S., and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution* **18**: 691-699.

Wuchter, C., Abbas, B., Coolen, M.J.L., Herfort, L., van Bleijswijk, J., Timmers, P. et al. (2006) Archaeal nitrification in the ocean. *Proceedings of the National Academy of Sciences* **103**: 12317-12322.

Xu, J., and Zhang, Y. (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**: 889-895.

Xu, S., Dai, Z., Guo, P., Fu, X., Liu, S., Zhou, L. et al. (2021) ggtreeExtra: Compact visualization of richly annotated phylogenetic data. *Molecular Biology and Evolution* **38**: 4039-4042.

Yang, R., Li, S., Li, Y., Yan, Y., Fang, Y., Zou, L., and Chen, G. (2021) Bactericidal effect of *Pseudomonas oryziphila* sp. Nov., a novel *Pseudomonas* species against *Xanthomonas oryzae* reduces disease severity of bacterial leaf streak of rice. *Frontiers in Microbiology* **12**: 759536.

Yang, T., Nian, Y., Lin, H., Li, J., Lin, X., Li, T. et al. (2024) Structure and mechanism of the osmoregulated choline transporter BetT. *Science Advances* **10**: eado6229.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.Y. (2017) ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**: 28-36.

Yu, T., Cui, H., Li, J.C., Luo, Y., Jiang, G., and Zhao, H. (2023) Enzyme function prediction using contrastive learning. *Science* **379**: 1358-1363.

Zeng, J., Tu, Q., Yu, X., Qian, L., Wang, C., Shu, L. et al. (2022) PCycDB: a comprehensive and accurate database for fast analysis of phosphorus cycling genes. *Microbiome* **10**: 101.

Zhao, M., Lee, W.-P., Garrison, E.P., and Marth, G.T. (2013) SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications. *PLoS One* **8**: e82138.

Zheng, J., Ge, Q., Yan, Y., Zhang, X., Huang, L., and Yin, Y. (2023) dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Research* **51**: W115-W121.

Zhou, Z., Tran, P.Q., Martin, C., Rohwer, R.R., Baker, B.J., McMahon, K.D., and Anantharaman, K. (2025) Unravelling viral ecology and evolution over 20 years in a freshwater lake. *Nature Microbiology* **10**: 231-245.

Zhou, Z., Tran, P.Q., Breister, A.M., Liu, Y., Kieft, K., Cowley, E.S. et al. (2022) METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome* **10**: 33.