

Decoding extremophiles: insights from bioinformatics, machine learning, and data-driven approaches

Maria N. Chasapi^{1,2,†}, Nicholas Kontis^{2,†}, Robert Lehmann^{3,†}, Ruqaiya Tasneem⁴, Niketan S. Patel², Sumeer A Khan³, Xabier Martínez de Morentin³, Iro N. Chasapi^{1,5}, Eleni Aplakidou^{1,6}, Alexandros Galaras¹, Lila Aldakheel², Minjing Su², Fotis A. Baltoumas¹, Kasthuri Venkateswaran⁷, Vincenzo Lagani^{3,8}, David Gómez-Cabrero³, Jesper Tegnér^{3,9,10,11,*}, Georgios A. Pavlopoulos^{1,12,*}, Alexandre Soares Rosado^{1,2,3,*}

¹Institute for Fundamental Biomedical Research, BSRC “Alexander Fleming”, 34 Fleming Str., Vari, 16672, Greece

²Biological and Environmental Science and Engineering Division, 4700 King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

³Biomedical Sciences Division, 4700 King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

⁴National Institute of Animal Biotechnology(NIAB), Opp. Journalist Colony, Near Gowlidoddy, Extended Q City Road, Gachibowli, Hyderabad 500032, Telangana, India

⁵Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Panepistimioupolis, Ilisia, 16122 Athens, Greece

⁶Division of Basic Sciences, University of Crete Medical School, Voutes, Heraklion 70013, Greece

⁷Department of Space Studies, University of North Dakota, Grand Forks 4149 University Avenue, Grand Forks, ND 58202ND 58202-9008, P 701.777.2480, United States

⁸Institute of Chemical Biology, Ilia State University, 3/5 K. Cholokashvili Ave, Tbilisi 0162, Georgia

⁹Unit of Computational Medicine, Department of Medicine, Center for Molecular Medicine, Karolinska University Hospital, Visiongatan 18, SE-171 76 Stockholm, Sweden

¹⁰Computer, Electrical and Mathematical Sciences and Engineering Division, 4700 King Abdullah University of Science and Technology (KAUST), 23955-6900 Thuwal, Saudi Arabia

¹¹Science for Life Laboratory, Karolinska Institutet, Tomtebodavagen 23A, SE-17165 Solna, Sweden

¹²Department of Computational Biology, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Building 1B, Masdar City, SE45 05, Abu Dhabi, United Arab Emirates

*Corresponding authors. Jesper Tegnér, Biomedical Sciences Division, 4700 King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia.

E-mail: jesper.tegner@kaust.edu.sa; Georgios A. Pavlopoulos, Institute for Fundamental Biomedical Research, BSRC “Alexander Fleming”, 34 Fleming str., Vari 16672, Greece.

E-mail: pavlopoulos@fleming.gr; Alexandre Soares Rosado, Biological and Environmental Science and Engineering Division, 4700 King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. E-mail: alexandre.rosado@kaust.edu.sa

[†]Maria N. Chasapi, Nicholas Kontis, and Robert Lehmann are co-first authors.

Abstract

Life thrives in Earth's most inhospitable environments, from boiling hydrothermal vents to hypersaline lakes and frozen polar deserts, thanks to the remarkable adaptations of extremophilic microorganisms. The study of these organisms has rapidly evolved from early cultivation-based discoveries to a data-rich discipline powered by advanced omics technologies. This review comprehensively outlines the current landscape and future directions in extremophile research, emphasizing the pivotal role of bioinformatics, machine learning (ML), and data-driven approaches. We begin by charting the evolution of methodologies, from innovative *in situ* cultivation techniques and robust biomolecule extraction protocols to modern multi-omics workflows (metagenomics, transcriptomics, proteomics, and metabolomics) that decode the genetic and functional basis of extremophiles. We then catalogue essential bioinformatics resources and specialized databases critical for annotating extremophile genomes and uncovering their unique adaptive strategies, including protein stabilization and syntrophic metabolic relationships. Finally, we explore the transformative potential of artificial intelligence (AI) and ML in overcoming fundamental challenges in the field. These include predicting the functions of uncharacterized “hypothetical” proteins, identifying novel extremozymes, modeling complex genotype–phenotype relationships, and guiding the targeted engineering of industrially relevant strains. By synthesizing insights across these domains, this review highlights how integrating computational biology and AI is poised to unlock the full biotechnological potential of extremophiles and redefine the boundaries of life itself.

Keywords extremophiles, cultivation, bioinformatics, metagenomics, multi-omics, machine learning

Received: November 25, 2025. **Revised:** February 20, 2026. **Accepted:** April 15, 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 License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

Over the past century, exploration of Earth's most inhospitable niches has transformed our understanding of where life can persist and how it does so. In 1969, Brock & Freeze isolated the first nonspore-forming thermophile, *Thermus aquaticus*, from a hot spring in Yellowstone Park, USA, demonstrating robust growth near 70°C in both natural and laboratory settings [1]. However, contemporaneous anthropocentrism at the time entrenched a human-centered physiological baseline, characterized by narrow temperature (~20°C–40°C), pH (6–8), salinity (~0.1–0.5 M NaCl), and pressure ranges, such that any deviation is labeled as “extreme.” This human notion guided early expeditions away from habitats that appeared only subtly outside these human norms, delaying the discovery of meso- to polyextremophilic microorganisms [2, 3]. Within a few years, Malceroy [3] formalized the concept of organisms that prefer such conditions, naming them “lovers” (“philos” in Greek) of extreme environments. Malceroy introduced the term “extremophiles” to distinguish these (“extreme-loving” organisms) from extremotolerant taxa that merely endure stresses without requiring them.

Meanwhile, a vast deep biosphere, largely invisible at the time, was coming into view. Prokaryotic cells extend down to 5 km beneath continents and several kilometers into the oceanic crust, with recent estimates indicating that 5×10^{29} – 1.2×10^{30} cells inhabit these dark sediments. Considering that this sheer number might even surpass microbes on the surface, extremophiles may well constitute one of the most abundant lifeforms on our planet. Based on recent studies, molecular clocks further suggest that hyperthermophilic and acidophilic lineages diverged early, consistent with an Archean Earth (~4.0–2.5 Ga) where extremophily may have been the modal lifestyle [4–6].

Today, extremotolerant and extremophilic microbes have been detected under a wide range of temperatures, pH, salinity, desiccation, radiation, pressure, energy, and nutrient-limited conditions, continually redefining the boundaries of conditions under which life can thrive. Extremophiles can persist in these challenging environmental conditions due to a long history of intense adaptations driven by episodic shifts in Earth's geochemistry and climate [7]. Given that “modern” surface conditions are relatively recent on geological timescales, it is plausible that extremophile strategies helped shape core metabolic processes long before mesophilic niches expanded [8, 9].

Extremophiles are defined by the physicochemical extremes at which they grow optimally, and many types satisfy more than one criterion (polyextremophiles) (Fig. 1). Thermophiles tend to thrive between temperatures of 50°C and 80°C, while hyperthermophiles exhibit optimum growth at 80°C or higher. They stabilize proteins and membranes via heat-shock chaperones, ionic networks, and archaeal ether lipids. Psychrophiles grow optimally below 15°C with a maximum temperature of 20°C and have even been found to thrive under 0°C. They tolerate low temperatures by using highly unsaturated membranes, cold-active enzymes, antifreeze proteins, and extracellular polysaccharides. Acidophiles prefer a pH below 3 and maintain extreme proton gradients through impermeable membranes and cytoplasmic buffering. In contrast, alkaliphiles tend to grow and thrive at pH values above 9 by exploiting sodium bioenergetics, pH homeostasis transporters, and alkalistable cell walls. Xerophiles thrive at low water activity (<~0.8) by accumulating compatible solutes (trehalose, ectoine) and vitrifying sugars. At the

same time, osmophiles tolerate very high osmotic pressures, typically those of concentrated sugars, via the uptake of osmoprotectants and specialized aquaporins. Halophiles require elevated salinity, ranging from moderate (3%–15%) to extreme (15%–30%). They can counterbalance ionic stressors by accumulating potassium and maintaining acidic, salt-stabilized proteomes. Radiophiles withstand intense ionizing and Ultraviolet (UV) radiation (several kilograys) through manganese-based antioxidant systems, robust DNA repair systems, and protective carotenoids. Metallophiles tolerate or utilize high concentrations of metals and metalloids using efflux pumps, chelation, redox transformations, and biomineralization. Barophiles or piezophiles grow best at high hydrostatic pressures (≥ 10 MPa), with hyperpiezophiles requiring hadal-trench pressures (≥ 70 –100 MPa). They employ pressure-adapted enzymes, piezolytes, and fluidized membranes. Other frequently encountered categories include anaerobes and microaerophiles (oxygen extremes), endoliths inhabiting rock matrices, oligotrophs adapted to chronic nutrient limitation, and polyextremophiles such as cold, hypersaline brine specialists [7, 10, 11] (Fig. 1).

As exploration has accelerated, so too has the catalogue of extreme habitats harboring microbial life. Microorganisms have been identified in sea ice permafrost and polar regions, cold seeps and mud volcanoes, shallow-water hydrothermal vents, fumaroles, hyperacidic lakes and volcanoes, deserts and arid environments, acid mine drainage, deep-sea anoxic lakes and brines, serpentinizing environments, deep-sea sediments and trenches, soda lakes and hypersaline lakes, nuclear contaminated sites, and ophiolites. Given this variety of environments, microbial life can be found virtually anywhere where liquid water is available. The continuous discovery of microbial life in previously unexplored extremes demonstrates the highly adaptive nature of extremophiles and underscores the importance of defining life's true upper and lower limits. The combined effects of multiple stressors reveal not only biotechnological innovations but also inform astrobiological models for habitable zones beyond Earth [12, 13].

In addition to naturally occurring extreme environments, engineered systems can expose microorganisms to stress conditions that are just as severe and at times more complex. Spaceflight-associated habitats, including the International Space Station (ISS) and space mission-associated environments including spacecraft assembly facilities (SAFs), represent some of the most stringently controlled and biologically challenging environments ever studied. These habitats function as polyextreme habitats, imposing chronic stressors such as microgravity, elevated radiation, desiccation, nutrient limitation, oxidative stress, and repeated chemical disinfection. Unlike classical natural extremes (e.g. hydrothermal vents or hypersaline lakes), ISS and SAF environments are engineered extremes, where microbial persistence occurs under continuous human occupation and aggressive contamination control. Over the past 25+ years, systematic microbial surveillance and cultivation-based programs have demonstrated that these environments harbor rare but viable extremophiles, many of which are poorly represented in terrestrial reference databases [14–17]. Microorganisms associated with the ISS and spacecraft assembly cleanrooms represent a unique class of high-stress survivors shaped by microgravity, radiation, desiccation, and stringent decontamination [18]. Genome-resolved investigations of space-associated taxa have uncovered novel biosynthetic gene clusters, antimicrobial resistance determinants, and stress-adaptation systems that are rare or absent in terrestrial reference sets, broadening the landscape of functional and chemical space. Leveraging large culture collections and

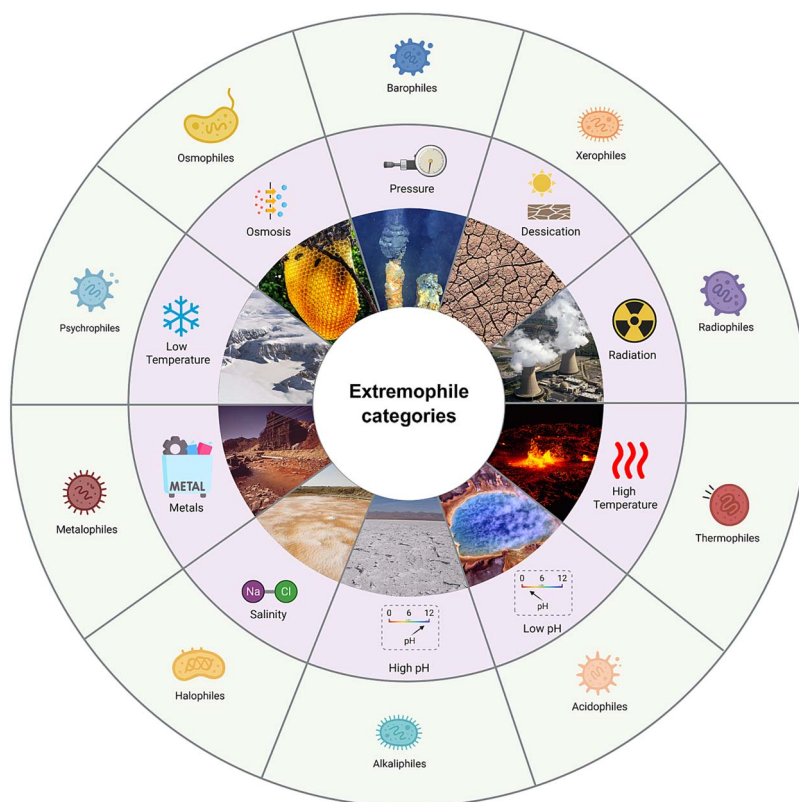


Figure 1 Main categories of extremophiles. This illustration depicts the 10 most common extremophile types, highlighting their defining physicochemical stressors and representative natural environments.

integrative data analytics allows predictive discovery of bioactive compounds and resistance mechanisms [19].

Due to the environments in which extremophiles reside, they have evolved distinctive amino acid compositions and protein structures. Such unique changes have been observed in the most prominent extremophiles categories, thermophiles, psychrophiles, halophiles, and piezophiles.

In thermophiles, proteins are often stabilized by increased hydrophobicity, forming larger hydrophobic cores, more disulfide bonds, and a greater number of ionic interactions. These features reduce protein flexibility and raise the melting temperatures, ultimately preventing denaturation at high temperatures. In contrast, psychrophiles follow the opposite trend. Their proteins are more flexible due to fewer stabilizing bonds, such as salt bridges and hydrogen bonds, while their amino acid composition favors smaller and/or polar residues. By incorporating extra surface loops alongside lower proline content, psychrophiles can avoid protein backbone solidification and maintain catalytic activity at low temperatures. On the other hand, halophiles are characterized by protein surfaces enriched in acidic amino acids, which allow them to remain soluble and functional in high salt concentrations. Interestingly, in many halophiles, proper protein folding and catalytic function can only occur at relatively high salt concentrations, which is further aided by reduced hydrophobicity and, occasionally, by the insertion of small loops that increase flexibility [20, 21].

The conserved and well-characterized metabolic enzyme malate dehydrogenase (MDH) was used to examine whether the predicted amino acid and protein structural adaptations in extremophiles are

also observed in this protein [22]. Thus, thermophilic MDH was characterized by an increase in large-chain and charged amino acids. This is correlated with a higher number of ionic interactions, which would lead to a higher overall protein stability. Halophilic adaptations were characterized by an increase in acidic, hydrophilic amino acids, accompanied by a decrease in basic amino acids. Psychrophilic adaptations were more challenging to evaluate due to the underrepresentation of MDH in psychrophiles. Despite the limited dataset, psychrophilic MDH was characterized by increased hydrophilic amino acids, which facilitate water coordination, preventing ice formation and ultimately increasing the overall flexibility through internal hydration.

While amino acid analyses can reveal trends of general protein adaptations, a structural analysis can provide deeper insights into how proteins have adapted to their environment. On average, thermophilic MDH structures are larger than their mesophilic counterparts, with only a few exceptions, suggesting that not all adaptations are used universally. Additionally, these structures were found to contain more basic residues on the surface. Conversely, all halophilic structures were smaller, very similar in volume, and included a higher number of acidic amino acids on the surface. The small psychrophile dataset still showed a significant increase in size compared to the mesophilic counterpart. The larger volume might have been caused by the rise in surface cavity and looser folding [21, 23].

In extreme habitats, the protein, amino acid, and structural adaptations mentioned above could affect ecosystem composition at the microbial community scale. These biochemical adaptations enable extremophiles to occupy niches inaccessible to ordinary organisms,

thereby shaping the microbial community by excluding organisms lacking niche-specific protein modifications.

At the microbial community scale, niche-specific adaptations among all community members have led to various metabolic dependencies, termed syntrophic relationships. Syntrophy is a cooperative interaction in which two or more microorganisms rely on each other to complete a metabolic process that neither could perform independently. This process is observed primarily in energy-poor environments, such as deep-sea sediments, hydrothermal vents, and hypersaline brines. The main interactions that support syntrophic interactions are hydrogen/formate interspecies electron transfer [24], direct interspecies electron transfer (DIET) [25], electron exchange via intercellular soluble redox shuttles such as quinones and flavins [26], and physical electron conduction through nanowires and vesicle-like appendages [27].

A fascinating example of a syntrophic relationship occurs in anaerobic methane oxidation, where anaerobic methanotrophic archaea (ANME) and sulfate-reducing bacteria (SRB) persist in deep-sea hydrothermal vents up to 60°C and in cold marine seeps. ANME can activate and oxidize methane via a reversed methanogenesis pathway, in which SRB must scavenge and reduce the intermediate products. This reaction is thermodynamically feasible only if the SRB performs the reduction reaction immediately. Due to the limited energy available in these extreme environments, DIET is more favorable for interaction between the two syntrophic partners [27].

Since their discovery, extremophiles have revealed a vast and largely underexplored reservoir of resources that could be of great benefit to biotechnology. Their long evolutionary history under extreme physicochemical conditions has allowed for the development of specialized adaptive mechanisms, enabling them to exploit the limited available resources utilizing unique enzymes, biomolecules, and secondary metabolites. As a result, extremophiles have attracted significant interest in biotechnology, contributing to the sustainability of industrial processes and supporting the transition toward a greener, bio-based economy [28]. Moreover, they have also been widely applied in bioremediation where they reduce, transform, immobilize, and detoxify contaminants in disturbed environments [29]. Across several sectors, extremophiles facilitate demanding industrial reactions using extremozymes. Their remarkable stability allows them to function under harsh conditions incompatible with conventional enzymes [9, 30]. Furthermore, the conformational resilience, flexibility, and overall robustness of extremoenzymes show promise in preventing protein misfolding and aggregation associated with neurodegenerative and cardiovascular diseases [31]. Extremophilic *Actinobacteria* have been found to produce enzymes and secondary metabolites with potent antibacterial activity against multidrug-resistant pathogens, as well as antifungal, anti-Human Immunodeficiency Virus (HIV), anticancer, and anti-inflammatory compounds [32]. Beyond their potential for biotechnological applications, the study of extremophiles can enhance our understanding of terrestrial analogue environments, refine hypotheses on the origins of life, and illuminate the evolutionary and physiological boundaries that define life on Earth [33–35]. Finally, since extremophiles were among the first pioneers of life on our planet, they remain the most promising astrobiological model organisms for pursuing extinct or extant life in extraterrestrial environments [36–39].

In this review, we first cover the evolution of methods for discovering extremophiles, ranging from classical cultivation and *in situ* techniques to advanced biomolecule extraction and multi-omics workflows. We then catalogue the critical bioinformatics

resources, datasets, and databases for interpreting the (meta)genomics of extremophiles and conclude with the transformative potential of machine learning (ML) and deep learning.

Field sampling and cultivation strategies

In situ

Extremophiles often require complex growth conditions, including interactions with other microbes and habitat-specific nutrients. Therefore, attempts to isolate them under standard laboratory conditions frequently result in the loss of many, if not all, novel extremophiles. *In situ* cultivation, although lacking the precision and control of laboratory setups, can provide access to extremely delicate extremophiles that cannot survive outside of their environment or have complex metabolic relationships with other members of their local microbial community (Fig. 2).

The process from sampling to assembly begins with environmental sampling. In the case of extremophiles, sites are targeted based on their extreme conditions and physicochemical parameters. Geothermal hot springs, hypersaline lakes, acidic rivers, highly alkaline soils, and nuclear-contaminated sites are among the extreme environments [7].

Before sampling, efforts to cultivate and capture extremophiles can take place at the location of interest. Solid and liquid media can be prepared in advance using both nutrient-rich and nutrient-poor formulations. This process can be enhanced by using portable culture chambers. By using these designs, physical conditions such as temperature, pH, and salinity can be carefully adjusted to further enhance the cultivation of fastidious microorganisms. In addition to controlled *in situ* inoculation under adjusted conditions, microbes can also be cultivated using the exact native environmental factors present at the site.

Diffusion chambers and biofilm traps are efficient and widely used methods for growing and capturing microbes directly in their natural environments. The most common diffusion systems are diffusion bioreactors and the iChip. The bioreactors are relatively large chambers with semipermeable membranes that allow microbes to grow, often in bulk, under near-natural conditions. This method enables microbes to maintain metabolic crosstalk, making it particularly useful for studying community dynamics and syntrophic relationships. The iChip employs a similar concept but is designed to contain hundreds of tiny wells on a microfabricated plate. Compared with bioreactors, this device promotes the growth of single, isolated microbial lineages, providing them with their own nutrient stream from the environment [40]. Biofilm traps, in contrast, are sterile abiotic surfaces (rocks, glass, steel, ceramics, or minerals) deployed directly into the environment. Although less complex, they offer an interesting cultivation method that targets *in situ* natural colonization by free-living microbes. This technique is mainly used to study early colonizers, biofilm-producing microbes, and microbe–substrate interactions under natural conditions [41].

Finally, *in situ* sequencing can also be performed using the well-established MinION nanopore technology, which not only enables immediate identification of novel taxa, genes, proteins, and metabolic pathways but also preserves the integrity of fragile DNA and RNA that would otherwise degrade during transport.

Cultivation

Cultivation in the laboratory can still yield significant isolation success when conducted rapidly (Fig. 2). Using only fresh samples, a few hours

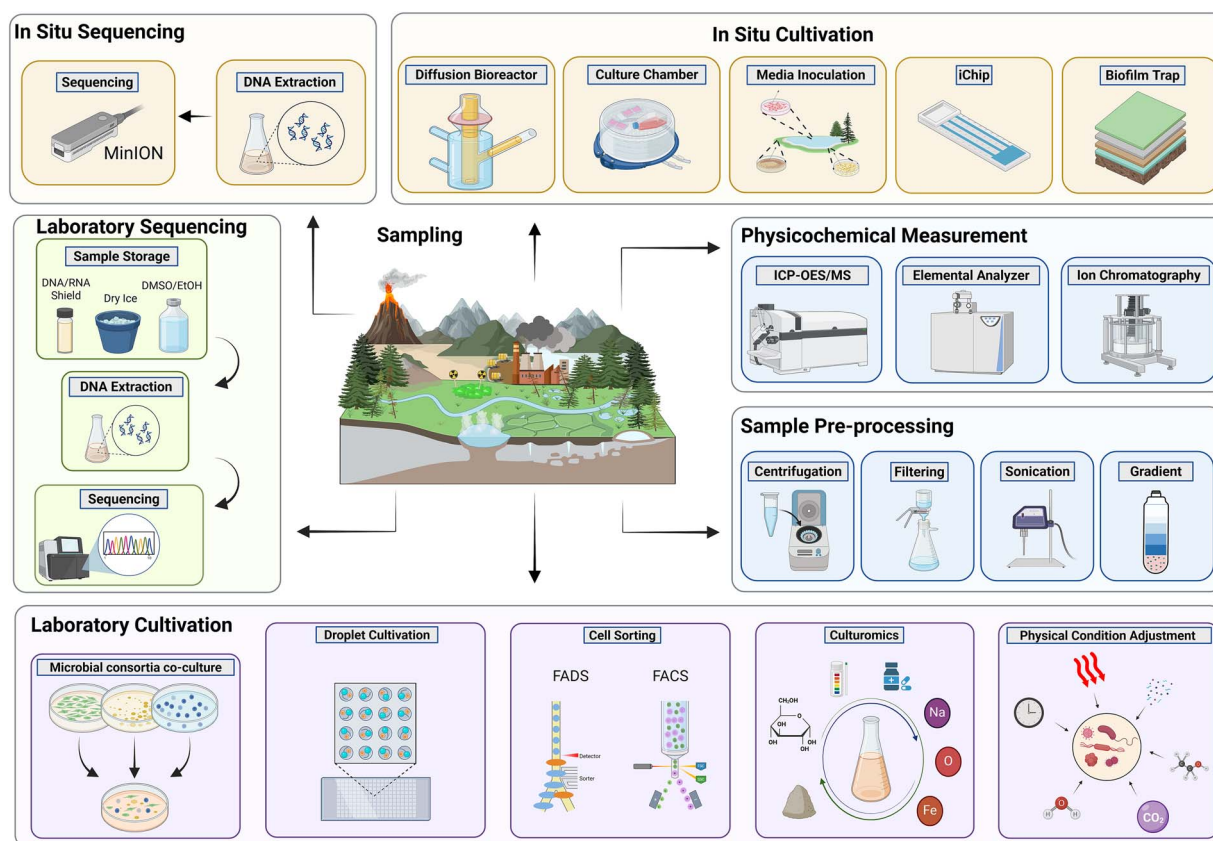


Figure 2 *In situ* and laboratory techniques for the cultivation and sequencing of extremophiles. Microbial cultivation and analysis can be performed either directly in the native environment or after samples are transported to the laboratory. *In situ* approaches include cultivation systems and nanopore-based sequencing. Physicochemical measurements and sample preprocessing provide information on sample composition and help reduce contaminants. DNA extraction and sequencing can be conducted in parallel with cultivation or independently, generating large-scale genetic and metabolic datasets that characterize native microbial communities. Laboratory cultivation strategies range from simple enrichment to advanced methods such as co-culture, droplet-based cultivation, fluorescence-activated cell sorting, and culturomics.

old, and knowing the site's physicochemical composition beforehand can significantly increase cultivation efficiency. Retrieved samples can include sediment, water, biofilm, or rock and should always be collected using sterile equipment and aseptic techniques to prevent non-native microbial contamination. The choice of sample storage depends on the downstream experiment. For cultivation purposes during the expedition, samples can be stored on ice or at ambient temperature. For sequencing, samples should be kept on dry ice and prepared in a nucleic acid stabilization solution when available [42].

To obtain fastidious microorganisms, recent efforts have focused on enhancing the culturable fraction by optimizing culture conditions. Strategies include refining physical parameters (temperature, pH, oxygen concentration, incubation time), adding selective or stimulatory compounds (e.g. antimicrobials, extracellular metabolites, pollutants, signaling molecules), applying single-cell sorting technologies and microfluidic cultivation, and designing synthetic microbial consortia with both native and non-native supportive microbes. Minor yet crucial adjustments of media composition, based on measured physicochemical parameters at the source environment, have also been shown to significantly improve microbial recovery [43].

Adjusting the physical and chemical conditions of culture media (temperature, pH, salinity, pressure, or unusual substrates) helps recreate the harsh, niche-specific conditions required by certain extremophiles [21]. Mimicking these extremes provides the selective

pressure needed for extremophiles to grow where standard media would prove to be inefficient.

Another significant factor is the presence of “helper” microbes. Microbial co-cultures can provide essential metabolites, detoxify the environment, and generate gradients strictly required by many syntrophy-dependent extremophiles. Tweaking physicochemical conditions and using co-cultures have resulted in substantial improvements in laboratory cultivation, and these can be further enhanced with microfluidic and cell-sorting approaches. Through droplet cultivation, microdroplets containing varying nutrients or conditions can encapsulate single microbial cells. This method can generate up to millions of parallel microscale incubators. This approach is analogous to the iChip, where each droplet forms a separate environment that allows slow-growing extremophiles to escape competition from fast-growing taxa [40, 44, 45]. Cell-sorting techniques such as Fluorescence-Activated Cell Sorting (FACS) and Fluorescence-Activated Droplet Sorting (FADS) can physically separate single cells based on fluorescence signals into wells or droplets. These methods enable precise single-cell isolation from complex samples. They can specifically enrich cells that show activity under stress conditions such as exposure to UV, high salinity, or heavy metals [46, 47].

Additionally, live/dead staining can be used to separate extremophiles from nonextremophiles. Other mechanisms include adding fluorescent substrates or probes linked to extremophile-specific

processes (e.g. sulfate reduction, compatible solute production) and stains that bind to protective molecules extremophiles produce under stress. Together, these act as filters to enrich extremophiles [48, 49].

Lastly, integrating multiple media formulations, physicochemical conditions, and incubation times can yield hundreds of combinations, enabling high-throughput culturing. Traditional culturomics does not require sequencing before cultivation. However, genome-informed culturomics significantly enhances the success rate by using sequencing data to design tailored media that compensate for missing pathways and supply essential metabolites. With this approach, the odds of culturing previously uncultivable extremophiles significantly increase (Fig. 2).

Biomolecule extraction

Once extremophile samples are collected and isolated, biomolecules such as DNA, RNA, proteins, and metabolites can be extracted for genomics, transcriptomics, proteomics, and metabolomics, respectively.

The initial step in all procedures involves cell lysis, achieved by the mechanical or chemical disruption of the cell wall. DNA can be isolated using silica-based columns, magnetic beads, or organic extractions followed by alcohol precipitation and phenol–chloroform separation. Next, purity and concentration can be assessed using spectrophotometric measurements and fluorometric assays, while fragment size distribution and integrity are more accurately evaluated with capillary electrophoresis systems [50].

DNA extraction from extremophiles generally follows the same routine protocols and kits used for other microbes. However, protocol modifications have been shown to improve the yield and quality of extremophile DNA. Many extremophiles, particularly archaea and spore-forming bacteria, possess robust cell envelopes [51], inhabit highly inaccessible niches (e.g. thermophiles in biofilms, halophiles within salt crystals) [7], and are often in the presence of inhibitory substances that can co-purify with DNA and interfere with downstream applications [52].

To improve the recovery of high-molecular-weight (HMW) DNA suitable for long-read sequencing, several adaptations can be employed. These include extended enzymatic lysis, the use of protective additives [53], gentle mechanical disruption [54], and agarose plug methods [55]. These modifications, either in place of or in combination with traditional DNA extraction techniques, have been shown to increase DNA yield and fragment size while reducing shearing.

Additionally, selective lysis strategies can enhance DNA recovery from diverse cell types. Tailored enzyme cocktails can be used to ensure comprehensive cell lysis (e.g. MetaPolyzyme) [56], and lysis conditions can be adapted to exploit natural vulnerabilities, such as inducing autolysis in halophilic archaea through osmotic shock [57]. Indirect extraction methods, in which cells are first pelleted and then lysed in a controlled buffer, can also reduce the co-extraction of inhibitors [58]. In exceptionally resilient taxa, such as deep-sea archaea, hot alkaline lysis has been employed effectively to recover DNA from otherwise inaccessible cells [59]. Environmental samples from extreme habitats often contain inhibitory substances, such as humic and fulvic acids, or high salt concentrations, which can co-purify with DNA and interfere with polymerase chain reaction (PCR), library preparation, and sequencing. To avoid or limit this, additional steps can be incorporated. These include pre-extraction washing and sterilization (e.g. surface sterilization of halite crystals with hydrochloric acid and bleach), desalting and concentration via size-exclusion

chromatography [60], and the use of detergents and precipitation agents such as cetyltrimethylammonium bromide, to remove humic substances [61]. The implementation and refinement of these strategies will continue to expand our understanding of extremophilic communities and our ability to study life in Earth's most challenging environments. After nucleic acid extraction, standard library preparation protocols [62] can be followed to proceed to sequencing. Once sequencing data are obtained, they can be processed for a wide range of downstream analyses, including assembly, annotation, and comparative genomics.

Protein extraction requires solubilization with detergents or chaotropic agents, followed by clarification through centrifugation or with various precipitation techniques. This is then followed by their extraction using mixtures of methanol, acetonitrile, chloroform, or water to capture both polar and nonpolar compounds. A range of analytical methods can be used to confirm the quantity and quality of the biomolecules of interest. Proteins can be quantified using colorimetric assays. Their quality can be verified through immunoprecipitation for specific target recovery and Western blotting for detection of expression and integrity. Metabolites are profiled and quantified by chromatography-coupled mass spectrometry or Nuclear Magnetic Resonance (NMR) spectroscopy.

Extremophiles in the omics era

In extremophile research, samples are collected from chemically harsh environments, where DNA degradation, contamination, and limited usable sequencing data are common. Shotgun metagenomics is the gold standard for uncovering and characterizing the taxonomic distribution of microbial communities in different environments. Recent advances in sequencing technologies and bioinformatic tools for decontamination, *de novo* assembly, binning, and taxonomic classification have enabled the detection of low-abundance and rare taxa with unprecedented resolution [63, 64].

Depending on DNA quantity, integrity, and environmental chemistry, researchers use short-read sequencing, such as Illumina, to obtain highly accurate, cost-effective data suitable for metagenomic surveys and taxonomic profiling, though these methods often produce fragmented assemblies. Conversely, long-read sequencing technologies, including Oxford Nanopore Technologies (ONT) and PacBio, generate reads spanning several kilobases, enabling improved genome assembly and the resolution of repetitive regions, plasmids, and structural variants. The ONT MinION platform, which performs both barcoding and shotgun metagenomic sequencing, is ideal for extreme environments where the microbial community is sensitive and can provide immediate feedback on sample quality and community composition. However, if the MinION platform is unavailable and higher sequencing accuracy is desired, established laboratory sequencing platforms are used. In challenging environments where DNA fragmentation or low biomass limits yield, hybrid sequencing approaches that combine long- and short-read data can maximize assembly continuity and accuracy [65]. Targeted sequencing approaches, such as 16S ribosomal ribonucleic acid (rRNA) gene amplicon sequencing on Illumina MiSeq or Ion Torrent platforms, are also widely used in extremophile studies to characterize microbial community composition when DNA yield is low or reconstruction is impractical [66].

Following sequencing, strict quality control (QC) thresholds and host decontamination are essential to remove bias. To achieve reliable downstream analyses, quality control is performed on

raw sequencing data first. Then reads are mapped to reference host genomes using alignment tools to remove contaminant DNA. Following preprocessing, high-quality reads are assembled into contigs to reconstruct longer genomic fragments, which are then grouped into metagenome-assembled genomes (MAGs) using binning algorithms [67]. Given the low biomass of the microbial community of interest, the ideal approach is a hybrid strategy combining long and short reads. However, depending on the environment's severity, fragmentation of the total DNA can be high, limiting sequencing to short reads. The completeness and reliability of MAGs are evaluated, and then dereplication is performed to generate a nonredundant set of high-quality genomes. MAGs and isolated genomes are then taxonomically classified using a variety of pipelines and viewers [68, 69]. Following metagenomic assembly and recovery of MAGs, functional insights are gained through genome annotation and multi-omics approaches, which provide the functional and evolutionary context needed to interpret extremophile adaptation [70].

Genome annotation integrates multiple reference resources to assign functions to predicted genes. The first step of the workflow is protein prediction. This process identifies potential protein-coding regions, known as Open Reading Frames, on MAGs or scaffolds using tools such as Prodigal. The next step is to identify and functionally characterize the coding sequences, as well as transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs). This step is particularly challenging in extremophile research because many genes encode highly divergent proteins with few homologs in existing databases. Functional assignment can be achieved by tools that perform sequence similarity searches against different databases. Still, in this case, annotation relies more heavily on hidden Markov model profile searches, which detect conserved protein domains and motifs [71, 72].

Once functionality is established, genes are mapped to known metabolic pathways and enzyme classifications, resulting in the reconstruction of metabolic pathways. This process highlights metabolic traits essential for surviving in extreme environments, such as stress tolerance, energy acquisition, or the production of extremozymes. It thus provides insight into the functional potential of extremophilic organisms. To place MAGs in a phylogenetic context, first conserved marker genes are identified and aligned, followed by multiple sequence alignment. Phylogenetic trees are then constructed using model selection to optimize accuracy and can provide insights into evolutionary relationships, horizontal gene transfer, and the identification of novel lineages. This type of analysis of extremophile MAGs often uncovers deeply branching lineages, lateral gene transfer events, and novel taxa, expanding our understanding of microbial diversity in extreme ecosystems [62].

Unlike shotgun metagenomics, which sequences all genomic material and often involves complex genome assembly, 16S rRNA gene amplicon sequencing follows a more targeted approach. Specifically, it focuses on amplifying and sequencing a specific, highly conserved genomic region, the 16S rRNA gene, that serves as a barcode for prokaryotic organisms. These data bypass genome assembly entirely; instead, the raw sequencing reads undergo denoising, which outputs a set of amplicon sequence variants. These variants are unique biological sequences that can be resolved down to a single-nucleotide difference and provide strain-level insights into the microbial community [73] (Fig. 3).

Even though metagenomic analyses remain the primary strategy for identifying novel genes and classifying microorganisms in extreme environments, the presence of a gene does not guarantee

its expression or translation into a functional protein. Therefore, metatranscriptomics and metaproteomics have emerged as complementary approaches to uncover the functional microbial landscape of extremophiles.

Metatranscriptomics enables the characterization of gene expression profiles of microbial communities, ultimately revealing the molecular cascades underlying stress responses and adaptation mechanisms in extreme conditions. A typical metatranscriptomic workflow includes quality control, trimming (if necessary), decontamination, *de novo* assembly, transcript-level taxonomy, and functional annotation [74]. Moreover, differential expression analysis using standard RNA-sequencing algorithms, such as DESeq2 [75] and edgeR [76], enables researchers to compare gene expression programs across different conditions. Despite their importance, metatranscriptomics remains technically challenging because intact RNA is difficult to isolate from these environmental samples. Nevertheless, they have been implemented in several studies. For instance, metatranscriptomic studies have elucidated the metabolic processes underlying carbon mineralization in active methanogenic communities [77], characterized altered metabolic activity in hydrothermal vents based on the distance from the vent [78], and demonstrated that archaea and bacteria follow distinct osmoadaptation strategies in hypersaline environments [79]. Collectively, these findings highlight the importance of metatranscriptomics as a complementary tool in addressing extremophiles (Fig. 3).

Accordingly, metaproteomics is used in combination with metagenomics to characterize the protein landscape of microbial communities. Samples are processed to extract proteins, which are then digested into peptides and analyzed by mass spectrometry (bottom-up approach). Then, the resulting spectra are matched to appropriate reference databases either from public repositories or based on metagenomic assembled sequences. The identified proteins can be further used for taxonomic assignment, functional annotation, quantification, and differential expression analysis [80]. Similarly, mass spectrometry can be used to profile microbial communities to characterize small-molecule metabolites and the underlying biochemical pathways involved in microbial adaptation to extreme conditions [81, 82].

Metaproteomics offers deeper insights into the functional landscape of microbial taxa, enabling the comprehensive cataloguing of community composition, characterization of the metabolic processes that support survival under extreme conditions, and the discovery of novel enzymes with potential biotechnological applications [64]. For instance, a previous study integrated metaproteomics and metagenomics to characterize the microbial biodiversity of an Andean high-altitude lake [83]. Similarly, metaproteomic analyses revealed unique ecophysiological adaptations in haloarchaea that enable them to efficiently utilize available light and survive under nutrient-deprived conditions in a hypersaline Antarctic lake [84]. In hydrothermal vent systems, metaproteomics have shown that Gammaproteobacteria dominate inactive chimney communities as autotrophic sulfide oxidizers capable of metal-sulfide dissolution via extracellular electron transfer [85]. In contrast, an earlier study applied metaproteomics to map the microbial communities inhabiting actively venting hydrothermal chimneys [86]. Additionally, multi-omic approaches could lead to an exhaustive characterization of extremophiles, as demonstrated for *T. filiformis*, where integrative analyses uncovered multiple thermostable proteins and revealed its adaptation mechanisms at different temperatures [87] (Fig. 3).

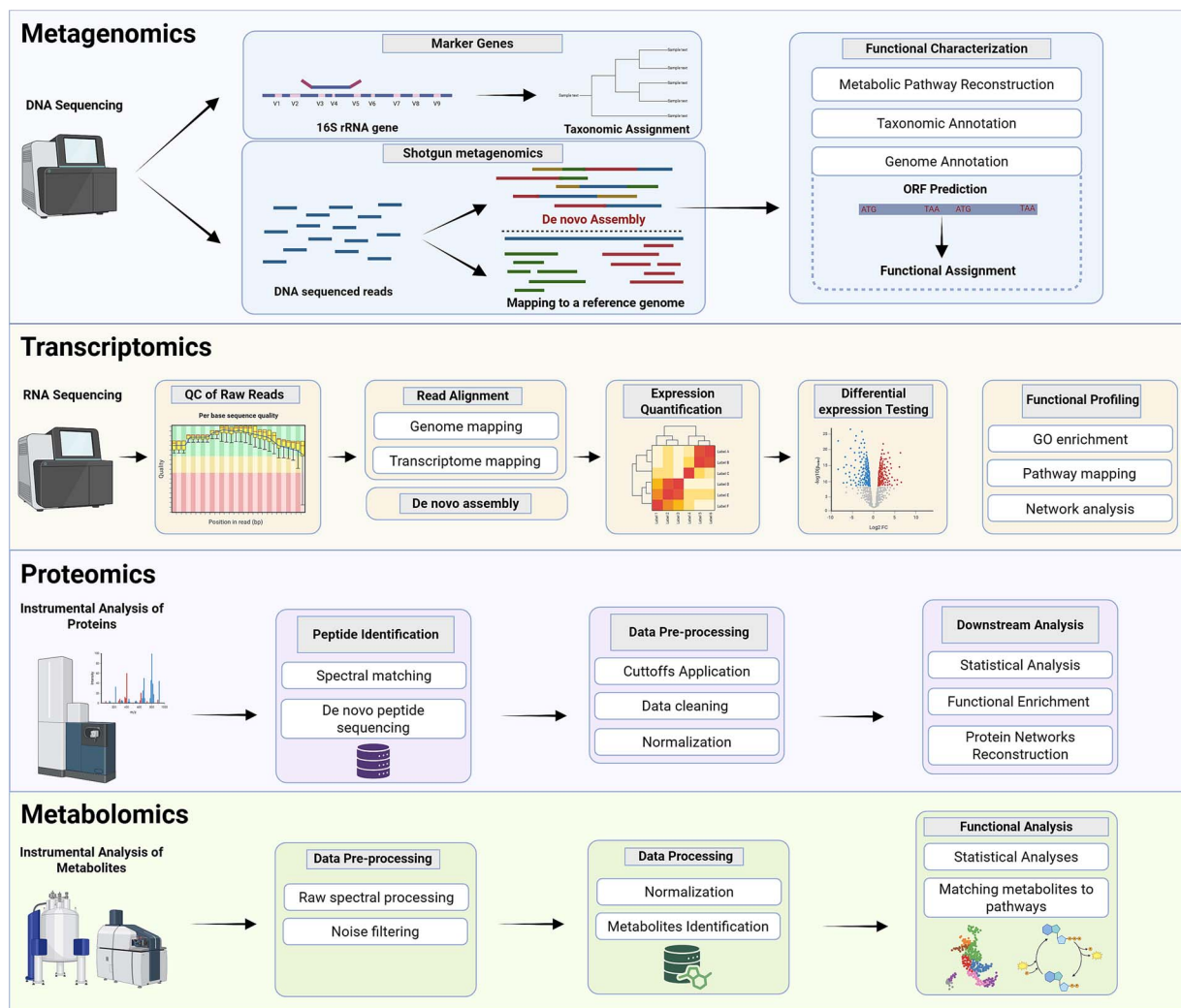


Figure 3 Multi-omics workflows for extremophile research. Metagenomic data can be processed either through shotgun metagenomics, enabling *de novo* assembly and genome-centric analyses, or through rRNA marker gene sequencing for taxonomic assignment. Reconstructed genomes are then subjected to functional annotation, metabolic pathway reconstruction, and phylogenomic analysis. Transcriptomics begins with RNA sequencing followed by read quality control, alignment or *de novo* assembly, expression quantification, differential expression testing, and functional profiling through GO enrichment, pathway mapping, and network analysis. Both proteomics and metabolomics are based on mass spectrometry. Metabolomic analysis includes spectral data pre-processing, metabolite identification, and functional analysis. Proteomics includes peptide identification, data preprocessing, and downstream statistical and functional analyses to characterize protein abundance and changes.

Bioinformatics resources

The workflows and techniques discussed so far are widely used for genome annotation, taxonomic classification, and functional characterization and are inherently dependent on reference databases (Table 1). These resources provide the necessary information to make predictions and assign annotations. They are crucial in extremophile research, where a lack of reference data often leads to genes and their products being characterized as “hypothetical.”

Database search is a preliminary step in identifying microbial communities found in any environment. The National Center for Biotechnology Information (NCBI) GenBank [127] and RefSeq [128] host nucleotide sequence data and standardized reference genomes, while UniProt [143], InterPro [114], Pfam [131], and TIGRFAMs [141] hold functional annotations of proteins and domains. At the same time, Gene Ontology (GO) [103] provides a functional classification of genes. KEGG [118] houses information about metabolic, signaling,

and regulatory pathways, and the Clusters of Orthologous Groups (COGs)/eggNOG database [97, 98] organizes proteins based on evolutionary relationships.

For community-level analyses, IMG/M [113] and MGnify [123] provide data and pipelines for metagenomic data across diverse environments, while JGI GOLD [116] tracks global genome and metagenome sequencing projects. The Sequence Read Archive (SRA) [135] and the European Nucleotide Archive (ENA) [101] store raw sequencing datasets for reuse. TerrestrialMetagenomeDB [139] and MarineMetagenomeDB [121] host metagenomic datasets from terrestrial and marine ecosystems, and the Marine Metagenomics Portal (MMP) [120] provides taxonomic and functional annotations for oceanic datasets. Comparative studies are enriched by tools such as GlobDB [104], which gathers data from over 14 microbial genome catalogs, and Ensembl Bacteria [100], which provides a genome browser for archaeal and bacterial genomes. SILVA [136]

Table 1 Genomic, proteomic, and metagenomic databases that support data-driven analyses of extremophiles and their adaptations.

Databases	Description	Last update date	API access	Batch download capability	Access barriers
1000springs [88]	Genomic, geochemical, and microbiome data from over 1000 hot spring samples in New Zealand.	2018	No	No	Authorized access only
AcidB [89]	Acidophile genomes.	2020	No	No	Open access
BacDive [90]	Strain-level database of bacteria and archaea, providing taxonomy, physiology, morphology, environment, growth conditions, and genomic metadata.	2024	Yes	Yes	Free registration required
BacMet [91]	Metal resistance genes.	2023	No	Yes	Open access
BRENDA [92]	Enzyme database with curated data on functions, substrates, kinetics, and organism-specific properties.	2024	Yes	Yes	Open access
BV-BRC [93]	Workspace that includes a database of pathogenic bacteria and viruses' genomes, transcriptomes, protein features, protein structures, and metabolic processes.	2024	Yes	Yes	Account required
CAMPR3 [94]	Database and predictor for AMP classes.	2015	No	Partial	Open access
CARD [95]	Antimicrobial resistance gene database.	2024	Yes	Yes	Open access
CAZy [96]	Carbohydrate-active enzymes.	2024	No	Partial	Open access
COG/eggNOG [97, 98]	A database that consists of phylogenetically classified proteins based on their evolutionary relationships, which form COGs.	2024	Yes	Yes	Open access
DisProt database [99]	Database for Intrinsically Disordered Proteins.	2024	No	Yes	Open access
Ensembl Bacteria [100]	Browser for bacterial and archaeal genomes.	2024	Yes	Yes	Open access
European Nucleotide Archive (ENA) [101]	Public repository for nucleotide sequencing data.	Continuous	Yes	Yes	Open access
EzBioCloud [102]	16S rRNA database with taxonomic hierarchy.	2024	No	Partial	Subscription required
Gene Ontology (GO) [103]	Assigns functions to the genes by defining classes of gene functions that have specified relations to each other, representing gene products' properties.	Continuous	Yes	Yes	Open access
GlobDB [104]	Combines and dereplicates over 14 major microbial genome catalogs.	2025	No	Yes	Open access
GTDB (Genome Taxonomy Database) [105]	Standardized microbial taxonomy based on phylogenomics.	2025	No	Yes	Open access
HaloDom [106]	Halophile database.	2024	No	No	Open access
HaloWeb [107]	Genomic and metabolic pathway data for halophiles.	2022	No	No	Open access
ICEberg 3.0 [108]	Integrative and conjugative elements.	2023	No	Yes	Open access
IMG/VR 4.0 [109] and MetaVR [110]	Repository of viral genome and metagenome data.	2025	Yes	Yes	Open access
IND-enzymes [111]	Thermophilic/psychrophilic hydrolytic enzymes.	2021	No	Yes	Open access
INDIGO [112]	Red Sea extremophile genomics with the AAMG pipeline.	2013	Yes	No	Open access
Integrated Microbial Genomes & Microbiomes (IMG/M) [113]	Platform for microbial genome and metagenome data with standardized annotations and comparative analysis tools.	2025	Yes	Yes	Account required
InterPro [114]	Integrates multiple protein signature databases to classify protein families, predict domains, and assign GO terms.	2025	Yes	Yes	Open access
ISFinder [115]	Database of manually curated prokaryotic Insert Sequence elements.	2025	No	Yes with written authorization	Open access
JGI GOLD [116]	Genome and metagenome sequencing projects, and their metadata.	2024	Yes	No	Account required
KBase (DOE Systems Biology Knowledgebase) [117]	Combines data, tools, and workflows for analyzing microbes, plants, and microbiomes.	2026	Yes	Yes	Account required
KEGG [118]	Reference resource hub for assigning genes or gene products to a specific pathway.	Continuous	Yes	Yes	FTP is paid subscription
KAUST Metagenomic Analysis Platform (KMAP) [119]	Platform for taxonomic and functional profiling of microbial communities	2025	No	Yes	Authorized access only
Marine Metagenomics Portal (MMP) [120]	Marine metagenomics data, integrating public datasets and tools for exploration, including taxonomic and functional annotations.	2018	No	Yes	Open access

(continued)

Table 1 Continued.

Databases	Description	Last update date	API access	Batch download capability	Access barriers
MarineMetagenomeDB [121]	Metagenomic datasets from marine ecosystems, including metadata like depth, salinity, and geographic coordinates.	Biannually	No	Yes	Open access
metaMicrobesOnline [122]	Integrated platform for comparative genomics, functional and expression analysis with built-in tools.	2011	No	Yes	Open access
MGnify [123]	Taxonomic, functional, and pathway annotations for raw reads and assembled data across diverse environments.	2023	Yes	Yes	Open access
MiBIG [124]	Secondary metabolite gene clusters.	2024	No	Yes	Open access
Microbe Atlas [125]	A global microbial biogeography platform for exploring environmental distributions of taxa based on public sequencing data.	2020	No	Yes	Open access
MicroScope [126]	Microbial genome annotation and comparative genomics.	2025	Yes	Yes	Account required
NCBI GenBank [127]	Reference sequences are maintained and standardized by NCBI.	Continuous	Yes	Yes	Open access
NCBI RefSeq [128]	Archival database of nucleotide sequences submitted directly by researchers, often with varying annotation quality.	Continuous	Yes	Yes	Open access
NMPFamsDB [129]	Novel protein families.	2023	Yes	Yes	Open access
PAZy [130]	Specialized database for plastic-active enzymes	2025	No	Yes	Open access
Pfam [131]	Protein families and domains.	2024	Yes	Yes	Open access via InterPro
PSORTdb [132]	Protein subcellular localization database for bacteria & archaea, helpful in studying adaptation mechanisms in extremophiles.	2021	No	Yes	Open access
SDADB [133]	Functional annotation database of structural domains in proteins.	2018	No	Yes	Open access
SEED [134]	A database that stores annotated genomes produced by the RAST web server, along with manually annotated genomes.	Continuous	Yes	Yes	Optional login
Sequence Read Archive (SRA) [135]	Archive for raw sequencing data.	Continuous	Yes	Yes	Open access
SILVA [136]	Database of rRNA sequences from bacteria, archaea, and eukaryotes.	2026	No	Yes	Open access
TADB 3.0 [137]	Database of bacterial toxin-antitoxin systems with curated loci, families, and functional annotations.	2024	No	Yes	Open access
TEMPURA [138]	Database of thermophilic and psychrophilic protein adaptations.	2020	No	Yes	Open access
TerrestrialMetagenomeDB [139]	Metagenomic datasets from terrestrial environments, annotated with metadata.	2021	No	Yes	Open access
ThermoBase [140]	Thermophile/hyperthermophile database.	2022	No	Yes	Open access
TIGRFAMS [141]	Database of protein family definitions.	2018	No	Yes	Open access on NCBI
Transporter Classification Database (TCDB) [142]	Classifies transporters and supports genome annotation.	2020	No	Yes	Open access
UniProt [143]	A database that stores protein sequences and functional information.	Continuous	Yes	Yes	Open access
VFDB [144]	Virulence factor database.	2024	No	Yes	Open access
WASPS [145]	Plasmid sequence database.	Continuous	No	No	Open access

and EzBioCloud [102] are resources that support 16S rRNA-based taxonomic profiling, followed by BacDive [90] for detailed strain-level metadata for bacterial isolates. KBase [117] and metaVR [110] further expand functional modeling and viral metagenomic analysis.

MiBIG [124] contains biosynthetic gene cluster data, DisProt [99] is a repository of intrinsically disordered proteins, and NMPFamsDB [129, 146] is a database of novel protein families from microbial metagenomes and metatranscriptomes. PSORTdb [132] and the Transporter Classification Database [142] provide information on

protein subcellular localization and membrane transport proteins, respectively.

The need for a specialized database arose from the unique adaptations observed in extremophiles. ThermoBase [140] contains descriptions of thermophilic or hyperthermophilic organisms, whereas TEMPURA [138] stores prokaryotic growth temperatures. Halophile genomic and metabolic data are available in HaloWeb [107] and HaloDom [106], while acidophile genomes are stored in AcIDB [89]. The 1000springs [88] repository contains genomic,

geochemical, microbial, and physicochemical data derived from hot-spring samples, whereas INDIGO [112] is explicitly focused on Red Sea extremophiles. IND-enzymes [111], BRENDA [92], CAZy [96], and SDADB [133] host not only extremophile genomes but also data on enzymes derived from extremophiles. PAZy [130] is a specialized database that contains only biochemically characterized plastic-active enzymes. Additionally, metaMicrobesOnline [122] is a platform for comparative genomics and functional analyses that includes tools for studying microbial adaptations across a variety of environments, including extreme environments.

Mobile genetic elements and resistance traits greatly influence extremophile resilience. ISFinder [115] and ICEberg 3.0 [108] contain information about insertion sequences, integrative elements, and transposons. Also, antimicrobial resistance and virulence data are available through CARD [95] and VFDB [144], respectively, while BacMet [91] hosts metal resistance genes, and WASPS contains plasmid sequence data.

The opportunity of artificial intelligence in extremophile research

As outlined above, advances in experimental high throughput “omics” technologies have generated a large volume of data, offering fertile ground for discovering novel biological mechanisms and industrially valuable enzymes. However, extremophile research is hampered by two persistent bottlenecks: most extremophilic organisms remain unculturable, leaving their genomes uncharacterized and underrepresented in current data collections. And even when input material is available, extremophile omics pipelines routinely contend with limited biomass, DNA damage, contamination, and fragmented assemblies. This implies that descriptive techniques and pipelines developed [147] for readily available mesophilic data are unlikely to generalize to extremophiles. Accordingly, the field would benefit from a paradigm shift toward data-driven, predictive models, posing the fundamental challenge of effectively training them when extremophile datasets are sparse and opportunities to leverage information from other biological systems are limited.

A recurring obstacle in extremophile research is the scarcity of high-quality labels: cultivation is difficult, functional assays are expensive, and environmental samples can be low in biomass and noisy. As a result, many tasks of interest (e.g. predicting thermostability/halotolerance, optimal growth conditions, enzyme pH/temperature optima, or mutation effects for extremozymes) operate in a regime where classical train-from-scratch supervised learning is unreliable. Three complementary ML strategies are particularly relevant in these settings:

(i) *Transfer learning*. Large protein language models (pLMs) pre-trained on massive sequence corpora provide general representations that can be fine-tuned with relatively small extremophile-specific datasets, improving performance when labeled examples are limited. For example, TemBERTure predicts thermostability class and melting temperature from protein sequences and emphasizes the importance of training diversity for robust generalization [148].

(ii) *Few-shot learning and meta-learning*. When only tens of labeled variants (or a small number of characterized proteins) are available, few-shot methods adapt models to a new target with minimal data. For instance, FSFP improves protein fitness prediction under extreme data scarcity using only tens of labeled mutants [149], a setting that aligns

well with extremophile enzyme engineering, where experimental labels are costly.

(iii) *Multi-task learning*. Extremophile datasets are often thin per task (few measurements per enzyme), but related properties co-vary (e.g. stability, kinetics, temperature/pH dependence, and solubility). Multi-task models can share statistical strength across endpoints, improving sample efficiency and calibration under data scarcity. For example, MPEK is a pretrained multitask framework that jointly predicts enzyme kinetic parameters while conditioning on factors such as pH and temperature [150].

Taken together, these strategies provide a practical path for data-scarce extremophile studies: pretrain broadly (self-supervised on sequences/reads), then adapt with transfer learning, specialize with few-shot or meta-learning when labels are minimal, and stabilize learning via multi-task objectives when multiple weakly labeled endpoints exist.

Machine learning and artificial intelligence in modern biology

ML, a specific subset of artificial intelligence (AI), has become a central paradigm in modern science, fundamentally reshaping how complex biological systems [151] are analyzed and modeled [152]. At their core, ML approaches involve algorithms that identify dependency patterns among measured variables, learn representations, and make predictions directly from data [153], rather than relying on explicit, manually crafted rules. Specifically, ML refers to algorithms that improve task performance by learning from data, typically by optimizing the parameters of neural networks, decision-tree ensembles, or probabilistic models to minimize a loss function in contrast to hypothesis-driven statistical inference, where predefined mechanistic hypotheses guide model formulation, ML methods directly learn complex relationships from data, postponing the interpretation of underlying mechanisms to subsequent analysis. Notably, recent advances have demonstrated that data-driven ML methods can, under certain conditions, uncover causal relationships [154]. This capability enables the model to detect hidden organization in datasets that are too large, multidimensional, or noisy for conventional approaches and human inspection [155]. Throughout this review, we use ‘ML’ to refer to data-driven modeling (including deep and foundation models) and ‘AI’ to refer to broader systems that integrate ML with hypothesis generation, experiment selection, or automation.

Even within the field of neural network research, a wide range of architectures has been developed to capture the hierarchical and relational structure of different biological data modalities (Table 2). For instance, Convolutional Neural Networks (CNNs) extract local features such as sequence motifs or spatial patterns in molecular structures; recurrent and transformer models capture long-range dependencies in genomic or proteomic sequences; and graph neural networks (GNNs) represent the topology of biological systems, from spatially resolved transcriptomic patterns and protein–protein interaction maps to metabolic pathways. These architectures—and hybrids combining their strengths—have enabled remarkable advances in protein-level prediction, linking amino-acid sequences to structural and functional properties through general protein language models such as AlphaFold [157], ESM-2 [193], and ProtTrans [194]. In parallel, application-specific architectures have been applied to variant and property prediction, estimating functions [195] (e.g. DeepFri), as

Table 2 Machine learning based tools with potential use in extremophile biology.

Tool	Description/use	Type	Input data type	Primary output/prediction	Limitations
AI4AMP [156]	Uses hybrid encoding and CNN to predict AMP activity.	Deep Learning	Protein sequence	Prediction scores, giving the probability of how much a particular peptide contains AMP activity.	Curated training sets required
AlphaFold [157]	Predicts protein 3D structure from sequence using transformer-based models.	Deep Learning	Protein sequence/alignments	3D protein structure	Does not capture conformational dynamics
AMPLify [158]	Identifies AMPs with long-range dependencies.	Deep learning	Protein sequence	AMP probability score	Performance is limited by lack of detailed AMP training data
AMPScanner [159]	Trains on known AMPs to predict novel AMP candidates from protein sequences.	Deep learning	Protein sequence	Sequences of AMP candidates	Sequences with size <50–60aa
Aurora [160]	Associates bacterial genotypes with environmental adaptation via GWAS-based ML.	Machine Learning	Pangenome matrix & phylogenetic distance matrices	Ranked genes based on association score & model performance tables	Usability is still dependent on the R ecosystem and user expertise
CONCOCT [161]	Uses Gaussian mixture models for contig binning from metagenomes.	Machine Learning	Subcontigs	Clustering file assigning contigs into bins, which can be converted into a single file consisting of a MAG	Contigs need to be cut up prior to binning; reads from all samples should be mapped to a single shared assembly
DeepARG [162]	Predicts antibiotic resistance genes from sequence data using deep neural networks.	Deep Learning	Nucleotide or protein sequences	ARGs with high and low probability	Performs better with assembled contigs
DeepBGC [163]	Detects BGCs in bacterial and fungal genomes	Deep Learning	Nucleotide sequences	BGC regions and class predictions	Sequences with length > 5000 bp
DeepEC [164]	Assigns EC numbers to enzymes using deep learning models.	Deep Learning	Protein sequences	EC numbers and statistics	Limited performance on multifunctional or novel enzymes
DeepGO [165]	Predicts protein functions using sequence and interaction networks, incorporating GO hierarchy.	Deep Learning	Protein sequences	GO term predictions	Easier to work on established annotated organisms
DeepGOMeta [166]	Predicts protein functions using sequence and interaction networks, incorporating GO hierarchy.	Deep Learning	OTU table of relative abundance & protein sequences	GO term predictions	Restricted training scope and incomplete benchmarking context
DeepMicrobes [167]	Taxonomic classification for metagenomics	Deep Learning	Reads & nucleotide sequences	Predicted taxon and abundance	Can be used with specific species; otherwise, it requires retraining
EukRep [168]	Separates eukaryotic and prokaryotic sequences.	Deep Learning	Nucleotide sequences	Gene predictions & Eukaryotic bins	Reduced accuracy on short or low-complexity contigs
evoMIL [169]	Predicts virus–host interactions with transformer architectures.	Deep Learning	Protein sequences	Host label and probability scores	Predictions are limited to hosts with sufficient known virus host associations
geNomad [170]	Identifies mobile genetic elements in metagenomic and genomic data.	Machine Learning	Nucleotide sequences	Viral and MGE annotations	May misclassify novel or highly fragmented elements
GLIMMER [171]	Identification of coding sequences within bacterial, archaeal, and viral genomes	Markov models	Nucleotide sequences	Predicted protein-coding genes (CDS)	Accuracy decreases on error prone or highly fragmented sequences, and the method is computationally intensive
GraphEC [172]	Assigns EC numbers to enzymes using deep learning models.	Deep Learning	Protein (enzyme) sequences	Predicted EC numbers, active sites, and optimum pH of the enzyme	Accuracy is constrained by predicted structural quality and available training features
LookingGlass [173]	Context-aware DNA embeddings predict enzyme function, including optimal temperature for extremophiles.	Deep Learning	Nucleotide sequences	DNA sequence embeddings	Requires task-specific fine-tuning and may show domain bias
Macrel [174]	AMP discovery and MIC prediction from metagenomes and genome sequences.	Machine Learning	Protein sequences	AMP probability score	Emphasizes high specificity at the cost of reduced sensitivity

(continued)

Table 2 Continued.

Tool	Description/use	Type	Input data type	Primary output/prediction	Limitations
MaxBin2 [175]	Binning using the expectation–maximization algorithm based on sequence composition and abundance.	Machine Learning	Metagenomic assembly and sequencing reads	Bin assignments and characteristics	Contigs should be preferably of 1000 bp or larger
MetaBAT2 [176]	Unsupervised binning using probabilistic distances; suited for complex communities.	Machine Learning	Contigs	Contig clusters and bins/draft genomes	Dependency on multiple samples
MetageNN [177]	Taxonomic classification for metagenomics	Deep Learning	<i>K</i> -mer profiles and DNA sequence	Genus-level classification	Sequence length bias
MetaMIC [178]	Detects misassemblies in metagenomic assemblies using deep neural networks.	Deep Learning	Raw reads, contigs	Misasassembly identification and corrected assemblies	metaMIC only corrects contigs that can be split into fragments longer than 1 kb.
MetaTransformer [179]	Taxonomic classification for metagenomics	Deep Learning	Raw DNA reads (short-read data)	Genus-level and species-level classification, and abundance estimation	Updateability and memory constraints (increasing <i>k</i> -mer increases memory consumption)
PanPhlAn [180]	Pan-genome-based strain-level microbial profiling.	Machine Learning	Raw reads	Gene presence/absence matrix based on the pangenome	Novel genes and species that might not be present in their pangenome database could be missed
PhyloPythiaS [181]	Performs taxonomic binning of metagenomic contigs using sequence composition and ML.	Machine Learning based	Assembled metagenome contigs	Taxonomic label for each contig	Based on assembled contigs longer than 1 kb.
PrDOS [182]	Prediction of natively disordered regions of a protein chain from its amino acid sequence.	Machine Learning	Amino acid sequence	Two-state prediction (order/disorder) and a disorder probability for each residue	The number of sequences in the multiple FASTA-formatted input is limited to 50
SemiBin/SemiBin2 [183]	Metagenomic binning	Deep Learning	Contigs and BAM files consisting of reads mapped to the contigs	Bins	Not optimized for eukaryotic data
ThermoNet [184]	Quantitative prediction of the changes in protein thermodynamic stability caused by single-point amino acid substitutions (3D-CNN)	Deep Learning	Protein 3D structure (either experimental or computational)	Predicted $\Delta\Delta G$ caused by the given mutation compared to the wild-type protein structure	Requires reliable protein structures and is computationally intensive
ThermoProt [185]	Python package to predict the thermostability of proteins (SVM)	Deep Learning	Protein sequences	Prediction of <i>T_m</i>	Trained on specific IMG-derived sequences
Tiara [186]	Classifies eukaryotic sequences from metagenomic assemblies using deep learning	Deep Learning	Nucleotide sequences	Microbial & organelle taxonomy	Classification accuracy decreases for short sequences (<3000 bp)
TolRad [187]	Random Forest Binary Classifier that uses Pfam domains in bacterial genomes to differentiate radiation-sensitive from radiation-tolerant species	Machine Learning	Nucleotide & Protein sequences	Predicts tolerance (<i>D</i> ₁₀ > 200 Gy) or radiosensitivity (<i>D</i> ₁₀ < 200Gy)	Accuracy is limited by a small and taxonomically biased training dataset
tRNAscan-SE 2.0 [188]	tRNA prediction	Machine Learning	Nucleotide sequences (DNA or RNA)	Prediction of tRNA genes and their features and predicted structure	Accuracy declines for highly atypical or lineage-specific tRNAs and for distinguishing functional genes from pseudogenes without additional contextual data
VAMB/AAMB [189]	Variational autoencoder–based tool for contig binning; uses deep learning	Deep Learning	Nucleotide sequences (contigs)	Contig bins (and optionally taxonomy)	May struggle with low-abundance or highly similar genomes

(continued)

Table 2 Continued.

Tool	Description/use	Type	Input data type	Primary output/prediction	Limitations
ViraMiner [190]	Classifies viral contigs using CNN and Long Short-Term Memory layers.	Deep Learning	Metagenomic contigs	Output a value between 0 and 1 indicating the probability of the sequence coming from a viral origin	Performance may decline for highly divergent or novel viruses
Whokaryote [191]	Separates eukaryotic and prokaryotic sequences using sequence signatures	Machine Learning	Metagenomic contig or genomic sequence.	Contigs headers that have been classified as Prokaryotic or eukaryotic in separate files	Contigs with <2 genes or shorter than 5000 bp are not classified
XenoBug [192]	Application of machine learning-based methods to identify novel bacterial enzymes capable of degrading a wide range of xenobiotics	Machine Learning	PubChem ID or an sdf file of the query molecule	Predicted enzymes that can degrade the input molecule, their metagenome source, and pathway information, as well as the Tanimoto threshold at which EC numbers were predicted	Class imbalance among EC enzyme categories restricts training data representation

well as features such as thermostability [184] (e.g. ThermoNet) and solubility [196] (e.g. DeepSol) from sequence or structural inputs.

Another significant contribution of ML lies in integrating multi-omics data. Integrative ML frameworks can learn joint embeddings that map heterogeneous data types into shared latent spaces, uncovering hidden correlations between molecular layers and environmental gradients. A key idea is the combined use of genomic, transcriptomic, proteomic, and metabolomic measurements to construct a virtual cell. This unified computational representation captures cellular state and function as closely as possible [197].

A significant recent development is the replacement of task-specific tools and pipelines by large foundation models. Foundation models are large-scale neural networks trained on vast and highly diverse datasets with multiple objectives to learn generalizable representations that can be adapted to a wide range of downstream tasks with minimal fine-tuning [198]. Genomic foundation models (gFMs) have emerged as a new paradigm for a wide range of applications. Models such as DNABERT-2 [199], Nucleotide Transformer [200], and HyenaDNA [201] have demonstrated broad utility across biological domains, including gene expression and splicing regulation, variant-effect prediction, promoter and enhancer classification, and non-coding variant prioritization. The recently published AlphaGenome model [202] predicts thousands of functional genomic tracks for up to 1 Mb DNA sequence windows and reports strong performance on variant-effect benchmarks. The obtained embeddings provide a versatile foundation for downstream tasks, often outperforming task-specific models in low-data settings. Moreover, pretraining on a large number of genome sequences ensures that models can generalize beyond what they have seen during training.

Recent advances in explainable AI (XAI) further strengthen the synergy between computational modeling and experimental validation [203, 204]. By interpreting the internal structure of models, identifying which residues, motifs, or pathways contribute most to predictions, researchers can generate testable hypotheses and guide targeted experiments [205]. This interpretability is particularly critical in biological systems, where mechanistic transparency determines scientific credibility.

Similar to most other fields in life sciences, metagenomics has seen rapid adaptation of deep learning techniques [206], enabling automated binning, annotation, and element classification across complex microbial communities. VAMB [189] uses a variational autoencoder framework to integrate sequence composition and co-abundance information to perform highly accurate metagenomic binning, recovering 29%–98% more near-complete genomes compared to baseline methods, and even from low-abundance species. Complementing such binning approaches, CheckM2 [207] employs deep neural networks to estimate genome completeness and contamination, improving quality assessment over heuristic marker-based metrics. Beyond assembly, several ML tools address functional annotation and element identification. DeepEC [208] predicts enzyme commission (EC) numbers directly from protein sequences using convolutional networks. On a set of held-out sequences, DeepEC achieved the highest precision/recall (0.920/0.455) compared to five baseline methods. DeepARG [162] classifies antibiotic resistance genes from environmental metagenomes with reported performance exceeding 97% precision and 90% recall across 30 resistance categories. Similarly, DeepBCG [163] identifies biosynthetic gene clusters in bacterial and fungal genomes, accelerating the discovery of secondary metabolites. For mobile genetic elements, PlasFlow [209] uses neural-network classifiers to distinguish plasmid DNA from chromosomal DNA with reported performance up to 96% classification accuracy, 85.98% recall, and 72.17% precision, and DeepVirFinder [210] detects viral sequences directly from assembled contigs, achieving Area Under the Receiver Operating Characteristic Curve (AUROC) scores between 0.93 and 0.98 depending on contig length. The recently proposed ViraLM [211] uses the DNABERT-2 foundation model architecture to improve the identification of viral sequences with low similarity to known viruses and performs robustly on fragmented metagenomic assemblies, which have been challenging for prior methods. In benchmarking, ViraLM achieved precision comparable to protein-based baseline tools while maintaining higher recall across host groups. Notably, it attains F1-scores of 0.91 and 0.92 on distinguishing viruses from bat- and human-origin contigs, respectively, while the evolutionarily closest annotated sequences in the training dataset are from insects.

A complementary trend is the move toward reference- and assembly-free inference directly from metagenomic reads. For example, the REMME DNA language model and its fine-tuned version REBEAN predict enzyme-class annotations for short sequencing reads [212], consistently identifying three- to six-fold more enzymatic reads compared to alignment-based tools. This is especially valuable when extreme-environment samples yield low biomass, fragmented DNA, or highly fragmented assemblies.

Together, these models exemplify how deep learning architectures can capture nonlinear patterns in high-dimensional genomic data, yielding more accurate and scalable predictions than traditional rule-based or alignment-dependent pipelines.

While the methods above illustrate the power of ML, their application to extremophile datasets raises reliability concerns because available training data are sampling-, cultivation-, and annotation-biased and can differ markedly from the target domain. Two broad failure modes are therefore important to consider: Large Language Model (LLM)/agentic tools used for literature synthesis or automated reporting can generate unsupported statements (“hallucinations”) unless outputs are grounded in traceable sources, whereas sequence/structure foundation models more often fail through overconfident extrapolation and bias amplification under domain shift, producing plausible, high-confidence predictions that reflect training-set correlations rather than true mechanistic determinants. To promote valid and reproducible use, we recommend: (i) careful dataset preparation including metadata and train–test split criteria, as well as bias documentation, (ii) domain-shift-aware training and evaluation (e.g. leave-one-clade/biome-out testing and leakage controls), (iii) uncertainty calibration with abstention and triangulation across complementary methods, and (iv) experimental validation. Accordingly, AI/ML outputs should be treated as hypothesis-generating and require orthogonal validation before guiding further decisions.

Machine learning in extremophile research

Predicting protein and enzyme adaptations

Adaptive traits often arise through alterations of existing proteins rather than the evolution of entirely novel sequences and functions. Comparative analyses have shown that extremophilic proteins maintain canonical folding patterns while incorporating stabilizing features. Alterations in amino acid composition, increased surface charge density, strengthened hydrophobic cores, or additional salt bridges enhance resilience under harsh conditions. These principles are well illustrated in studies of thermophilic and halophilic enzymes, such as α -amylases and proteases, where subtle sequence variations confer substantial improvements in thermostability and solvent tolerance [213–215].

ML methods now offer systematic ways to generalize and predict such adaptive molecular properties directly from sequence or structure. Models trained on large protein datasets have achieved accurate predictions of thermostability [216], halotolerance [217], and acid stability [218], demonstrating ML’s capacity to learn nonlinear relationships between residue composition and environmental performance. One recent example, the thermostability prediction approach TemBERTure [219], emphasizes the careful construction of the training dataset with respect to sequence diversity, which is particularly relevant when transferring to underrepresented extremophile lineages. Benchmarking of TemBERTure yielded a comparable F1 score of 0.9 for the prediction of protein thermal class.

Furthermore, pretrained protein language models such as the ESM family or ProtT5 can be used in a transfer learning setting by fine-tuning on extremophile datasets to capture adaptation-relevant sequence embeddings, indicating which sequence patterns or global features correlate with environmental adaptation, thereby helping to identify, e.g. molecular determinants of stability without requiring explicit structural data. This approach can help overcome limited data availability for extremophilic organisms by leveraging the extensive general protein knowledge in foundation models (Fig. 4).

Bioprospecting novel functions and pathways

In addition to adaptations in existing proteins, some extreme environments required the innovation of new protein functions. The systematic exploration of such biomolecules and biocatalysts, known as bioprospecting, has become a central task in extremophile research. Both function-based and sequence-based metagenomic strategies have enabled the discovery of novel genes and enzymes directly from environmental samples, thereby overcoming the cultivation bottlenecks inherent to many extremophilic microorganisms [220]. gFMs have already been adapted in recent studies to model metagenomic sequences to provide informative DNA embeddings for segregating species and discovering multigenic biosynthetic gene clusters [221], as well as to improve the identification of coding regions from fragmented and mixed DNA sequences [222]. To link enzymatic function to ecological context, another recent report, XenoBug [192] uses random forests to predict the EC numbers of bacterial enzymes capable of biodegrading environmental contaminants. However, a significant limitation persists: a substantial fraction of recovered sequences shows no detectable homology to known proteins, rendering standard homology-based function-prediction approaches inapplicable. This has motivated method development beyond general property prediction, using deep learning architectures to classify enzyme function and predict catalytic residues. Tools such as DeepEC [208], DeepEC-Transformer [164], and GraphEC [172] combine convolutional, transformer, and GNN approaches to predict EC numbers and active sites from sequence and structural features. Importantly, these methods often generalize well even when trained predominantly on nonextremophilic data, benefiting from the broad representational power of deep learning models. Looking ahead, multimodal protein foundation models fine-tuned on extremophile datasets may further accelerate exploration of this functional “dark matter” (Fig. 4).

Modeling ecological and metabolic interactions

Extremophilic organisms often form complex, interdependent microbial communities whose collective metabolism enables survival under otherwise prohibitive conditions. Understanding the ecological and metabolic interactions can therefore be essential for understanding how energy, nutrients, and stress-response mechanisms are linked among community members. A previous study [223] employed machine learning to predict syntrophic and metabolic interdependencies among thermophilic consortia, revealing cooperative nutrient exchange and redox coupling strategies that sustain community function at high temperatures.

More generally, network-based modeling approaches could be used to model microbial community structures and functional organization across multiple scales [224]. GNNs [225] can represent species, genes, or metabolites as nodes connected by learned interaction edges, enabling the prediction of emergent community properties such as

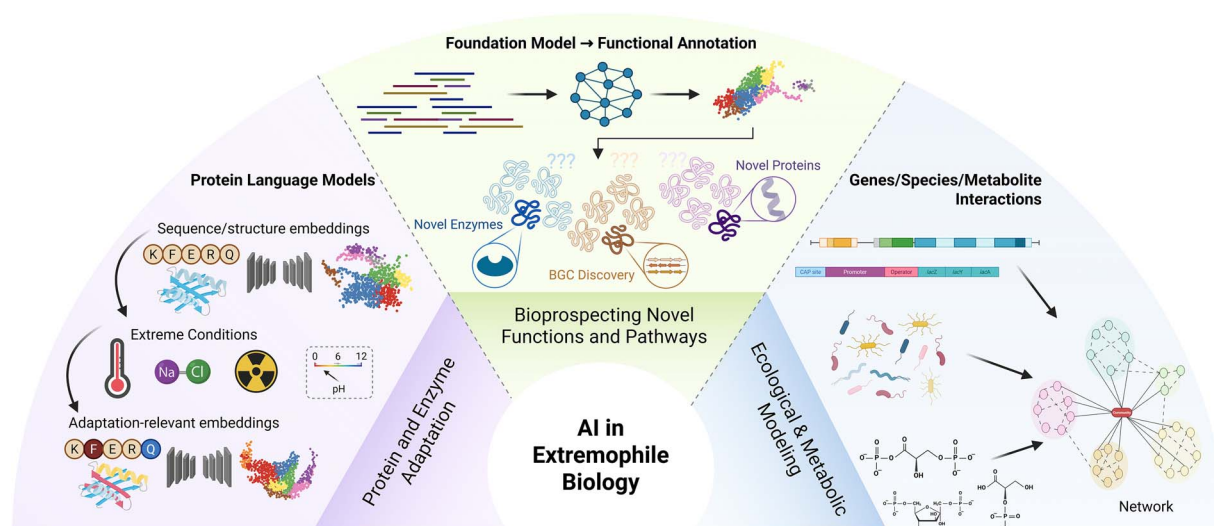


Figure 4 ML in extremophile biology. Protein language models capture sequence and structure features relevant to adaptation under extreme conditions, enabling the prediction of thermostability, halotolerance, and other traits. Foundation models support functional annotation of poorly characterized extremophile genomes, including the discovery of novel enzymes and biosynthetic gene clusters, and integrative AI frameworks model gene-species-metabolite interactions, revealing ecological organization and metabolic exchanges in extreme environments.

stability and resilience. Such integrative models offer a data-driven path to understanding how extremophile ecosystems maintain homeostasis (Fig. 4).

Discussion

Through long-term adaptation, extremophiles have evolved resistance mechanisms that allow them to thrive in some of the most inhospitable environments on Earth, including high temperatures, acidity, salinity, pressure, and radiation. These microorganisms have not only enhanced our understanding of microbial biology and the limits of life but also represent a significant untapped source of metabolic and enzymatic diversity. The earliest discoveries in the field of extremophiles led to the identification of essential biomolecules, such as ‘*Taq* polymerase’, enabling the development of revolutionary technologies. The comprehensive characterization of extremophiles will reveal an unexplored pool of biomolecules with promising biotechnological potential. Several studies have already started exploiting the unique characteristics of the extremophiles.

One of the most widely used biotechnological applications regarding enzymes derived from extremophiles is the development of VENUCEANE™. This active ingredient is used in the cosmetics industry, in creams, serums, and gels for the prevention and treatment of infrared-related skin damage. It originates from the extremophilic bacterium *T. thermophilus* strain GY1211. This thermophilic microorganism was first discovered in deep-sea hydrothermal vents and was cultivated under fermentation conditions engineered to yield a thermostable ferment extract rich in enzymatic antioxidants and extremozymes. The resulting ingredient retains activity at elevated temperatures and under solar exposure, making it highly effective against infrared-A- and ultraviolet-induced skin damage [226].

In the bioremediation field, a case of extremophile-based innovation is the Multi-strain Mixed Microbial Application Process (MMMAP). MMMAP is a microbial consortium comprising hyperthermophilic, barophilic, acidogenic, anaerobic bacteria (notably *Thermoanaerobacterium* sp., *Thermotoga* sp., and *Thermococcus* sp.) specialized in enhanced oil recovery. In this innovation, researchers collected

anaerobic formation fluid samples from high-temperature oil wells (70°C–90°C) in Mehsana, Gujarat, India, sampling under strict anoxic conditions. This consortium produces a combination of biosurfactants, organic acids, gases, and solvents that work synergistically to reduce oil viscosity, increase reservoir permeability, and improve hydrocarbon mobilization under extreme heat and pressure [227].

Another significant innovation in extremophile-based bioremediation biotechnology is demonstrated in a study evaluating diesel-contaminated soil near the Brazilian Antarctic Station. Indigenous, cold-adapted hydrocarbon-degrading bacteria, including species of *Pseudomonas*, *Sphingomonas*, and *Acinetobacter*, were isolated and used to create a bioaugmentation consortium tailored to Antarctic conditions. When applied *in situ*, these microbial consortia, along with biostimulation treatments, significantly altered community structure and enhanced degradation of total petroleum hydrocarbons, even at near-freezing temperatures. Bioaugmentation proved especially effective in soils with high diesel loads, highlighting the potential of native psychrotolerant strains for polar bioremediation efforts [228].

Despite growing significance, extremophile research remains in its infancy, with several ecosystems and microbial taxa left unexplored due to technical, logistical, and cultivation challenges. AI enables transforming this new field from an observational science into a predictive and generative framework. The synergy between foundation models, multi-omics datasets, and AI-guided experimentation can reveal adaptation principles innate to extremophiles, serving as natural experiments in robustness and adaptability. Key goals include AI-assisted bioprospecting and strain engineering, improving genotype–phenotype mapping in recombinant systems, and, ultimately, anticipating how overexpressed genes or modules perturb cellular state [229, 230]. Insights into metabolic interactions in extreme habitats may inform the design of synthetic consortia optimized for industrial processes under harsh conditions. Beyond these applications, AI could enable the systematic exploitation of extremophiles for bioremediation and terrestrial analogue research. ML-guided screening of metagenomic and proteomic data could identify stress-tolerant enzymes and

pathways involved in contaminant transformation, including heavy-metal resistance and degradation of recalcitrant compounds under extreme physicochemical conditions. At the ecosystem level, AI-driven models may optimize the composition and functional balance of extremophile-based microbial consortia, predicting collective behavior and metabolic interactions in harsh environments. Similar approaches could also be applied to terrestrial analogue studies. In such contexts, AI-assisted environmental and metabolic modeling of extremophile communities inhabiting hypersaline, acidic, or radiation-rich ecosystems (conditions that resemble key geothermal and physical features of Mars) could generate insights relevant both to contaminated environments on Earth and to Mars-analog systems.

Nevertheless, the application of AI and machine-learning methods to extremophile research is not without challenges. Available training data are overwhelmingly dominated by mesophilic organisms, and this may impede the proper modeling of distribution shifts associated with important biological differences in extremophiles, e.g. unique physicochemical signatures. To mitigate these limitations, possible strategies include leveraging transfer learning and fine-tuning to adapt representations learned from mesophilic data, while domain adaptation and physics-informed modeling can explicitly incorporate extremophile-specific constraints, such as enhanced structural rigidity in hyperthermophiles or surface charge enrichment in halophiles, thereby improving robustness beyond purely evolutionary similarity.

Finally, using extremophiles as a stress test for model generalization is attractive, but requires carefully designed benchmarks. These should include edge cases such as short or fragmented contigs and rare taxa, standardized environmental covariates, and held-out sites or time points to detect overfitting. Research community priorities should include curated and versioned benchmark suites with quality thresholds for metagenome assemblies and annotated proteins, prospective challenge tasks (e.g. enzyme property or function prediction), and interoperable repositories linking sequence, structure predictions, and experimental outcomes. These efforts will accelerate the impact of AI on extremophile research while ensuring robust, reproducible advances. Conclusively, combining multi-omic approaches with AI to study extremophiles will help us expand the catalogue of life and understand novel biochemical processes, ultimately fueling biotechnological innovation.

ecology, adaptation, and biotechnological applications, including industrial bioprocessing and bioremediation.

Funding

G.A.P. 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 (Grant ID: 23592-EMISSION). F.A.B. was supported by the 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); M.C. was awarded a fellowship through the KAUST Visiting Student Research Program (VSRP). King Abdullah University of Science and Technology also provided institutional support for this work through the KAUST Baseline grant BAS/1/1096-01-01, awarded to A.S.R.

Data availability

This study does not produce or analyze new data.

References

1. Brock TD, Freeze H. *Thermus aquaticus* gen. n. and sp. n., a Nonsporulating Extreme Thermophile. *Journal of Bacteriology* 1969;**98**:289–97. <https://doi.org/10.1128/jb.98.1.289-297.1969>
2. Rothschild LJ, Mancinelli RL. Life in extreme environments. *Nature* 2001;**409**:1092–101. <https://www.nature.com/articles/35059215>
3. Macelroy RD. Some comments on the evolution of extremophiles. *Biosystems*. 1974;**6**:74–5. [https://doi.org/10.1016/0303-2647\(74\)90026-4](https://doi.org/10.1016/0303-2647(74)90026-4)
4. Pedersen K. Exploration of deep intraterrestrial microbial life: current perspectives. *FEMS Microbiol Lett* 2000;**185**:9–16. <https://doi.org/10.1111/j.1574-6968.2000.tb09033.x>
5. Ruff SE, de Angelis IH, Mullis M *et al.* A global comparison of surface and subsurface microbiomes reveals large-scale biodiversity gradients, and a marine-terrestrial divide. *Science Advances* 2024;**10**:eadq0645. <https://www.science.org/doi/10.1126/sciadv.adq0645>
6. Battistuzzi FU, Feijao A, Hedges SB. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evolutionary Biology* 2004;**4**:44. <https://link.springer.com/article/10.1186/1471-2148-4-44>
7. Merino N, Aronson HS, Bojanova DP *et al.* Living at the Extremes: Extremophiles and the Limits of Life in a Planetary Context. *Frontiers in Microbiology* 2019;**10**:780. <https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2019.00780/full>. <https://doi.org/10.3389/fmicb.2019.00780>
8. Capece MC, Clark E, Saleh JK *et al.* Polyextremophiles and the Constraints for Terrestrial Habitability. In: Seckbach J, Oren A, Stan-Lotter H (eds.), *Polyextremophiles: Life Under Multiple Forms of Stress*. Dordrecht: Springer Netherlands; 2013, 3–59. https://link.springer.com/10.1007/978-94-007-6488-0_1
9. Rawat M, Chauhan M, Pandey A. Extremophiles and their expanding biotechnological applications. *Arch Microbiol* 2024;**206**:247. <https://doi.org/10.1007/s00203-024-03981-x>

Key Points

- Extremophiles exhibit unique physiological and molecular adaptations that enable survival under extreme temperature, pH, salinity, radiation, and pressure conditions.
- Advances in multi-omics technologies, metagenomics, transcriptomics, proteomics, and metabolomics have transformed extremophile research, enabling detailed characterization of uncultured and rare taxa.
- Bioinformatics resources and specialized databases are essential for interpreting extremophile genomes, identifying extremozymes, and exploring novel metabolic pathways.
- Machine learning approaches are increasingly central, supporting functional annotation, prediction of adaptive protein traits, metagenomic binning, and discovery of novel enzymes and biosynthetic gene clusters.
- Integrating artificial intelligence with multi-omics approaches provides a predictive framework for understanding extremophile

10. Schröder C, Burkhardt C, Antranikian G. What we learn from extremophiles. *ChemTexts* 2020;**6**:8. <https://doi.org/10.1007/s40828-020-0103-6>
11. Harrison JP, Gheeraert N, Tsigelnitskiy D *et al.* The limits for life under multiple extremes. *Trends Microbiol* 2013;**21**:204–12. <https://doi.org/10.1016/j.tim.2013.01.006>
12. Billi D. Desert cyanobacteria under space and planetary simulations: a tool for searching for life beyond earth and supporting human space exploration. *Int J Astrobiol* 2019;**18**:483–9. <https://doi.org/10.1017/S147355041800037X>
13. Cowan DA, Albers SV, Antranikian G *et al.* Extremophiles in a changing world. *Extremophiles*. 2024;**28**:26. <https://doi.org/10.1007/s00792-024-01341-7>
14. Monsieurs P, Mijndonckx K, Provoost A *et al.* Draft genome sequences of *Ralstonia pickettii* strains SSH4 and CW2, isolated from space equipment. *Genome Announc* 2014;**2**:e00887–14. <https://doi.org/10.1128/genomeA.00887-14>
15. Checinska A, Probst AJ, Vaishampayan P *et al.* Microbiomes of the dust particles collected from the international Space Station and spacecraft assembly facilities. *Microbiome*. 2015;**3**:50. <https://doi.org/10.1186/s40168-015-0116-3>
16. Minich JJ, Zhu Q, Janssen S *et al.* KatharoSeq enables high-throughput microbiome analysis from low-biomass samples. *mSystems*. 2018;**3**:e00218–7. <https://doi.org/10.1128/mSystems.00218-17>
17. Singh NK, Wood JM, Karouia F *et al.* Correction to: succession and persistence of microbial communities and antimicrobial resistance genes associated with international Space Station environmental surfaces. *Microbiome*. 2018;**6**:214. <https://doi.org/10.1186/s40168-018-0609-y>
18. La Duc MT, Venkateswaran K, Conley CA. A genetic inventory of spacecraft and associated surfaces. *Astrobiology*. 2014;**14**:15–23. <https://doi.org/10.1089/ast.2013.0966>
19. Sivabalan SKM, Vijayakumar V, Sengupta P *et al.* Unveiling Hidden Microbial Diversity in Mars 2020 Mission Assembly Cleanrooms with Molecular Insights into the Persistent and Perseverance of Novel Species Defying Metagenome Sequencing. 2025:2025.01.21.633868. <https://www.biorxiv.org/content/10.1101/2025.01.21.633868v1>. <https://doi.org/10.1101/2025.01.21.633868>
20. Fukuchi S, Yoshimune K, Wakayama M *et al.* Unique amino acid composition of proteins in halophilic bacteria. *J Mol Biol* 2003;**327**:347–57. [https://doi.org/10.1016/S0022-2836\(03\)00150-5](https://doi.org/10.1016/S0022-2836(03)00150-5)
21. Kohli I, Joshi NC, Mohapatra S *et al.* Extremophile – an adaptive strategy for extreme conditions and applications. *Curr Genomics* 2020;**21**:96–110. <https://doi.org/10.2174/1389202921666200401105908>
22. Brininger C, Spradlin S, Cobani L *et al.* The more adaptive to change, the more likely you are to survive: Protein adaptation in extremophiles. *Seminars in Cell & Developmental Biology* 2018;**84**:158–69. (SI: Antigen presentation). <https://doi.org/10.1016/j.semcdb.2017.12.016>
23. Coker JA. ‘All about’ extremophiles. *Fac Rev* 2023;**12**:27. <https://doi.org/10.12703/r/12-27>
24. Walker DJF, Nevin KP, Holmes DE *et al.* Syntrophus conductive pili demonstrate that common hydrogen-donating syntrophs can have a direct electron transfer option. *The ISME Journal* 2020;**14**:837–46. <https://doi.org/10.1038/s41396-019-0575-9>
25. Valentin MT, Luo G, Zhang S *et al.* Direct interspecies electron transfer mechanisms of a biochar-amended anaerobic digestion: a review. *Biotechnol Biofuels Bioprod* 2023;**16**:146. <https://doi.org/10.1186/s13068-023-02391-3>
26. Gralnick JA, Bond DR. Electron transfer beyond the outer membrane: putting electrons to rest. *Ann Rev Microbiol* 2023;**77**:517–39. <https://doi.org/10.1146/annurev-micro-032221-023725>
27. Wegener G, Krukenberg V, Riedel D *et al.* Intercellular wiring enables electron transfer between methanotrophic archaea and bacteria. *Nature* 2015;**526**:587–90. <https://doi.org/10.1038/nature15733>
28. Krüger A, Schäfers C, Schröder C *et al.* Towards a sustainable biobased industry – Highlighting the impact of extremophiles. *New Biotechnology* 2018;**40**:144–53. (Bioeconomy). <https://doi.org/10.1016/j.nbt.2017.05.002>
29. Extremophilic microbes for environmental bioremediation: a review. *Biointerface Research in Applied Chemistry* 2024;**14**:103. <https://doi.org/10.33263/BRIAC145.103>
30. Coker JA. Extremophiles and biotechnology: current uses and prospects. *F1000Research* 2016;**5**:F1000 Faculty Rev-396. <https://doi.org/10.12688/f1000research.7432.1>
31. Uppar AP, Vanti GL, Poondla N *et al.* Applications of Extremophiles in Therapeutics. In: Shah MP, Dey S (eds.), *Trends in Biotechnology of Polyextremophiles*. Cham: Springer Nature Switzerland; 2024, 231–42. https://doi.org/10.1007/978-3-031-55032-4_10
32. Hui ML-Y, Tan LT-H, Letchumanan V *et al.* The extremophilic actinobacteria: from microbes to medicine. *Antibiotics*. 2021; **10**(6):682. <https://www.mdpi.com/2079-6382/10/6/682>. <https://doi.org/10.3390/antibiotics10060682>
33. Damer B, Deamer D. The Hot Spring Hypothesis for an Origin of Life. *Astrobiology* 2020;**20**:429–52. https://journals.sagepub.com/.https://doi.org/10.1089/ast.2019.2045?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed
34. Pikuta EV, Hoover RB, Tang J. Microbial Extremophiles at the Limits of Life. *Critical Reviews in Microbiology* 2007;**33**:183–209. https://www.tandfonline.com/.https://doi.org/10.1080/10408410701451948?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed
35. Islas S, Velasco AM, Becerra A *et al.* Hyperthermophily and the origin and earliest evolution of life. *Int Microbiol* 2003;**6**:87–94. <https://link.springer.com/article/.https://doi.org/10.1007/s10123-003-0113-4>
36. Schultz J, dos Santos A, Patel N *et al.* Life on the edge: bioprospecting extremophiles for astrobiology. *Journal of the Indian Institute of Science* 2023;**103**:721–37. <https://doi.org/10.1007/s41745-023-00382-9>
37. Carré L, Zaccari G, Delfosse X *et al.* Relevance of Earth-Bound Extremophiles in the Search for Extraterrestrial Life. *Astrobiology* 2022;**22**:322–67. https://journals.sagepub.com/.https://doi.org/10.1089/ast.2021.0033?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed
38. McClain CR, Bryant SR, Hanks G *et al.* Extremophiles in Earth’s Deep Seas: A View Toward Life in Exo-Oceans. *Astrobiology* 2022;**22**:1009–28. https://journals.sagepub.com/.https://doi.org/10.1089/ast.2021.0120?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed
39. Lopez JV, Peixoto RS, Rosado AS. Inevitable future: space colonization beyond earth with microbes first. *FEMS Microbiol Ecol* 2019;**95**:fiz127. <https://doi.org/10.1093/femsec/fiz127>
40. Yang Z-W, Lian Z-H, Liu L *et al.* Cultivation strategies for prokaryotes from extreme environments. *iMeta* 2023;**2**:e123. <https://onlinelibrary.wiley.com/.https://doi.org/10.1002/imt2.123>
41. Rock Surface Fungi in Deep Continental Biosphere—Exploration of Microbial Community Formation with Subsurface In Situ Biofilm Trap | MDPI. [accessed 2026 Feb 18]. <https://www.mdpi.com/2076-2607/9/1/64>

42. Frontiers | Shedding light on the composition of extreme microbial dark matter: alternative approaches for culturing extremophiles. [accessed 2026 Feb 18]. <https://www.frontiersin.org/journals/microbiology/articles/https://doi.org/10.3389/fmicb.2023.1167718/full>
43. Kapinusova G, Lopez Marin MA, Uhlik O. Reaching unreachables: Obstacles and successes of microbial cultivation and their reasons. *Front Microbiol* 2023;14. <https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2023.1089630/full>. <https://doi.org/10.3389/fmicb.2023.1089630>
44. Stewart EJ. Growing Unculturable Bacteria. *Journal of Bacteriology* 2012;194:4151–60. <https://journals.asm.org/>. https://doi.org/10.1128/jb.00345-12?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossr ef.org&rfr_dat=cr_pub++0pubmed
45. Lewis WH, Tahon G, Geesink P *et al*. Innovations to culturing the uncultured microbial majority. *Nat Rev Microbiol* 2021;19:225–40. <https://www.nature.com/articles/s41579-020-00458-8>
46. Rinke C, Lee J, Nath N *et al*. Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat Protoc* 2014;9:1038–48. <https://doi.org/10.1038/nprot.2014.067>
47. Baret J-C, Miller OJ, Taly V *et al*. Fluorescence-activated droplet sorting (FADS): efficient microfluidic cell sorting based on enzymatic activity. *Lab Chip* 2009;9:1850–8. <https://doi.org/10.1039/B902504A>
48. Ambriz-Aviña V, Contreras-Garduño JA, Pedraza-Reyes M. Applications of Flow Cytometry to Characterize Bacterial Physiological Responses. *Biomed Res Int* 2014;2014:461941. <https://onlinelibrary.wiley.com/>. <https://doi.org/10.1155/2014/461941>
49. Stress response assessment of *Lactobacillus sakei* strains selected as potential autochthonous starter cultures by flow cytometry and nucleic acid double <https://www.zotero.org/google-docs/?EW8D9i-staining> analyses | Journal of Applied Microbiology | Oxford Academic. [accessed 2026 Feb 18]. <https://academic.oup.com/jambio/article-abstract/115/3/786/6716751?redirectedFrom=fulltext&login=false>
50. Olson ND, Morrow JB. DNA extract characterization process for microbial detection methods development and validation. *BMC Res Notes* 2012;5:668. <https://link.springer.com/article/>. <https://doi.org/10.1186/1756-0500-5-668>
51. Leuko S, Goh F, Ibáñez-Peral R *et al*. Lysis efficiency of standard DNA extraction methods for Halococcus spp. in an organic rich environment. *Extremophiles*. 2008;12:301–8. <https://doi.org/10.1007/s00792-007-0124-8>
52. Chauhan G, Arya M, Kumar V *et al*. An improved protocol for metagenomic DNA isolation from low microbial biomass alkaline hot-spring sediments and soil samples. *3 Biotech* 2024;14:34. <https://link.springer.com/article/>. <https://doi.org/10.1007/s13205-023-03824-4>
53. Improvements in Extraction Methods of High-molecular-weight DNA from Soils by Modifying Cell Lysis Conditions and Reducing Adsorption of DNA onto Soil Particles. [accessed 2026 Feb 18]. https://www.jstage.jst.go.jp/article/jsme2/36/3/36_ME21017/_article
54. Bürgmann H, Pesaro M, Widmer F *et al*. A strategy for optimizing quality and quantity of DNA extracted from soil. *J Microbiol Methods* 2001;45:7–20. [https://doi.org/10.1016/S0167-7012\(01\)00213-5](https://doi.org/10.1016/S0167-7012(01)00213-5)
55. Moreira D. Efficient removal of PCR inhibitors using agarose-embedded DNA preparations. *Nucleic Acids Research* 1998;26:3309–10. <https://academic.oup.com/nar/article/26/13/3309/2386765?login=false>
56. Tighe S, Afshinnekoo E, Rock TM *et al*. Genomic methods and microbiological Technologies for Profiling Novel and Extreme Environments for the extreme microbiome project (XMP). *Journal of Biomolecular Techniques : JBT* 2017;28:31–9. <https://doi.org/10.7171/jbt.17-2801-004>
57. Oren A. Molecular ecology of extremely halophilic archaea and bacteria. *FEMS Microbiol Ecol* 2002;39:1–7. <https://doi.org/10.1111/j.1574-6941.2002.tb00900.x>
58. Gabor EM, de Vries EJ, Janssen DB. Efficient recovery of environmental DNA for expression cloning by indirect extraction methods. *FEMS Microbiol Ecol* 2003;44:153–63. <https://academic.oup.com/femsec/article/44/2/153/546907?login=false>
59. Morono Y, Terada T, Hoshino T *et al*. Hot-Alkaline DNA Extraction Method for Deep-Subseafloor Archaeal Communities. *Appl Environ Microbiol* 2014;80:1985–94. <https://journals.asm.org/>. <https://doi.org/10.1128/aem.04150-13>
60. Sankaranarayanan K, Timofeeff MN, Spathis R *et al*. Ancient Microbes from Halite Fluid Inclusions: Optimized Surface Sterilization and DNA Extraction. *PLOS ONE* 2011;6:e20683. <https://journals.plos.org/plosone/article?id=>. <https://doi.org/10.1371/journal.pone.0020683>
61. Minas K, McEwan NR, Newbold CJ *et al*. Optimization of a high-throughput CTAB-based protocol for the extraction of qPCR-grade DNA from rumen fluid, plant and bacterial pure cultures. *FEMS Microbiol Lett* 2011;325:162–69. <https://academic.oup.com/femsle/article-abstract/325/2/162/566576?redirectedFrom=fulltext&login=false>
62. 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>
63. Han J-R, Li S, Li W-J *et al*. Mining microbial and metabolic dark matter in extreme environments: a roadmap for harnessing the power of multi-omics data. *Advanced Biotechnology* 2024;2:26. <https://doi.org/10.1007/s44307-024-00034-8>
64. Fongaro G, Maia GA, Rogovski P *et al*. Extremophile microbial communities and enzymes for bioenergetic application based on multi-omics tools. *Curr Genomics* 2020;21:240–52. <https://doi.org/10.2174/1389202921999200601144137>
65. de la Haba RR, Antunes A, Hedlund BP. Editorial: Extremophiles: Microbial genomics and taxogenomics. *Front Microbiol* 2022;13. <https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2022.984632/full>. <https://doi.org/10.3389/fmicb.2022.984632>
66. Johnson JS, Spakowicz DJ, Hong B-Y *et al*. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun* 2019;10:5029. <https://www.nature.com/articles/s41467-019-13036-1>
67. Oulas A, Pavludi C, Polymenakou P *et al*. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinform Biol Insights* 2015;9:75–88. <https://doi.org/10.4137/BBI.S12462>
68. 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>
69. 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>

70. Uritskiy G, DiRuggiero J. Applying genome-resolved metagenomics to deconvolute the halophilic microbiome. *Genes*. 2019;**10**:220. <https://doi.org/10.3390/genes10030220>
71. Chen K, Pachter L. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol* 2005;**1**:106–12. <https://doi.org/10.1371/journal.pcbi.0010024>
72. Chasapi MN, Chasapi IN, Aplakidou E *et al.* metagRoot: a comprehensive database of protein families associated with plant root microbiomes. *Nucleic Acids Res* 2025;**54**:D1733–42. <https://doi.org/10.1093/nar/gkaf862>
73. Langille MGI, Zaneveld J, Caporaso JG *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* 2013;**31**:814–21. <https://www.nature.com/articles/nbt.2676>
74. Shakya M, Lo C-C, Chain PSG. Advances and challenges in Metatranscriptomic analysis. *Front Genet* 2019;**10**:904. <https://doi.org/10.3389/fgene.2019.00904>
75. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. <https://doi.org/10.1186/s13059-014-0550-8>
76. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;**26**:139–40. <https://doi.org/10.1093/bioinformatics/btp616>
77. Yan W, Wang D, Wang Y *et al.* Metatranscriptomics-guided genome-scale metabolic reconstruction reveals the carbon flux and trophic interaction in methanogenic communities. *Microbiome*. 2024;**12**:121. <https://doi.org/10.1186/s40168-024-01830-z>
78. Polinski JM, Rodrigue M, Meyer JD *et al.* Drifting in the deep: Metatranscriptomics and metabarcoding reveal sustained metabolic activity and community composition in hydrothermal vent plume microbial communities. *Frontiers in Marine Science* 2023;**10**.
79. Mirete S, Lamprecht-Grandío M, de Figueras CG *et al.* Domain-specific osmoadaptation revealed by metatranscriptomic analysis in hypersaline environments. *Sci Rep* 2025;**15**:23320. <https://doi.org/10.1038/s41598-025-04148-4>
80. Van Den Bossche T, Armengaud J, Benndorf D *et al.* *The microbiologist's guide to metaproteomics iMeta* 2025;**4**:e70031. <https://doi.org/10.1002/imt2.70031>
81. Thomas SC, Payne D, Tamadonfar KO *et al.* Genomics, Exometabolomics, and Metabolic Probing Reveal Conserved Proteolytic Metabolism of Thermoflexus hugenholtzii and Three Candidate Species From China and Japan. *Front Microbiol* 2021;**12**. <https://www.frontiersin.org/journals/microbiology/articles/https://doi.org/10.3389/fmicb.2021.632731>
82. Bilova T, Golushko N, Frolova N *et al.* Strain-specific features of primary metabolome characteristic for extremotolerant/extremophilic cyanobacteria under long-term storage. *Int J Mol Sci* 2025;**26**:2201. <https://doi.org/10.3390/ijms26052201>
83. Exploring Andean high-altitude Lake extremophiles through advanced Proteotyping |. *J Proteome Res* 2025 [accessed 2025 Nov 14];**23**:891–904. <https://pubs.acs.org/doi/https://doi.org/10.1021/acs.jproteome.3c00538>
84. Tschitschko B, Williams TJ, Allen MA *et al.* Ecophysiological distinctions of haloarchaea from a hypersaline Antarctic Lake as determined by Metaproteomics. *Appl Environ Microbiol* 2016 [accessed 2025 Nov 14];**82**:3165–73. <https://pubmed.ncbi.nlm.nih.gov/26994078/https://doi.org/10.1128/AEM.00473-16>
85. Meier DV, Pjevac P, Bach W *et al.* Microbial metal-sulfide oxidation in inactive hydrothermal vent chimneys suggested by metagenomic and metaproteomic analyses. *Environ Microbiol* 2019;**21**:682–701. <https://doi.org/10.1111/1462-2920.14514>
86. Pjevac P, Meier DV, Markert S *et al.* Metaproteogenomic Profiling of Microbial Communities Colonizing Actively Venting Hydrothermal Chimneys. *Frontiers in Microbiology* 2018;**9**. <https://www.frontiersin.org/journals/microbiology/articles/https://doi.org/10.3389/fmicb.2018.00680>
87. Mandelli F, Couger MB, Paixão D a. A *et al.* Thermal adaptation strategies of the extremophile bacterium *Thermus filiformis* based on multi-omics analysis. *Extremophiles* 2017;**21**:775–88. <https://link.springer.com/article/https://doi.org/10.1007/s00792-017-0942-2>
88. Power JF, Carere CR, Lee CK *et al.* Microbial biogeography of 925 geothermal springs in New Zealand. *Nat Commun* 2018;**9**:2876. <https://doi.org/10.1038/s41467-018-05020-y>
89. Neira G, Cortez D, Jil J *et al.* AcIDB 1.0: a database of acidophilic organisms, their genomic information and associated metadata. *Bioinformatics*. 2020;**36**:4970–1. <https://doi.org/10.1093/bioinformatics/btaa638>
90. Schober I, Koblit J, Sardà Carbasse J *et al.* BacDive in 2025: the core database for prokaryotic strain data. *Nucleic Acids Research* 2025;**53**:D748–56. <https://doi.org/10.1093/nar/gkaf959>
91. Pal C, Bengtsson-Palme J, Rensing C *et al.* BacMet: Antibacterial biocide and metal resistance genes database. *Nucleic Acids Res* 2014;**42**:D737–43. <https://doi.org/10.1093/nar/gkt1252>
92. Chang A, Jeske L, Ulbrich S *et al.* BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res* 2021;**49**:D498–508. <https://doi.org/10.1093/nar/gkaa1025>
93. Olson RD, Assaf R, Brettin T *et al.* Introducing the bacterial and viral bioinformatics resource Center (BV-BRC): a resource combining PATRIC. *IRD and ViPR Nucleic Acids Research* 2023;**51**:D678–89. <https://doi.org/10.1093/nar/gkac1003>
94. Waghv FH, Barai RS, Gurung P, Idicula-Thomas S. CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Research* 2016;**44**:D1094–7. <https://doi.org/10.1093/nar/gkv1051>
95. Alcock BP, Huynh W, Chalil R *et al.* CARD 2023: expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2023;**51**:D690–9. <https://doi.org/10.1093/nar/gkac920>
96. Drula E, Garron M-L, Dogan S *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res* 2022;**50**:D571–7. <https://doi.org/10.1093/nar/gkab1045>
97. Galperin MY, Vera Alvarez R, Karamycheva S *et al.* COG database update 2024. *Nucleic Acids Res* 2024;**53**:D356–63. <https://doi.org/10.1093/nar/gkaf983>
98. Cantalapiedra CP, Hernández-Plaza A, Letunic I *et al.* eggNOG-mapper v2: functional annotation, Orthology assignments, and domain prediction at the metagenomic scale Tamura K, editor. *Mol Biol Evol* 2021;**38**:5825–9. <https://doi.org/10.1093/molbev/msab293>
99. Aspromonte MC, Nugnes MV, Quaglia F *et al.* DisProt in 2024: improving function annotation of intrinsically disordered proteins. *Nucleic Acids Research* 2024;**52**:D434–41. <https://doi.org/10.1093/nar/gkad928>
100. Harrison PW, Amode MR, Austine-Orimoloye O *et al.* Ensembl 2024. *Nucleic Acids Res* 2024;**52**:D891–9. <https://doi.org/10.1093/nar/gkad1049>
101. Burgin J, Ahamed A, Cummins C *et al.* The European nucleotide archive in 2022. *Nucleic Acids Res* 2023;**51**:D121–5. <https://doi.org/10.1093/nar/gkac1051>

102. Chalita M, Kim YO, Park S *et al.* EzBioCloud: a genome-driven database and platform for microbiome identification and discovery. *Int J Syst Evol Microbiol* 2024;**74**:006421. <https://doi.org/10.1099/ijsem.0.006421>
103. The Gene Ontology Consortium, Aleksander SA, Balhoff J *et al.* The gene ontology knowledgebase in 2023. *Genetics*. 2023;**224**:iyad031. <https://doi.org/10.1093/genetics/iyad031>
104. Speth DR, Pullen N, Aroney STN *et al.* GlobDB: A comprehensive species-dereplicated microbial genome resource. 2025. <http://arxiv.org/abs/2506.11896>. <https://doi.org/10.48550/arXiv.2506.11896>
105. Chaumeil P-A, Mussig AJ, Hugenholtz P *et al.* GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*. 2022;**38**:5315–6. <https://doi.org/10.1093/bioinformatics/btac672>
106. Loukas A, Kappas I, Abatzopoulos TJ. HaloDom: a new database of halophiles across all life domains. *J Biol Res* 2018;**25**:2. <https://doi.org/10.1186/s40709-017-0072-0>
107. DasSarma SL, Capes MD, DasSarma P *et al.* HaloWeb: the haloarchaeal genomes database. *Saline Systems* 2010;**6**:12. <https://doi.org/10.1186/1746-1448-6-12>
108. Wang M, Liu G, Liu M *et al.* ICEberg 3.0: functional categorization and analysis of the integrative and conjugative elements in bacteria. *Nucleic Acids Research* 2024;**52**:D732–7. <https://doi.org/10.1093/nar/gkad935>
109. Camargo AP, Nayfach S, Chen I-MA *et al.* IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res* 2022;**51**:D733–43. <https://doi.org/10.1093/nar/gkac1037>
110. Fiamenghi MB, Camargo AP, Chasapi IN *et al.* Meta-virus resource (MetaVR): expanding the frontiers of viral diversity with 24 million uncultivated virus genomes. *Nucleic Acids Res* 2026;**54**:D801–12. <https://doi.org/10.1093/nar/gkaf1283>
111. Sunny JS, Nisha K, Natarajan A *et al.* IND-enzymes: a repository for hydrolytic enzymes derived from thermophilic and psychrophilic bacterial species with potential industrial usage. *Extremophiles*. 2021;**25**:319–25. <https://doi.org/10.1007/s00792-021-01231-2>
112. Alam I, Antunes A, Kamau AA *et al.* INDIGO - INtegrated data warehouse of microbial genomes with examples from the red sea extremophiles. *PLoS One* 2013;**8**:e82210. <https://doi.org/10.1371/journal.pone.0082210>
113. 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>
114. Paysan-Lafosse T, Blum M, Chuguransky S *et al.* InterPro in 2022. *Nucleic Acids Res* 2023;**51**:D418–27. <https://doi.org/10.1093/nar/gkac993>
115. Siguier P, Perochon J, Lestrade L *et al.* ISfinder: the reference Centre for bacterial insertion sequences. *Nucleic Acids Res* 2006;**34**:D32–6. <https://doi.org/10.1093/nar/gkj014>
116. Mukherjee S, Stamatis D, Li CT *et al.* Twenty-five years of genomes OnLine database (GOLD): data updates and new features in v.9. *Nucleic Acids Res* 2023;**51**:D957–63. <https://doi.org/10.1093/nar/gkac974>
117. Arkin AP, Cottingham RW, Henry CS *et al.* KBase: the United States Department of Energy Systems Biology Knowledgebase. *Nat Biotechnol* 2018;**36**:566–9. <https://doi.org/10.1038/nbt.4163>
118. Kanehisa M, Sato Y. KEGG mapper for inferring cellular functions from protein sequences. *Protein Science: A Publication of the Protein Society* 2020;**29**:28–35. <https://doi.org/10.1002/pro.3711>
119. Alam I, Kamau AA, Ngugi DK *et al.* KAUST metagenomic analysis platform (KMAP), enabling access to massive analytics of re-annotated metagenomic data. *Sci Rep* 2021;**11**:11511. <https://doi.org/10.1038/s41598-021-90799-y>
120. Klemetsen T, Raknes IA, Fu J *et al.* The MAR databases: development and implementation of databases specific for marine metagenomics. *Nucleic Acids Res* 2018;**46**:D692–9. <https://doi.org/10.1093/nar/gkx1036>
121. Nata'ala MK, Avila Santos AP, Coelho Kasmanas J *et al.* MarineMetagenomeDB: a public repository for curated and standardized metadata for marine metagenomes. *Environ Microbiome* 2022;**17**:57. <https://doi.org/10.1186/s40793-022-00449-7>
122. Chivian D, Dehal PS, Keller K *et al.* MetaMicrobesOnline: phylogenomic analysis of microbial communities. *Nucleic Acids Res* 2013;**41**:D648–54. <https://doi.org/10.1093/nar/gks1202>
123. Richardson L, Allen B, Baldi G *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res* 2023;**51**:D753–9. <https://doi.org/10.1093/nar/gkac1080>
124. Zdouc MM, Blin K, Louwen NLL *et al.* MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Research* 2025;**53**:D678–90. <https://doi.org/10.1093/nar/gkae1115>
125. Matias Rodrigues JF, Tackmann J, Malfertheiner L *et al.* The MicrobeAtlas database: Global trends and insights into Earth's microbial ecosystems. *Cell* 2026;**189**:2092–2107.e17. <https://doi.org/10.1016/j.cell.2026.01.021>
126. Vallenet D, Calteau A, Cruveiller S *et al.* MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. *Nucleic Acids Res* 2017;**45**:D517–28. <https://doi.org/10.1093/nar/gkw1101>
127. Sayers EW, Cavanaugh M, Clark K *et al.* GenBank. *Nucleic Acids Res* 2022;**50**:D161–4. <https://doi.org/10.1093/nar/gkab1135>
128. Haft DH, Badretdin A, Coulouris G *et al.* RefSeq and the prokaryotic genome annotation pipeline in the age of metagenomes. *Nucleic Acids Res* 2024;**52**:D762–9. <https://doi.org/10.1093/nar/gkad988>
129. 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>
130. Buchholz PCF, Feuerriegel G, Zhang H *et al.* Plastics degradation by hydrolytic enzymes: the plastics-active enzymes database-PAZy. *Proteins*. 2022;**90**:1443–56. <https://doi.org/10.1002/prot.26325>
131. 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>
132. Lau WYV, Hoad GR, Jin V *et al.* PSORTdb 4.0: expanded and redesigned bacterial and archaeal protein subcellular localization database incorporating new secondary localizations. *Nucleic Acids Res* 2021;**49**:D803–8. <https://doi.org/10.1093/nar/gkaa1095>
133. Zeng C, Zhan W, Deng L. SDADB: a functional annotation database of protein structural domains. *Database*. 2018;**2018**:bay064. <https://doi.org/10.1093/database/bay064>
134. Overbeek R, Olson R, Pusch GD *et al.* The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res* 2014;**42**:D206–14. <https://doi.org/10.1093/nar/gkt1226>
135. Kodama Y, Shumway M, Leinonen R *et al.* On behalf of the international nucleotide sequence database collaboration. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res* 2012;**40**:D54–6. <https://doi.org/10.1093/nar/gkr854>

136. Chuvochina M, Gerken J, Frentrop M *et al.* SILVA in 2026: a global core biodata resource for rRNA within the DSMZ digital diversity. *Nucleic Acids Research* 2026;**54**:D334–41. <https://doi.org/10.1093/nar/gkaf1247>
137. Guan J, Chen Y, Goh Y-X *et al.* TADB 3.0: an updated database of bacterial toxin–antitoxin loci and associated mobile genetic elements. *Nucleic Acids Res* 2024;**52**:D784–90. <https://doi.org/10.1093/nar/gkad962>
138. Sato Y, Okano K, Kimura H *et al.* TEMPURA: database of growth TEMPeratures of usual and RAre prokaryotes. *Microbes Environ* 2020;**35**:ME20074. <https://doi.org/10.1264/jsme2.ME20074>
139. Corrêa FB, Saraiva JP, Stadler PF *et al.* TerrestrialMetagenomeDB: a public repository of curated and standardized metadata for terrestrial metagenomes. *Nucleic Acids Res* 2019;**48**:D626–32. <https://doi.org/10.1093/nar/gkz994>
140. DiGiacomo J, McKay C, Davila A. ThermoBase: a database of the phylogeny and physiology of thermophilic and hyperthermophilic organisms. *PLoS One* 2022;**17**:e0268253. <https://doi.org/10.1371/journal.pone.0268253>
141. Haft DH, Selengut JD, Richter RA *et al.* TIGRFAMs and genome properties in 2013. *Nucleic Acids Res* 2013;**41**:D387–95. <https://doi.org/10.1093/nar/gks1234>
142. Saier MH, Reddy VS, Moreno-Hagelsieb G *et al.* The transporter classification database (TCDB): 2021 update. *Nucleic Acids Res* 2021;**49**:D461–7. <https://doi.org/10.1093/nar/gkaa1004>
143. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Res* 2025;**53**:D609–17. <https://doi.org/10.1093/nar/gkae1010>
144. Zhou S, Liu B, Zheng D *et al.* VFDB 2025: an integrated resource for exploring anti-virulence compounds. *Nucleic Acids Res* 2025;**53**:D871–7. <https://doi.org/10.1093/nar/gkae968>
145. Badel C, Da Cunha V, Catchpole R *et al.* WASPS: web-assisted symbolic plasmid synteny server. *Bioinformatics*. 2020;**36**:1629–31. <https://doi.org/10.1093/bioinformatics/btz745>
146. 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>
147. Hofmann-Apitius M, Ball G, Gebel S *et al.* Bioinformatics Mining and Modeling Methods for the Identification of Disease Mechanisms in Neurodegenerative Disorders. *International Journal of Molecular Sciences* 2015;**16**:29179–206. <https://www.mdpi.com/1422-0067/16/12/26148>
148. Rodella C, Lazaridi S, Lemmin T. TemBERTure: advancing protein thermostability prediction with deep learning and attention mechanisms. *Bioinformatics. Advances*. 2024;**4**:vbae103. <https://doi.org/10.1093/bioadv/vbae103>
149. Zhou Z, Zhang L, Yu Y *et al.* Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning. *Nat Commun* 2024;**15**:5566. <https://doi.org/10.1038/s41467-024-49798-6>
150. Wang J, Yang Z, Chen C *et al.* MPEK: a multitask deep learning framework based on pretrained language models for enzymatic reaction kinetic parameters prediction. *Briefings in Bioinformatics* 2024;**25**:bbae387. <https://doi.org/10.1093/bib/bbae387>
151. Wang H, Fu T, Du Y *et al.* Scientific discovery in the age of artificial intelligence. *Nature* 2023;**620**:47–60. <https://www.nature.com/articles/s41586-023-06221-2>
152. Zhang Y, Khan SA, Mahmud A *et al.* Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence* 2025;**1**:14. <https://www.nature.com/articles/s44387-025-00019-5>
153. Frontiers | The future of fundamental science led by generative closed-loop artificial intelligence. [accessed 2026 Feb 18]. <https://www.frontiersin.org/journals/artificial-intelligence/articles/>. <https://doi.org/10.3389/frai.2026.1678539/full>
154. Lagani V, Triantafillou S, Ball G *et al.* Probabilistic Computational Causal Discovery for Systems Biology. *Uncertainty in Biology: A Computational Modeling Approach* 2016;33–73. https://doi.org/10.1007/978-3-319-21296-8_3
155. Dwivedi SK, Tjörnberg A, Tegnér J *et al.* Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder. *Nature Communications* 2020;**11**:856. <https://doi.org/10.1038/s41467-020-14666-6>
156. Lin T-T, Yang L-Y, Lu I-H *et al.* AI4AMP: an antimicrobial peptide predictor using physicochemical property-based encoding method and deep learning. *mSystems*. 2021;**6**:e0029921. <https://doi.org/10.1128/mSystems.00299-21>
157. 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>
158. Li C, Sutherland D, Hammond SA *et al.* AMPLify: attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens. *BMC Genomics* 2022;**23**:77. <https://doi.org/10.1186/s12864-022-08310-4>
159. Veltri D, Kamath U, Shehu A *et al.* Deep learning improves antimicrobial peptide recognition. *Bioinformatics* 2018;**34**:2740–7. <https://doi.org/10.1093/bioinformatics/bty179>
160. Bujdoš D, Walter J, O'Toole PW. aurora: a machine learning gwas tool for analyzing microbial habitat adaptation. *Genome Biology* 2025;**26**:66. <https://doi.org/10.1186/s13059-025-03524-7>
161. Alneberg J, Bjarnason BS, Bruijn I de *et al.* CONCOCT: Clustering cONTigs on COverage and ComposiTiOn. 2013. <http://arxiv.org/abs/1312.4038>. <https://doi.org/10.48550/arXiv.1312.4038>
162. Arango-Argoty G, Garner E, Pruden A *et al.* DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*. 2018;**6**:23. <https://doi.org/10.1186/s40168-018-0401-z>
163. Hannigan GD, Prihoda D, Palicka A *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res* 2019;**47**:e110. <https://doi.org/10.1093/nar/gkz654>
164. Kim GB, Kim JY, Lee JA *et al.* Functional annotation of enzyme-encoding genes using deep learning with transformer layers. *Nat Commun* 2023;**14**:7370. <https://doi.org/10.1038/s41467-023-43216-z>
165. Kulmanov M, Hoehndorf R. DeepGOZero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms. *Bioinformatics*. 2022;**38**:i238–45. <https://doi.org/10.1093/bioinformatics/btac256>
166. Tawfiq R, Niu K, Hoehndorf R *et al.* DeepGOMeta for functional insights into microbial communities using deep learning-based protein function prediction. *Sci Rep* 2024;**14**:31813. <https://doi.org/10.1038/s41598-024-82956-w>
167. Liang Q, Bible PW, Liu Y *et al.* DeepMicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genom Bioinform* 2020;**2**:lqaa009. <https://doi.org/10.1093/nargab/lqaa009>
168. 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>
169. Liu D, Young F, Lamb KD *et al.* Prediction of virus-host associations using protein language models and multiple instance learning. *PLoS Comput Biol* 2024;**20**:e1012597. <https://doi.org/10.1371/journal.pcbi.1012597>

170. 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>
171. Kelley DR, Liu B, Delcher AL *et al*. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res* 2012;**40**:e9. <https://doi.org/10.1093/nar/gkr1067>
172. Song Y, Yuan Q, Chen S *et al*. Accurately predicting enzyme functions through geometric graph learning on ESMFold-predicted structures. *Nat Commun* 2024;**15**:8180. <https://doi.org/10.1038/s41467-024-52533-w>
173. Hoarfrost A, Aptekmann A, Farfaiuk G *et al*. Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nature Communications* 2022;**13**:2606. <https://www.nature.com/articles/s41467-022-30070-8>
174. Santos-Júnior CD, Pan S, Zhao X-M *et al*. Macrel: antimicrobial peptide screening in genomes and metagenomes. *PeerJ*. 2020;**8**:e10555. <https://doi.org/10.7717/peerj.10555>
175. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*. 2016;**32**:605–7. <https://doi.org/10.1093/bioinformatics/btv638>
176. Edwards R. *Computational Genomics Manual* 2023 [accessed 2025 Nov 10]. <https://github.com/linsalrob/ComputationalGenomicsManual>. <https://doi.org/10.5281/zenodo.8178842>
177. Peres da Silva R, Suphavitai C, Nagarajan N. MetageNN: a memory-efficient neural network taxonomic classifier robust to sequencing errors and missing genomes. *BMC bioinformatics* 2024;**25**:153. <https://doi.org/10.1186/s12859-024-05760-3>
178. Lai S, Pan S, Sun C *et al*. metaMIC: reference-free misassembly identification and correction of de novo metagenomic assemblies. *Genome Biol* 2022;**23**:242. <https://doi.org/10.1186/s13059-022-02810-y>
179. Wichmann A, Buschong E, Müller A *et al*. MetaTransformer: deep metagenomic sequencing read classification using self-attention models. *NAR Genomics and Bioinformatics* 2023;**5**:lqad082. <https://doi.org/10.1093/nargab/lqad082>
180. Beghini F, McIver LJ, Blanco-Míguez A *et al*. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*. 2021;**10**:e65088. <https://doi.org/10.7554/eLife.65088>
181. Gregor I, Dröge J, Schirmer M *et al*. PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ*. 2016;**4**:e1603. <https://doi.org/10.7717/peerj.1603>
182. Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Research* 2007;**35**:W460–4. <https://doi.org/10.1093/nar/gkm363>
183. Pan S, Zhao X-M, Coelho LP. SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinformatics* 2023;**39**:i21–9. <https://doi.org/10.1093/bioinformatics/btad209>
184. Li B, Yang YT, Capra JA *et al*. Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. *PLoS Comput Biol* 2020;**16**:e1008291. <https://doi.org/10.1371/journal.pcbi.1008291>
185. Erickson E, Gado JE, Avilán L *et al*. Sourcing thermotolerant poly(ethylene terephthalate) hydrolase scaffolds from natural diversity. *Nat Commun* 2022;**13**:7850. <https://doi.org/10.1038/s41467-022-35237-x>
186. Karlicki M, Antonowicz S, Karnkowska A. Tiara: deep learning-based classification system for eukaryotic sequences. *Bioinformatics* 2022;**38**:344–50. <https://doi.org/10.1093/bioinformatics/btab672>
187. Sweet P, Burroughs M, Jang S *et al*. TolRad, a model for predicting radiation tolerance using Pfam annotations, identifies novel radiosensitive bacterial species from reference genomes and MAGs. *Microbiol Spectrum* 2024;**12**:e0383823. <https://doi.org/10.1128/spectrum.03838-23>
188. Chan PP, Lin BY, Mak AJ *et al*. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* 2021;**49**:9077–96. <https://doi.org/10.1093/nar/gkab688>
189. Nissen JN, Johansen J, Allesøe RL *et al*. Improved metagenome binning and assembly using deep variational autoencoders. *Nat Biotechnol* 2021;**39**:555–60. <https://doi.org/10.1038/s41587-020-00777-4>
190. Tampuu A, Bzhhalava Z, Dillner J *et al*. ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. *PLOS ONE* 2019;**14**:e0222271. <https://doi.org/10.1371/journal.pone.0222271>
191. Pronk LJ, Medema MH. Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. *Microb Genom* 2022;**8** [accessed 2022 Sept 13]: <https://www.microbiologyresearch.org/content/journal/mgen/>. <https://doi.org/10.1099/mgen.0.000823>
192. Malwe AS, Longwani U, Sharma VK. XenoBug: machine learning-based tool to predict pollutant-degrading enzymes from environmental metagenomes. *NAR Genomics and Bioinformatics* 2025;**7**:lqaf037. <https://doi.org/10.1093/nargab/lqaf037>
193. Lin Z, Akin H, Rao R *et al*. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. <https://doi.org/10.1126/science.ade2574>
194. Elnaggar A, Heinzinger M, Dallago C *et al*. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2022;**44**:7112–27. <https://doi.org/10.1109/TPAMI.2021.3095381>
195. Gligorijević V, Renfrew PD, Kosciolk T *et al*. Structure-based protein function prediction using graph convolutional networks. *Nat Commun* 2021;**12**:3168. <https://doi.org/10.1038/s41467-021-23303-9>
196. Khurana S, Rawi R, Kunji K *et al*. DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*. 2018;**34**:2605–13. <https://doi.org/10.1093/bioinformatics/bty166>
197. Bunne C, Roohani Y, Rosen Y *et al*. How to build the virtual cell with artificial intelligence: priorities and opportunities. *Cell* 2024;**187**:7045–63. <https://doi.org/10.1016/j.cell.2024.11.015>
198. Szalata A, Hrovatin K, Becker S *et al*. Transformers in single-cell omics: a review and new perspectives. *Nat Methods* 2024;**21**:1430–43. <https://doi.org/10.1038/s41592-024-02353-z>
199. Zhou Z, Ji Y, Li W *et al*. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. 2024. <http://arxiv.org/abs/2306.15006>. <https://doi.org/10.48550/arXiv.2306.15006>
200. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J *et al*. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nat Methods* 2025;**22**:287–97. <https://doi.org/10.1038/s41592-024-02523-z>
201. Nguyen E, Poli M, Faizi M *et al*. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. 2023. <http://arxiv.org/abs/2306.15794>. <https://doi.org/10.48550/arXiv.2306.15794>

202. Avsec Ž, Latysheva N, Cheng J *et al.* Advancing regulatory variant effect prediction with AlphaGenome. *Nature*. 2026;**649**:1206–18. <https://doi.org/10.1038/s41586-025-10014-0>
203. Karim MR, Islam T, Shajalal M *et al.* Explainable AI for Bioinformatics: Methods, Tools and Applications. *Briefings in Bioinformatics* 2023;**24**:bbad236. <https://doi.org/10.1093/bib/bbad236>
204. Toussaint PA, Leiser F, Thiebes S *et al.* Explainable artificial intelligence for omics data: a systematic mapping study. *Brief Bioinform* 2024;**25**:bbad453. <https://doi.org/10.1093/bib/bbad453>
205. de Weerd HA, Guala D, Gustafsson M *et al.* Latent space arithmetic on data embeddings from healthy multi-tissue human RNA-seq decodes disease modules. *Patterns*. 2024;**5**:101093. <https://doi.org/10.1016/j.patter.2024.101093>
206. Roy G, Prifti E, Belda E *et al.* Deep learning methods in metagenomics: a review. *Microb Genom* 2024;**10**:001231. <https://doi.org/10.1099/mgen.0.001231>
207. Chklovski A, Parks DH, Woodcroft BJ *et al.* CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* 2023;**20**:1203–12. <https://doi.org/10.1038/s41592-023-01940-w>
208. Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc Natl Acad Sci U S A* 2019;**116**:13996–4001. <https://doi.org/10.1073/pnas.1821905116>
209. Krawczyk PS, Lipinski L, Dziembowski A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res* 2018;**46**:e35. <https://doi.org/10.1093/nar/gkx1321>
210. Ren J, Song K, Deng C *et al.* Identifying viruses from metagenomic data using deep learning. *Quantitative Biology (Beijing, China)* 2020;**8**:64–77. <https://doi.org/10.1007/s40484-019-0187-4>
211. Peng C, Shang J, Guan J *et al.* ViralLM: empowering virus discovery through the genome foundation model. *Bioinformatics* 2024;**40**:btac704. <https://doi.org/10.1093/bioinformatics/bt704>
212. Prabakaran M. Intelligence in the Pluriverse: Critical Ethology and the Undoing of WEIRD Rationality. 2025 [accessed 2026 May 12]. <https://papers.ssrn.com/abstract=5341216>. <https://doi.org/10.2139/ssrn.5341216>
213. Ahmad A, Rahamtullah MR, Mishra R. Structural and functional adaptation in extremophilic microbial α -amylases. *Biophys Rev* 2022;**14**:499–515. <https://doi.org/10.1007/s12551-022-00931-z>
214. Kumar A, Alam A, Tripathi D *et al.* Protein adaptations in extremophiles: an insight into extremophilic connection of mycobacterial proteome. *Semin Cell Dev Biol* 2018;**84**:147–57. (SI: Antigen presentation). <https://doi.org/10.1016/j.semcdb.2018.01.003>
215. Panja AS, Maiti S, Bandyopadhyay B. Protein stability governed by its structural plasticity is inferred by physicochemical factors and salt bridges. *Sci Rep* 2020;**10**:1822. <https://doi.org/10.1038/s41598-020-58825-7>
216. Pud I, Urbaitis T, Gasiunas G *et al.* TemStaPro: protein thermostability prediction using sequence representations from protein language models.
217. Narang K, Nath A, Hemstrom W *et al.* HaloClass: salt-tolerant protein classification with protein language models. *Protein J* 2024;**43**:1035–44. <https://doi.org/10.1007/s10930-024-10236-7>
218. Gado JE, Knotts M, Shaw AY *et al.* Machine learning prediction of enzyme optimum pH. 2024:2023.06.22.544776. <https://www.biorxiv.org/content/10.1101/2023.06.22.544776v2>. <https://doi.org/10.1101/2023.06.22.544776>
219. El Ouaty O, El M'rini A, Nachite D *et al.* Sediment Quality Indices for the Assessment of Heavy Metal Risk in Nador Lagoon Sediments (Morocco) Using Multistatistical Approaches. *Sustainability* 2024;**16**:1921. <https://doi.org/10.3390/su16051921>
220. Sysoev M, Grötzinger SW, Renn D *et al.* Bioprospecting of Novel Extremozymes From Prokaryotes—The Advent of Culture-Independent Methods. *Frontiers in Microbiology* 2021;**12**. <https://doi.org/10.3389/fmicb.2021.630013>
221. Z Zhou, R Riley, S Kautsar, W Wu, R Egan, S Hofmeyr, S Goldhaber-Gordon, M Yu, H Ho, Liu F, *Et al.* GenomeOcean: An Efficient Genome Foundation Model Trained on Large-Scale Metagenomic Assemblies. 2025 [accessed 2025 Oct 30]. <https://www.biorxiv.org/content/>. <https://doi.org/10.1101/2025.01.30.635558>
222. Li Q, Wu W, Zhu Y *et al.* Wang Z. A Genomic Foundation Model for Metagenomic Annotation: Generanno, 2025. [accessed 2025 Oct 30]. <https://www.biorxiv.org/content/10.1101/2025.06.04.656517v3>, <https://doi.org/10.1101/2025.06.04.656517>
223. Peng X, Wang S, Wang M *et al.* Metabolic interdependencies in thermophilic communities are revealed using co-occurrence and complementarity networks. *Nat Commun* 2024;**15**:8166. <https://doi.org/10.1038/s41467-024-52532-x>
224. Bidkhorji G, Shoaie S. MIGRENE: the toolbox for microbial and individualized GEMs, Reactobiome and community network modelling. *Metabolites*. 2024;**14**:132. <https://doi.org/10.3390/metabo14030132>
225. Andersen KS, Zhao K, Agerskov A d L *et al.* Predicting microbial community structure and temporal dynamics by using graph neural network models. *Nature. Communications*. 2025;**16**:9124. <https://doi.org/10.1038/s41467-025-64175-7>
226. Mondon P, Ringenbach C, Marchand T *et al.* Thermus thermophilus fermentation medium for use in prevention and treatment of infrared related skin damages. 2017. <https://patents.google.com/patent/EP2986305B1/en>
227. Lal B, Reddy MRV, Agnihotri A *et al.* Process for enhanced recovery of crude oil from oil wells using novel microbial consortium. 2009. <https://patents.google.com/patent/US7484560B2/en>
228. de Jesus HE, Peixoto RS, Cury JC *et al.* Evaluation of soil bioremediation techniques in an aged diesel spill at the Antarctic peninsula. *Appl Microbiol Biotechnol* 2015;**99**:10815–27. <https://doi.org/10.1007/s00253-015-6919-0>
229. Alsulami R, Lehmann R, Khan SA *et al.* PrePR-CT: Predicting Perturbation Responses in Unseen Cell Types Using Cell-Type-Specific Graphs. 2024. <https://www.biorxiv.org/content/10.1101/2024.07.24.604816v1>. <https://doi.org/10.1101/2024.07.24.604816>
230. Nikolados E-M, Oyarzún DA. Deep learning for optimization of protein expression. *Curr Opin Biotechnol* 2023;**81**:102941. <https://doi.org/10.1016/j.copbio.2023.102941>