frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 23 September 2024
DOI 10.3389/fmicb.2024.1433224

OPEN ACCESS

EDITED BY
Ivan A. Berg,
University of Münster, Germany

REVIEWED BY
Nicole Buan,
University of Nebraska-Lincoln, United States
Rafael Bargiela,
Instituto de Catálisis y Petroleoquímca (ICP-CSIC), Spain

*CORRESPONDENCE
Filipa L. Sousa
✉ filipa.sousa@univie.ac.at

RECEIVED 15 May 2024
ACCEPTED 28 August 2024
PUBLISHED 23 September 2024

CITATION
Karavaeva V and Sousa FL (2024) Navigating the archaeal frontier: insights and projections from bioinformatic pipelines.
Front. Microbiol. 15:1433224.
doi: 10.3389/fmicb.2024.1433224

Navigating the archaeal frontier: insights and projections from bioinformatic pipelines

Val Karavaeva[1,2] and Filipa L. Sousa[1]*

[1]Genome Evolution and Ecology Group, Department of Functional and Evolutionary Ecology, University of Vienna, Vienna, Austria, [2]Vienna Doctoral School of Ecology and Evolution, University of Vienna, Vienna, Austria

Archaea continues to be one of the least investigated domains of life, and in recent years, the advent of metagenomics has led to the discovery of many new lineages at the phylum level. For the majority, only automatic genomic annotations can provide information regarding their metabolic potential and role in the environment. Here, genomic data from 2,978 archaeal genomes was used to perform automatic annotations using bioinformatics tools, alongside synteny analysis. These automatic classifications were done to assess how good these different tools perform in relation to archaeal data. Our study revealed that even with lowered cutoffs, several functional models do not capture the recently discovered archaeal diversity. Moreover, our investigation revealed that a significant portion of archaeal genomes, approximately 42%, remain uncharacterized. In comparison, within 3,235 bacterial genomes, a diverse range of unclassified proteins is obtained, with well-studied organisms like *Escherichia coli* having a substantially lower proportion of uncharacterized regions, ranging from <5 to 25%, and less studied lineages being comparable to archaea with the range of 35–40% of unclassified regions. Leveraging this analysis, we were able to identify metabolic protein markers, thereby providing insights into the metabolism of the archaea in our dataset. Our findings underscore a substantial gap between automatic classification tools and the comprehensive mapping of archaeal metabolism. Despite advances in computational approaches, a significant portion of archaeal genomes remains unexplored, highlighting the need for extensive experimental validation in this domain, as well as more refined annotation methods. This study contributes to a better understanding of archaeal metabolism and underscores the importance of further research in elucidating the functional potential of archaeal genomes.

KEYWORDS

archaea, energy metabolism, microbial dark matter, microbial ecology, microbial diversity

1 Introduction

Archaea, as a domain of life, has been a source of continual surprises (Cavicchioli, 2010; Jarrell et al., 2011), with ongoing discoveries helping us to understand the processes conserved in all domains of life and revealing novel types and unique features of metabolism (DiMarco et al., 1990; Weiss and Thauer, 1993; Verhees et al., 2003; Siebers and Schönheit, 2005; Thauer et al., 2008), unique structural features (Nickell et al., 2003; Moissl et al., 2005; Walsby, 2005; Matsumi et al., 2011; Zillig et al., 1981), and new genomes (Mara

et al., 2023; Zhang I. H. et al., 2024), and novel lineages at the highest rank. Examples of those would be the new supergroups of archaea such as DPANN (Dombrowski et al., 2019; Zhang R. Y. et al., 2024) and Asgard (Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017; Imachi et al., 2020; Rodrigues-Oliveira et al., 2023). Moreover, the archaeal domain continuously reveals new sides to already known aspects of microbial metabolism, with novel metabolic capabilities for old enzymes, such as the case of McrA, previously a marker for methanogenesis and anaerobic methane oxidation (Friedrich, 2005), whose role in metabolism of higher carbon compounds was shown (Goenrich et al., 2004; Scheller et al., 2013; Laso-Pérez et al., 2016; Musat et al., 2024) in agreement with early *in vitro* experiments with higher carbons (Gunsalus et al., 1978), or the discovery of the metabolic potential of *Ca.* Korarchaeota for dissimilatory sulfite reduction (McKay et al., 2019). Notably, Archaea have played a pivotal role in the evolution of eukaryotes, indicating their significance in the history of life on Earth (Spang et al., 2015; Mac Leod et al., 2019; Imachi et al., 2020; López-García and Moreira, 2020; Eme et al., 2023; Spang, 2023, to name a few). The presence of eukaryotic signature proteins within Asgard genomes also led to an increased interest in the archaeal cell biology in the last few years, with a myriad of papers published on the topic (van Wolferen et al., 2022; Charles-Orszag et al., 2024; Makarova et al., 2024).

Archaea are everywhere, including the gut and skin of humans and other animals (Thomas C. M. et al., 2022; Moissl-Eichinger et al., 2017), with possibly a beneficial role. Yet, primarily due to methodological limitations (Taffner et al., 2019; Song et al., 2019), and possibly the biases in funding towards pathogens or biotechnologically relevant organisms (Lasken and McLean, 2014), the role of archaea with their host and other microorganisms remains largely unknown. Furthermore, archaea are present in plants, where, besides ammonia-oxidizing archaea (AOA) in the rhizosphere and leaves (Taffner et al., 2019; Song et al., 2019), methanogens (Taffner et al., 2019; Taffner et al., 2018) and halophilic (salt-loving) archaea can also be found (Taffner et al., 2018; Yadav et al., 2015; Al-Mailem et al., 2010). Thus, besides the impact of AOA and their role in increasing plant yield, using metagenomic sequencing techniques, indirect roles in plant growth-promoting traits, such as auxin production and production of secondary metabolites to aid against pathogens, abiotic, and biotic stress, were proposed (Taffner et al., 2018). These examples clearly show the archaeal versatile roles across different ecosystems. Whether thriving in extreme environments (*Sulfolobales, Halobacteria*) or existing in more common settings (*Nitrososphaerota*; Chaban et al., 2006), archaea remain enigmatic due to their unique adaptations and historical research biases towards the study of (pathogenic) bacteria. For instance, so far, no one really knows all enzymes involved in archaeal ammonia oxidation (Schleper and Nicol, 2010).

Since the origin of life on our planet, archaeal microorganisms continue to be fundamental to biogeochemical cycles, profoundly influencing ecosystems and environmental processes (Falkowski et al., 2008; Zhang C. et al., 2023; Qi et al., 2024; Lyons et al., 2024; Baker et al., 2020). Archaea contribute significantly to cycles involving sulfur (S; Offre et al., 2013; Neukirchen et al., 2023), nitrogen (N; Huang et al., 2021; Offre et al., 2013; Leigh, 2000), carbon (C; Boetius et al., 2000; Offre et al., 2013; Fuchs, 2011; Zhang X. et al., 2023; Justice et al., 2012; Kaster et al., 2011; Thauer et al., 1977; Thauer et al., 2008), oxygen (O; Bandeiras et al., 2005; Teske, 2018; Luo et al., 2024), iron (Fe, Dong et al., 2021; Auernik and Kelly, 2008; Malik and Hedrich, 2022), and arsenic (As; Zhang C. et al., 2023; van Lis et al., 2013) across various habitats. Their metabolic versatility and resilience in extreme environments make archaea indispensable for maintaining the equilibrium of these elemental cycles, impacting nutrient availability, greenhouse gas emissions, and overall ecosystem health (Falkowski et al., 2008; Zhang X. et al., 2023; Qi et al., 2024; Lyons et al., 2024).

Methanogens, halophiles, thermophilic *Euryarchaeota* and *Thermoproteota* have become valuable model systems in molecular biology and biotechnology (Allers and Ngo, 2003; Kletzin, 2007; Soppa, 2006; Leigh et al., 2011; Costa and Whitman, 2023; De Lise et al., 2023; Pfeifer et al., 2021; Aparici-Carratalá et al., 2023), and currently these four groups of archaea boast well-established genetic systems. This advancement renders them ideal for use as model organisms and facilitates the expanded exploration of the functions of archaeal genes. However, the biotechnological potential of recently discovered archaeal lineages remains to be explored.

At the heart of archaeal diversity lies their genomic repertoire, comprising a finite set of protein building blocks, organized into pathways that facilitate biochemical reactions. One prominent example is methanogenesis, a pathway wherein certain archaea produce methane through anaerobic metabolism, essential for carbon cycling in environments like wetlands, alkaline hydrothermal vents, and animal digestive tracts (Angle et al., 2017; Jones et al., 1983; ver Eecke et al., 2012; Thomas P. D. et al., 2022), and that is proposed to have had an important role at the origin of Life (Martin and Russel, 2007). Additionally, many archaea engage in chemolithotrophy, deriving energy by oxidizing inorganic compounds such as hydrogen, sulfur, or iron (Thauer et al., 1977; Edwards et al., 2000; Pereira et al., 2011; Colman et al., 2020).

With the advent of metagenomics, many novel lineages have been discovered, for which mainly only metagenomic information is available for metabolic reconstructions using functional annotation pipelines. However, most of these are biased toward bacterial knowledge, with archaeal proteins many times falling out of the established cutoffs due to their natural diversity. Thus, it is important to assess how much of this diversity can be retrieved semi-automatically using functional annotation pipelines. Moreover, this approach can, in a systematic way, pinpoint gaps in knowledge, driving for the experimental characterization of archaeal proteins, as well as a redefinition of model design. Several studies regarding microbial dark matter, particularly Archaea, have been put forward, where the ratios vary between 30 and 80% (e.g., Makarova et al., 2019; Rinke et al., 2013; Jiao et al., 2020). More recently, deep learning was applied to genomes to get insights from microbial dark matter, showing how relevant the characterization of microbial dark matter is (Hoarfrost et al., 2022).

The question put out in this paper is: do existing automated prediction tools perform as well at assigning gene functions to archaea as to bacteria? Thus, to deepen our understanding of archaeal biology and metabolism, we performed a comprehensive mapping of genomic data from 2,978 archaeal genomic assemblies, belonging to 27 phyla (including unclassified Archaea) and compared the results to the ones obtained from a similar number of bacterial assemblies (175 phyla). This initiative aims to assess the gaps in predicted knowledge about archaea, and compare it to bacteria. Through

systematic exploration and analysis, we can pinpoint gaps in predictive knowledge and guide experimental studies with the aim of further understanding the diverse metabolic capabilities and ecological significance of archaea.

2 Materials and methods

2.1 Genomic dataset

A subset of our in-house dataset (over 190,000 genomes, 2,629 of which are archaeal; downloaded from NCBI in November 2019 with two *Acidianus ambivalens* and one *Ca. Lokiarchaeum ossiferum* assemblies added later; Supplementary Table 1; Rodrigues-Oliveira et al., 2023) was created by filtering these assemblies by completeness and contamination (calculated using the "Rinke method," Rinke et al., 2013), excluding all with contamination >20%. In addition, assemblies containing more than 10% contamination were excluded unless there were only two or less representatives per genus. Assemblies with low contamination were filtered for completeness: if the genus had more than two representatives, those with <40% completeness were excluded. In case of DPANN archaea, genomes were excluded only if their completeness was <20% and they had more than one representative per genus. To capture the recent sequenced diversity of archaea, additional genomes were downloaded from JGI (1,731 genomes). For this study, the total of 2,978 archaeal assemblies, belonging to 27 phyla (incl. "unclassified Archaea") were obtained. In addition, a set of 3,235 bacterial genomes, belonging to 175 phyla (2 representatives per genus) used for comparison (Supplementary Table 1).

2.2 Functional annotation

The 2,978 archaeal and the 3,235 bacterial genomic assemblies were functionally annotated using KEGG HMM profiles (version 2024-02-28, Kanehisa and Goto, 2000; using HMMER version 3.4, hmmer.org). The resulting hits were filtered first by cutoffs provided by KEGG for each model, and second, by lowering the KEGG cutoffs by 20% for most models, except for cytochrome *bc1* complex models, where the previously established in-house cutoffs were used (Supplementary Table 2). The cutoffs were lowered to account for the fact that the standard KEGG cutoffs do not always work for the archaeal sequences. If no KEGG cutoff was provided for a model, a cutoff of 50 was used to ensure the hits for these KOs of acceptable quality were still included in the analysis. The KEGG name, module and pathway information was mapped to the resulting annotations.

The dataset was additionally annotated using Interproscan (version 5.66–98.0; Jones et al., 2014), which includes the following databases: CDD (NCBI Conserved Domain Database; Lu et al., 2020), PFAM (Mistry et al., 2007), Gene3D (Lees et al., 2012), PANTHER (Thomas C. M. et al., 2022), SUPERFAMILY (Gough and Chothia, 2002; Wilson et al., 2009), ProSitePatterns and ProSiteProfiles (Expasy Prosite; Sigrist et al., 2013), NCBIfam (also known and further referred to as TIGRFAM, Li et al., 2021), FunFam (Sillitoe et al., 2013), Hamap (Pedruzzi et al., 2015), PIRSF (Wu et al., 2004), Coils (Lupas et al., 1991), MobiDB-lite (Necci et al., 2021), SMART (Letunic et al., 2021), PRINTS (Attwood et al., 2012). PANTHER annotations were further filtered to eliminate uncharacterized proteins, domains of unknown functions (DUF) and annotations solely as "membrane protein" or "conserved protein."

Furthermore, the archaeal genomes were annotated using the information obtained from DiSCo (Neukirchen and Sousa, 2021). Diamond Blast searches were also performed to assign arCOG (Makarova et al., 2015; Liu et al., 2021) classification to all genomes, by selecting best hits using as cutoffs >= 25% identity and E-value of <= 0^{-10}.

2.3 Sequence classification into "characterized" and "uncharacterized"

The resulting annotated hits were split into "characterized" and "uncharacterized" sets using the following strategy (as described in Supplementary Figure 1): If the sequence has a KEGG annotation with a KEGG pathway annotation "Function unknown," then it is classified as "uncharacterized"; if the KEGG pathway annotation is different, then the sequence is classified as "characterized." If the sequence has no KEGG annotation, then the PANTHER annotation is checked. If a PANTHER annotation is present and it is not in the curated list of uncharacterized PANTHERs (Supplementary Table 3), the sequence is classified as "characterized"; if it is in the list, the sequence is "uncharacterized." If no PANTHER annotation is present, then the NCBIfam (TIGRFAM) annotation is checked. If a TIGRFAM annotation is present in the curated list of "uncharacterized TIGRFAMs" (Supplementary Table 3), then the sequence is classified as "uncharacterized"; otherwise, it is assigned as "characterized." If no TIGRFAM annotation is available, the Hamap annotation is checked: if it is present, the sequence is classified as "characterized," otherwise, it is classified as "uncharacterized."

Sequences without any annotations were automatically classified as "uncharacterized." The order of the steps is partially arbitrary, and, starting with KEGG annotations, the classification steps can be run in a different order if preferred. The reason for selecting KEGG as initial step is three-fold: KEGG is a widely used database in which metabolic maps were constructed manually, and KEGG orthology is usually based on characterized enzymes or proteins. Lastly, KEGG provides modules and higher classifications of metabolism which are of interest for this analysis. This pipeline is available at https://github.com/valkaravaeva/protein_classification_tool.

2.4 Analysis of "uncharacterized" sequences

The mean, median, maximum, and minimum numbers of uncharacterized sequences were calculated per taxon (in percent of uncharacterized vs. total CDS per genome) at a phylum level. The PFAM annotations of archaea were analyzed, in terms of most common occuring domains per taxon (supergroup or phylum). The values per lineage were plotted as a boxplot using "ggplot2" package in R.

2.5 Comparison between "uncharacterized" archaeal sequences and ArCOGs

ArCOG annotation was used as a comparison to the pipeline in terms of uncharacterized proteins. Sequences without arCOG

annotation or with the functional model belonging to "S_Function_unknown," "4_Poorly_Characterised" and "R_General_Function_Prediction_only" category, or having no category were classified as "uncharacterized." The mean, median, max, and min percentages per phylum of "uncharacterized" sequences based on arCOGs were computed and plotted, as described in section 2.4. The intersection between uncharacterized proteins between both methods as well as the method specific were analyzed. The values per lineage were plotted as a boxplot using "ggplot2" package in R.

2.6 Analysis of "characterized" sequences

The set of "characterized" sequences was analyzed in terms of KEGG module completeness (computed in percent; accounting for alternative KOs and for complexes—see pipeline documentation and files at https://github.com/valkaravaeva/protein_classification_tool and additional files at FigShare: 10.6084/m9.figshare.25782123). Briefly, per assembly, each module, including the different alternatives for each step was considered complete if there were identified proteins for at least one route (100%). If one or more proteins were missing, the ratio of identified proteins versus the number of pathway proteins needed was calculated and multiplied by 100. In the case of complexes, a similar approach was taken, in this case, using the number of identified subunits as numerator. This information was used to analyze the metabolic potential of each genome and *a posteriori*, aggregated by phylum. Further analyses were focused on cofactor biosynthesis and energy metabolism. For this, KOs of selected gene markers were used to represent types of energy metabolism. In addition, in specific cases, existing KEGG modules were manually modified, or created, by either joining several modules for the same pathway or complex, or, as in case of riboflavin biosynthesis, since no KEGG module for the archaeal version is available, by using the BioCyc database entry for *M. jannaschii* (and corresponding KO annotations; Karp et al., 2019). Completeness of these manual modules was assessed in the same way as for original KEGG modules. Completeness of selected modules was plotted as a stacked barchart using "ggplot2" package in R. Taxonomic distribution of selected marker genes was plotted as a heatmap using R package "Pretty heatmaps" (https://cran.r-project.org/web/packages/pheatmap/pheatmap.pdf) and beautified in Inkscape.

3 Results

To determine how much of archaeal proteomes fall into the category of uncharacterized, a pipeline with several different steps was employed (see Supplementary Figure 1 and Materials and Methods). In total, 2,451,799 (40.7%, lowered KEGG cutoffs) out of 6,029,057 of proteins fall into the uncharacterized category, from where newly discovered lineages, such as *Ca.* Heimdallarchaeota and *Ca.* Woesearchaeota, have a mean of ~50% of proteins classified as uncharacterized (Figure 1; Supplementary Table 4). Within Archaea, 16 out of 27 groups (59%; including unclassified Archaea) have more than 40% of its proteins classified as uncharacterized, and only in two groups this ratio falls shortly below 30%. The average of uncharacterized proteins across all analyzed archaeal genomes is 42%. When examining the percentage of uncharacterized proteins per phylum, in bacteria, only 45 out of 175 phyla (25%) have more than 40% uncharacterized proteins (Figure 2, Candidate phyla in Supplementary Figure 2). When comparing model organisms from both domains, and even excluding *E. coli* (12%), there are 31 bacterial phyla where at least one organism has less than 25% uncharacterized proteins. In contrast, among archaea, only the *Candidatus* Bathyarchaeota and *Euryarchaeota* have at least one assembly with less than 25% uncharacterized proteins. Moreover, while lowering KEGG model cutoffs induced a change in the number of archaeal unclassified proteins, it did not affect the number of uncharacterized bacterial proteins, indicating that the models are optimized for this domain (Supplementary Table 4).

The two assemblies of the archaeal group with lowest median percentage of uncharacterized proteins, *Ca.* Nezhaarchaeota, have a low number of proteins (fewer than 1,700) and have completeness scores of 88 and 93%, and contamination of 0.6 and 4.3%, respectively. Their reduced genome, potentially associated with a symbiotic lifestyle, could explain the median percentage of uncharacterized proteins being below 30%. In any case, this value is still roughly three times the one found for model bacteria. This pinpoints the problems in reconstructing the metabolism of newly sequenced archaeal lineages.

Within the 2,451,799 unclassified proteins, 33.8% have PFAM annotation, while 66.2% (1,622,446) lack any annotation. Remarkably, with the exception of *Ca.* Nezhaarchaeota, *Ca.* Hadarchaeota, and *Ca.* Verstraetearchaeota, where 46.6, 55.4, and 59.0% of uncharacterized proteins lack PFAM annotations, respectively, all other archaeal phyla exhibit over 60% of uncharacterized proteins devoid of PFAM annotations, leaving even their domains unidentified. The uncharacterized proteins with PFAM domains have their annotations spread over 16,689 different PFAM entries, from where 3,725 correspond to domains or proteins with unknown functions. The remaining 12,964 PFAM domains are found in 829,353 proteins (33.8%), with 6,602 present in less than 10 proteins.

Notably, the prevalent PFAM among those with annotations is the PIN domain, characterized by three conserved acidic residues but limited conservation otherwise, which in eukaryotes is associated with ribonucleases (Arcus et al., 2011), and in prokaryotes, it is a component of the toxin-antitoxin system (TA; Arcus et al., 2011). In fact, ~44% of those PIN domains are in the proximity of genes annotated as nosB or Vap, being potentially part of a toxin-antitoxin system (TA, Arcus et al., 2011; Bunker et al., 2008) or close to CRISPR-Cas systems. The remaining PIN domains are in the vicinity of enzymes, ribosomal proteins or other uncharacterized genes. The large superfamily of PIN domain proteins was divided into families (Matelska et al., 2017) and a role as endo/exonucleases and/or part of the defense arsenal proposed (Matelska et al., 2017). The second most frequent domain is "LexA-binding inner membrane-associated putative hydrolase," which is found in phospholipases and in proteins belonging to the SOS network, which rescues cells from DNA damage (Zhang and Lin, 2012). The third most frequent domain is the "halobacterial output domain 1," which is specific for haloarchaea and haloviruses, and possibly involved in regulatory processes (Galperin et al., 2018). The fourth most frequent domain overall is the helix-turn-helix domain, usually found in transcriptional regulatory proteins and involved in DNA binding that, in some cases, can also be found in multidomain proteins for nucleotide recruitment, or involved in protein–protein interactions (Menon and Lawrence, 2013). In fact, among the 20 most frequent PFAMs, additional

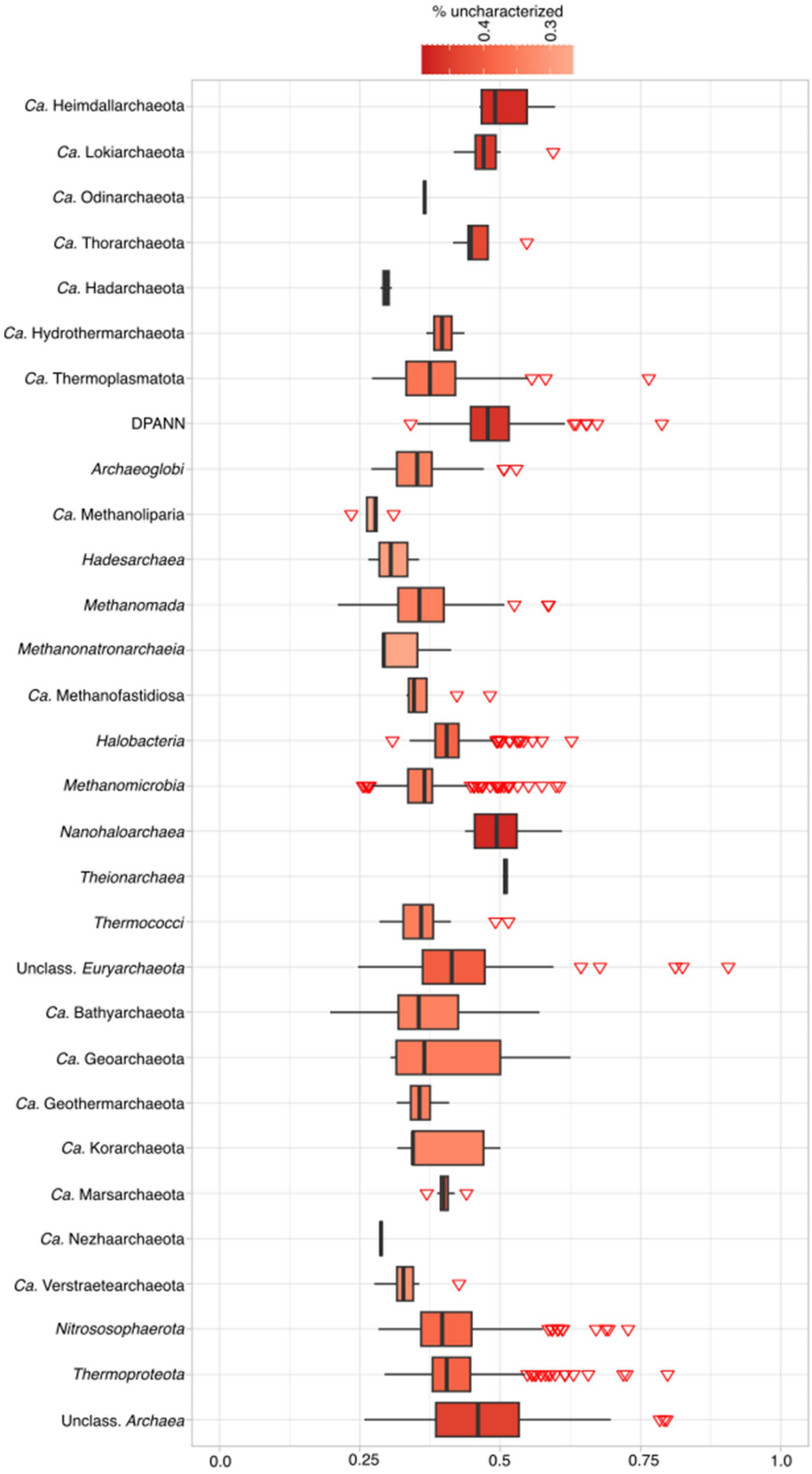

FIGURE 1
Percentage of archaeal unclassified proteins according to the pipeline classification per phylum. For complete taxonomic information see Supplementary Table 1. For exact percentages, see Supplementary Table 4.

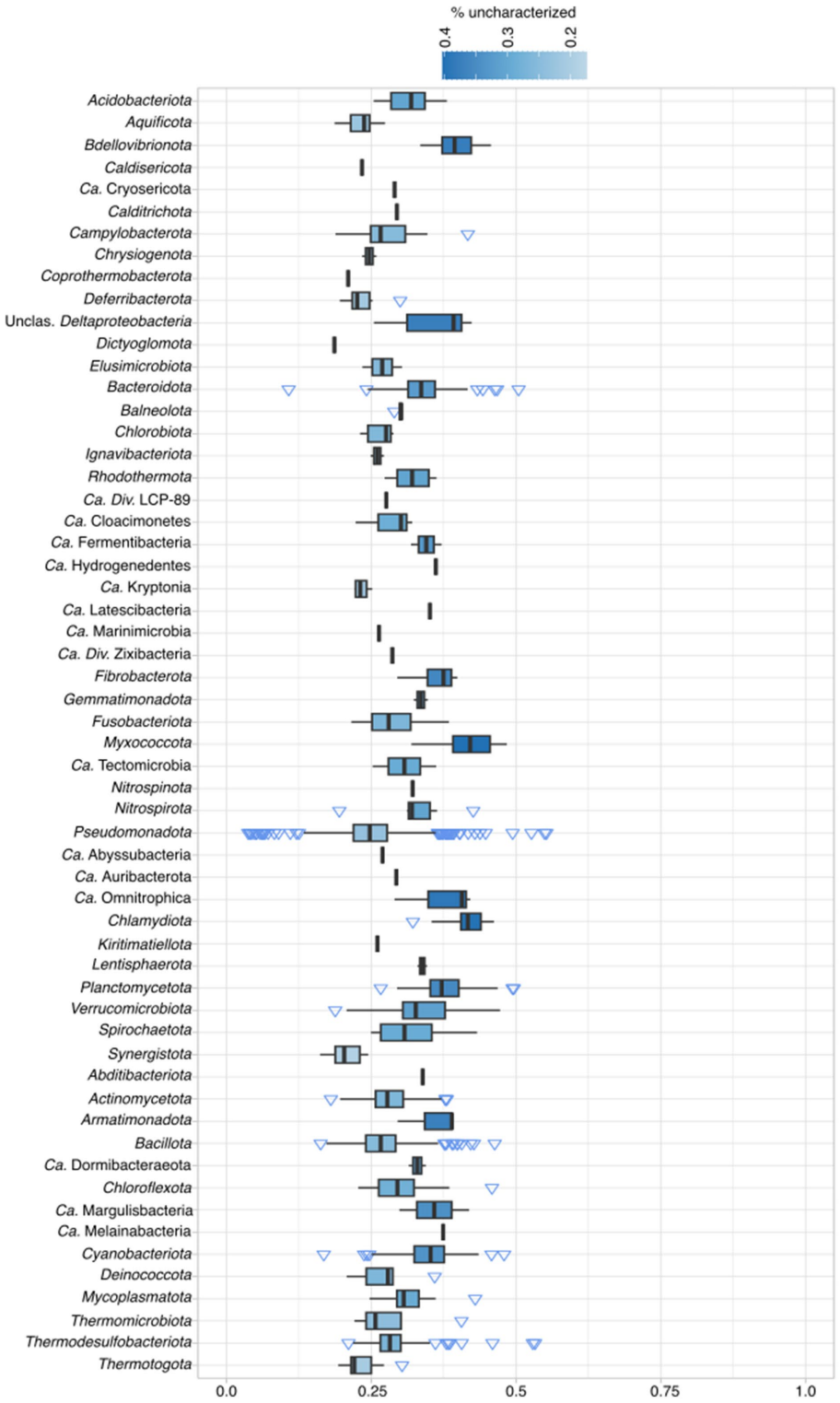

FIGURE 2
Percentage of unclassified bacterial proteins according to pipeline classification (see Materials and Methods) per phylum. For complete taxonomic information see Supplementary Table 1. For exact percentages, see Supplementary Table 4.

DNA-binding annotations emerge, including two winged helix-turn-helix domains, and the transcriptional regulator TrmB (Supplementary Table 5). TrmB, a sugar-specific transcriptional regulator of the trehalose/maltose ABC transporter from the hyperthermophilic archaeon *Thermococcus litoralis*, was previously characterized (Lee et al., 2003). Also, the *H. salinarum* reactive oxidative species regulator (RosR arCOG00006), which was experimentally characterized (Sharma et al., 2012) and whose

crystallographic structure is available (Kutnowski et al., 2019) is annotated as hypothetical protein and has no annotation within KEGG. These examples point to the misidentification of archaeal regulatory networks, in some cases, due to lack of models, in others, due to lack of characterization, as in the case of the ArsR/SmtB family (Lemmens et al., 2019). Moreover, the 6,602 PFAMs identified in fewer than 10 uncharacterized proteins, underscore the vast potential for innovation and diversity within this domain (Supplementary Table 5). Additionally, when examining proteins typically associated with metabolism, over 450 PFAMs (excluding radical SAM enzymes), corresponding to approximately 14,000 proteins, have annotations indicating the presence of hemes, FAD, NADH, molybdopterins, iron–sulfur clusters, oxidoreductases, or quinone-binding. This suggests that a portion of the archaeal metabolism remains not fully understood. A typical example would be the case of molybdopterin enzymes, which are ubiquitously present in prokaryotes, though the function of some is not known (Wells et al., 2020; Roy and Adams, 2002; Bevers et al., 2005).

The synteny analysis of unclassified proteins has shown that in 2,866 assemblies (96.2% of the dataset), there is at least one stretch of five or more genes without any available annotation (no PFAM). This number is even larger when considering the existence of pseudogenes in between uncharacterized ones. As a result, significant portions of the archaeal genomes remain without biological predictions, due to various factors such as the absence of models, assembly artifacts such as technical fusions, fissions or erroneous sequences (Padalko et al., 2024), inadequate CDS predictive methods for archaea (Dimonaco et al., 2022; Meng et al., 2022), or simply lack of biological knowledge. Notably, the uncharacterized genes within these regions are not necessarily involved in the same biological process, as genomic rearrangements frequently occur within genomes (Bobay and Ochman, 2017; Tillier and Collins, 2000; Darmon and Leach, 2014). When focusing on uncharacterized proteins for which PFAMs are available, particularly those which could, *a priori*, give some indication regarding energy metabolism, we observe that, for some cases, the uncharacterized protein's PFAM agrees with the surrounding genes, e.g., PF00507 and PF00420 NADH–ubiquinone/plastoquinone_ oxidoreductase, _chain_3 and 4L from complex I surrounded by other Complex I subunits (Supplementary Table 6). This indicates that their nonidentification by other methods might be due to the model not accounting for the full range of sequence diversity. In this case, the full predicted complex could, with thorough analysis, be recovered. In other cases, putative complexes have no attributed annotation except PFAM, making their identification more difficult. Those are the cases, for instance, for Complex IV subunits in proximity of each other in known aerobic organisms, such as *Halobacteria*, where both subunit I and subunit II (the catalytic ones) are found within a distance of four or less genes devoid of further annotations. While subunit II tends to be a transmembrane short protein, devoid of cofactors (for exceptions see Pereira et al., 2011; Murali et al., 2022), subunit I is composed of a conserved set of 12 transmembrane helices, containing the ligands for the low-spin heme and for the binuclear center, composed of a high-spin heme and a copper ion. This subunit, outside of the HCO family, has homology only with nitric oxide reductases (Pereira et al., 2011). Thus, the subunit I fold is specific to these enzymes, and, possibly due to sequencing artifacts, falls below the usual model cutoffs. In this way, the complex IV, previously described to be present in *Ca.* Heimdallarchaeota assemblies (Spang et al., 2019; Bulzu et al., 2019), could not be identified. Even though *Halobacteria* thrive in oxic environments (Grant and Ross, 1986; Oren, 1994; Oren and Litchfield, 1999; Cui and Dyall-Smith, 2021), and several Asgard assemblies have been obtained from oxic conditions (Bulzu et al., 2019), additional experimental characterizations are necessary to ascertain whether these "HCOs" can reduce O_2, utilize alternative terminal electron acceptors, or even function effectively.

We compared the results of our pipeline (available at https://github.com/valkaravaeva/protein_classification_tool) with the functional classification given by arCOGs (Makarova et al., 2015; Liu et al., 2021), a tool developed specifically for the identification of archaeal clusters of orthologous groups. Depending on the lineages, either arCOG (18; Figure 3) or our pipeline (4) has less uncharacterized proteins (Figure 1), with 5 phyla achieving similar results (differences below 1%). However, overall, arCOG outperforms our pipeline by identifying approximately 350,000 fewer uncharacterized proteins in total (see Figure 4). This advantage is also evident in lineages with lower overall numbers, such as *Ca.* Woesearchaeota and *Ca.* Heimdallarchaeota, which have mean proportions of 47 and 48% "unclassified" proteins, respectively, compared to 51% for both lineages using our pipeline. Additionally, four out of 27 archaeal phyla show a ratio of unclassified proteins just below 30% using arCOGs, whereas 9 out of 27 have more than 40% uncharacterized sequences.

The large majority of the proteins only classified by arCOGs belong to informational (transcription, translation, replication), defense, mobilome and cellular processes (74%), in agreement with the effort of the authors of arCOGs in improving those modules (Makarova et al., 2015; Liu et al., 2021) combined with the underdevelopment of KEGG in those modules. On the other hand, within the over 350,000 proteins only annotated by our pipeline, and focusing on the ones with KEGG annotations (corresponding to 45% of the proteins only annotated by the pipeline), 62% belong to the metabolism category, with signaling and cellular processes (20%) and informational processes (17%) as following categories. Among the proteins from metabolism are, for instance, 2000 involved in methane metabolism, including several acetyl-CoA synthase and formylmethanofuran dehydrogenase subunits, over 8,300 proteins involved in energy metabolism such as sulfide:quinone oxidoreductases (Brito et al., 2009), thiosulfate:quinone oxidoreductases (Müller et al., 2004), and V/A-type H+/Na+-transporting ATPase subunits, from where the *Methanobrevibacter ruminantium* complex was experimentally validated (McMillan et al., 2011). Therefore, to avoid running both approaches and to standardize the data, the choice of annotation strategy should depend on the specific goals of the study.

However, arCOGs are built from a graph method in which, due to, e.g., gene losses, paralogues can be grouped together. Moreover, there is no relationship between KOs and arCOGs, which renders the mappings of pathways using arCOGs for a database as large as the one used in this paper, an "Herculean" task. Thus, we continued the analysis using KEGG annotations.

The functional classification of archaeal proteins allows to reconstruct their metabolic potential and pinpoint possible gaps within pathways to be further experimentally characterized. Using KEGG modules combined with a strategy to count for their completeness (see Materials and Methods), the full reconstruction of the metabolism of 2,978 genomes is presented in Supplementary Table 7. Out of the 479 modules, 115 had no hit for

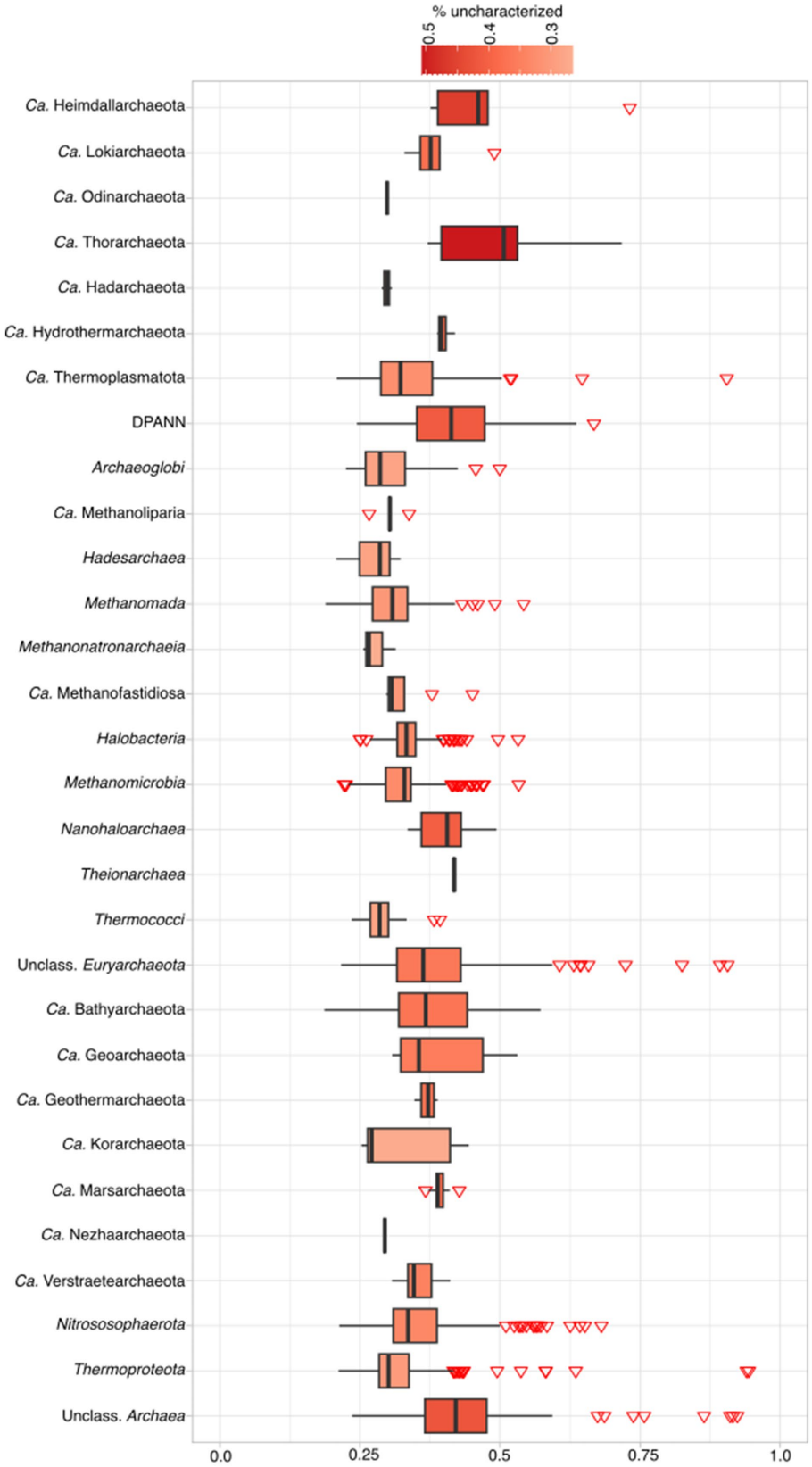

FIGURE 3
Percentage of unclassified archaeal proteins according to arCOG classification per phylum. For complete taxonomic information see Supplementary Table 1. For exact percentages, see Supplementary Table 4.

FIGURE 4
Number of proteins classified as "characterized" only by pipeline (green), only by arCOGs (pink), and by both methods (intersection).

archaeal proteins and only 20 were found to be complete or at least 75% complete in more than 50% of the assemblies in our dataset. These mainly correspond to the building blocks of life, such as nucleotides, amino acids and cofactors biosynthesis, ATP synthesis, lipid biosynthesis as well as carbon metabolism. The rationale behind setting a 75% completeness threshold includes instances where, despite adjustments, specific module components remain elusive. This also accounts for possible assembly incompleteness. Within bacteria, only 44 modules were not present in any of the genomes, and 60 modules found in more than 50% of the bacterial assemblies. The modules specific for bacteria are, for instance, gamma-aminobutyric acid (GABA) biosynthesis, a pathway present in several bacteria (Iorizzo et al., 2023) that has homologues within archaea which perform different functions (Tomita et al., 2014; Falb et al., 2008), or anoxygenic photosynthesis, a process that is absent in archaea (Hohmann-Marriott and Blankenship, 2011). Other modules address specific bacterial systems, such as antibiotic resistance or the synthesis of secondary metabolites unique to certain lineages.

In addition, several KEGG modules were modified to fill in gaps regarding archaeal metabolism not included in the original modules. For instance, in the cobalamin biosynthesis module, the decision to include CbiX, a homologue of CbiK that in some organisms performs the same reaction (Raux et al., 2003), was made to enhance the module's completeness, since this protein was initially absent. However, this adjustment, combined with lowered cutoffs, did not result in the increase of completeness of this module as expected, since there was no identification of other expected proteins within the module, such as the adenosylcobinamide kinase/adenosylcobinamide-phosphate guanylyltransferase CobP/CobU (K02231) because archaea do not have CobP/CobU but rather use CobY (K19712; Rodionov et al., 2003), which is not a part of the KEGG cobalamin biosynthesis modules. Even considering that some archaea might not have cobalamin biosynthesis, this scenario highlights the difficulty of defining a cutoff that accurately reflects the presence of all essential components, especially in complex biosynthetic pathways. Moreover, not all of complexes are part of KEGG modules. This is the case of the Ech and Ehb membrane hydrogenases (Marreiros et al., 2013; Marreiros et al., 2016) present in many methanogens, or the thiol:fumarate reductases (Heim et al., 1998), a complex whose subunits are homologous to the catalytic subunits of Complex II (Lancaster, 2002; Karavaeva and Sousa, 2023). Another problem is the existence of several modules for the same complexes, as, e.g., in the cases of succinate dehydrogenases/fumarate reductases, heme-copper oxygen reductases, and the *bc*1 complex. This leads to the existence of many archaeal complexes that have chimeric classifications according to KEGG modules, i.e., one subunit being part of one module and the other(s) belonging to another module of the same complex. This leads to modular incompleteness and hinders the usage of KEGG modules as a proxy of archaeal metabolism. For the cases mentioned above, we considered the module present if the subunits were identified, regardless of the KO module classification, meaning KEGG modules were merged, and different possible KOs would represent the same subunit. Completing this information with TIGRFam/NCBIFam and BioCyc information for selected modules (as described in Materials and Methods) led to an increase in module and pathway completeness. Still, in most of the cases, modules fall below the 75% completeness cutoffs (Supplementary Table 7). These results suggest that our understanding of the metabolic diversity and the distribution of biosynthetic pathways among archaea is still not included into databases, and the known existing gap between Bacteria and Archaea knowledge is even more pronounced at the level of automatic annotations.

Looking in detail to the different pathways for coenzymes and cofactors biosynthesis, we can observe that regarding heme biosynthesis in archaea (Supplementary Figure 3), the siroheme-dependent route is the most widely distributed, with the coproporphyrin-dependent pathway found to be complete in some *Halobacteria*, as already described (Dailey et al., 2017), as well as in one genome of *Ca.* Hydrothermarchaeota. Interestingly, within *Ca.* Heimdallarchaeota and some unclassified Euryarchaeota, the protoporphyrin-dependent heme biosynthesis was found. *Ca.* Heimdallarchaeota organisms have a mitochondrial-like electron-transport chain, being able to respire oxygen (Zaremba-Niedzwiedzka et al., 2017). This is not found in the majority of the other Asgard lineages and might be the result of HGT events. Since *Ca.* Heimdallarchaeota is also one of the few archaeal groups with protoporphyrin-dependent heme biosynthesis, this pathway might also have been acquired by HGT. Previously, several studies have reported on large events of interdomain HGT for archaea (Koonin and Wolf, 2008; Nelson-Sathi et al., 2012; Nelson-Sathi et al., 2015), and *Ca.* Heimdallarchaeota might be one of these cases. Of note, within our dataset, many other archaea were found to contain partial protoporphyrin-dependent heme biosynthesis pathways. However, this module also contains the universal tetrapyrrole biosynthesis part, common to the biosynthesis of all tetrapyrroles cofactors (heme, cobalamin, siroheme, F_{430}) that are all present in Archaea.

Regarding cobalamin biosynthesis, a full pathway is found in *Ca.* Thermoplasmatota, *Archaeoglobi*, *Ca.* Methanoliparia, *Methanomada, Methanonatronarchaeia, Halobacteria, Methanomicrobia*, *Ca.* Marsarchaeota, *Nitrososphaerota, Thermoproteota*, and unclassified Archaea. However, in most of these lineages, there are genomes that contain only a partial pathway, due to either not passing the cutoffs (especially in the case of CbiJ) or having no KO annotation for a fused protein (e.g., fusions of CbiK/CbiX chelatase and HmbS/HemC in certain *Archaeoglobi* genomes have only the KO annotation for the last protein). Fusion and fission events are a process common in Archaea, as shown in recent large-scale analysis (Padalko et al., 2024).

Complete pathways for the biosynthesis of menaquinone were found in *Ca.* Thermoplasmatota, *Archaeoglobi*, DPANN, *Thermoproteota, Methanomicrobia*, and unclassified Archaea

(Supplementary Figure 4). However, only in *Archaeoglobi* and *Ca.* Hydrothermarchaeota, they were present in the majority of the taxon assemblies. The presence of menaquinone in Archaea has been previously reported for *Thermoproteus tenax* (Thurl et al., 1985). As expected, no archaeal organisms have the complete pathway for ubiquinone biosynthesis. However, many have partial pathways, indicating the presence of several enzymes, homologous to those involved in ubiquinone biosynthesis. Within Archaea, besides menaquinones, several organisms use Caldariella (Schäfer et al., 2002) or sulfoquinone (Elling et al., 2016) as main quinone. Since the biosynthesis of these alternative quinones remains, to our knowledge, not fully resolved, it is not clear if the ubiquinone biosynthesis homologues found in those lineages might play a role in other quinone biosyntheses, and those are good candidates for further experimental validations.

Contrary to menaquinone biosynthesis, riboflavin (incl. FMN/FAD; Figure 5 and Supplementary Figure 4) biosynthesis is found to be partially present in many archaeal lineages, being complete within several lineages, such as *Archaeoglobi, Halobacteria, Methanomada, Theionarchaea, Nitrososphaerota*, and *Thermococci*. FMN/FAD biosynthesis enzymes are present in all lineages, including DPANN. Even with our improved module for FAD biosynthesis, we noticed that the enzyme(s) responsible for converting GTP to 2,5-Diamino-6-(1-D-ribosylamino)pyrimidin-4(3H)-one-5′-phosphate are absent in most archaea, indicating a gap in knowledge that possibly only experimentalists can fill. The biosynthesis of F_{430} is, as expected, present in several methanogenic groups (Figure 6) being less spread than the biosynthesis of F_{420} that besides methanogens, is also found in *Archaeoglobi*, *Ca.* Heimdallarchaeota, *Ca.* Lokiarchaeota, *Halobacteria* and *Theionarchaea*. The dihydrofolate reductase, used as a marker for folate biosynthesis, is mainly found in most assemblies from *Halobacteria* and the related group *Nanohaloarchaea*.

Various types of energy metabolism were investigated using gene markers for arsenic, nitrogen, oxygen and sulfur metabolism (Figure 7). Our findings indicate that organisms capable of detoxifying arsenate include *Methanomada,* as well as unclassified *Euryarchaeota* and *Nitrososphaerota*. Regarding oxygen metabolism, both *bd* oxidases and heme-copper oxidases (Pereira et al., 2001) were detected in *Ca.* Heimdallarchaeota, *Ca.* Thermoplasmatota*, Halobacteria, Methanomicrobia, Ca.* Geoarchaeota, *Nitrososphaerota, Thermoproteota,* unclassified *Euryarchaeota*, and unclassified Archaea. Some lineages have genes that encode only *bd* oxidase (DPANN, *Archaeoglobi, Thermococci, Ca.* Korarchaeota, *Ca.* Geothermarchaeota), while others have only HCO genes (*Ca.* Marsarchaeota). However, many of the hits did not pass the cutoffs, even lowered cutoffs, such as the case for HCOs in *Ca.* Heimdallarchaeota, where, despite earlier evidence of presence of these enzymes in this specific lineage (Spang et al., 2019; Bulzu et al., 2019), only one out of five genomes recovered a partial HCO complex.

Six marker proteins/complexes were selected to cover the diversity of nitrogen metabolism, although not including the ammonia monooxygenase AmoA, which shares a KO with the methane monooxygenase PmoA (K10944), and hence they are difficult to differentiate (Holmes et al., 1995). The hits for K10944 were nonetheless found in the dataset, in *Nitrososphaerota* (ammonia-oxidizers; Pester et al., 2011), as expected. The distribution of the Nif nitrogenase, used as protein marker for nitrogen fixation, recovered a similar distribution to that described in Baker et al. (2020), being found in *Ca.* Thermoplasmatota, *Methanomicrobia*, and *Theionarchaea*. However, in our case, additional 11 lineages had hits for nitrogenases, such as *Archaeoglobi*, *Ca.* Methanoliparia*, Methanomada*. Possible explanations for this difference could be the inclusion of vanadium-dependent nitrogenase Vnf in our results, or our search for all Nif subunits, as compared to Baker et al. (2020) using only NifH as a marker. Other cases, such as nitrite reductases NirK/NirS, did not overlap. For example, none of the lineages analyzed in Baker et al. (2020) were reported to contain NirS, and only *Aigarchaeota* and *Nitrososphaerota* were said to contain NirK. However, while our dataset did not include *Aigarachaeota*, other lineages had a hit for NirK in our dataset (e.g., *Ca.* Heimdallarchaeota, *Ca.* Thermoplasmatota*, Thermoproteota*), and NirS was found in *Halobacteria* (a lineage not included in Baker et al., 2020 analysis). It is possible, however, that additional NirK hits are in fact false positives due to the NirK homology to multicopper oxidases (Bento et al., 2005; Solomon et al., 1996).

To cover dissimilatory sulfur oxidation and reduction in archaea, seven protein markers were selected, ranging from sulfur oxygenase reductase (SOR; Urich et al., 2004; Urich et al., 2006) and thiosulphate:quinone oxidoreductase (TQO; Müller et al., 2004), first characterized in *Acidianus ambivalens* and representing chemolithoautotrophic sulfur-oxidizing metabolism in *Thermoproteales* (*Sulfolobales*, *Acidobales*), to the DsrAB and Qmo proteins to mark the Dsr-dependent dissimilatory sulfate/sulfite reduction in Archaea. The results recapture the known diversity within this dataset (Neukirchen and Sousa, 2021; Anantharaman et al., 2018; Friedrich et al., 2001; Ghosh and Dam, 2009). However, some of the newly discovered archaeal lineages with metabolic potential for Dsr-dependent sulfur metabolism, such as *Ca. Methanodesulfokores washburnensis* (McKay et al., 2019) or Dsr-containing Aigarchaeota (Hua et al., 2018) are absent from our dataset, explaining why this metabolism was not found in those groups. On the contrary, sulfide:quinone oxidoreductases were found to be present across 16 different archaeal groups, such as *Ca.* Heimdallarchaeota, *Halobacteria*, *Ca.* Korarchaeota, and *Thermoproteota*. So far, archaeal Sqrs have only been characterized from *A. ambivalens* (Brito et al., 2009) and *C. maquilingensis* (Lencina et al., 2013), and due to their sequence homology with Ndh-II (Brito et al., 2009), it cannot be excluded that some of these results are false positives, and the distribution of Sqr in Archaea is, in fact, smaller.

Using gene markers for terminal oxidoreductases or central complexes to pinpoint metabolic traits, while effective in uncovering the potential for certain types of energy metabolism in Archaea, falls short of presenting a comprehensive view of the possible variability within energy metabolic strategies. This approach may overlook the emergence of novel complexes formed through the rearrangement of modular protein components into unique architectures, not accounted for in these types of analyses.

4 Discussion

The aim of this paper was straightforward: to conduct a large-scale investigation into what is known and what is yet to be discovered within the archaeal domain, and to assess how much of archaeal metabolism can be reconstructed automatically using computational approaches. However, this turned out to be a much more complex

FIGURE 5
Presence of cofactor biosynthesis in archaea (per phylum for most lineages, per class for Euryarchaeota, with Methanobacteria, Methanococci, and Methanopyri grouped into the Methanomada supergroup), based on modified KEGG modules. **(A)** Presence of NAD biosynthesis (via both Tryptophan and Aspartate). **(B)** Presence of FMN + FAD biosynthesis. **(C)** Presence of cobalamin biosynthesis (excluding the lower ligand synthesis). Dark purple indicates that the full module is present, light purple marks the presence of the incomplete module, white shows the absence of the module in a lineage.

analysis than initially thought, due to the biases of knowledge regarding the other two domains of life, the different pathways of Archaea, and the fact that, with the exponential increase in sequencing projects and discovery of new lineages, their sequence divergency (real or due to sequencing artifacts) cannot be scaled up/incorporated in real time to existing databases. It is well-known that most of the current biological knowledge is based on Bacteria and Eukaryotes, with little attention given to incorporating Archaea and their differences into metabolic modules and pathways. Archaeal metabolism and information processing can be different from the ones present in Bacteria and Eukaryotes, and archaeal unique biochemical pathways enables them to thrive in extreme environments

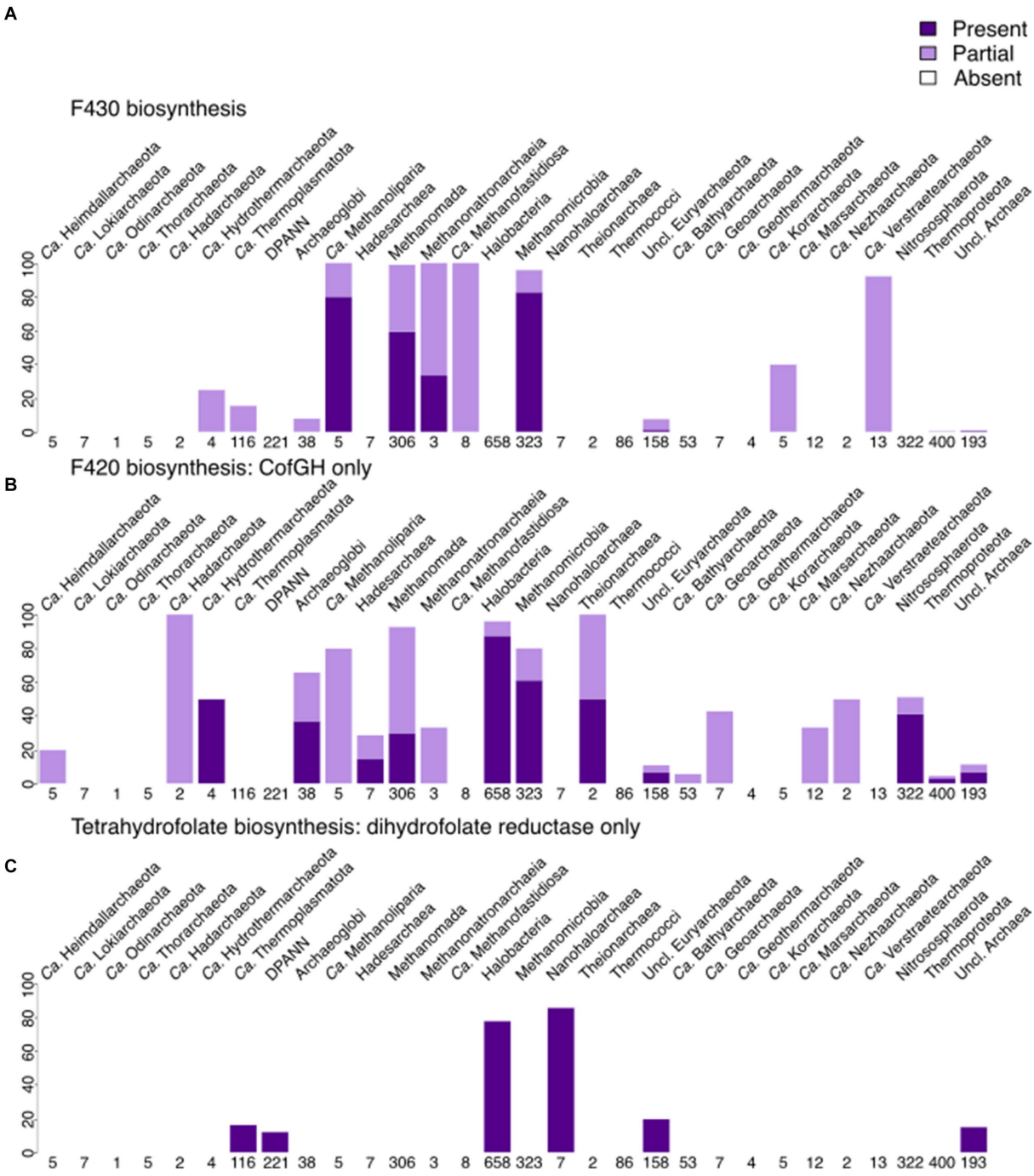

FIGURE 6
Presence of cofactor biosynthesis in archaea (per phylum for most lineages, per class for Euryarchaeota, with Methanobacteria, Methanococci, and Methanopyri grouped into the Methanomada supergroup), based on modified KEGG modules. **(A)** Presence of F_{430} biosynthesis. **(B)** Presence of F_{420} biosynthesis (includes only CofG + CofH as markers). **(C)** Presence of tetrahydrofolate biosynthesis (includes only dihydrofolate reductase as a marker). Dark purple indicates that the full module is present, light purple marks the presence of the incomplete module, white shows the absence of the module in a lineage.

and utilize diverse substrates, often relying on coenzymes and cofactors that necessitate entirely different enzymatic reactions compared to bacterial metabolic pathways. One prominent example is the incorporation of selenocysteine (Stadtman, 1974), often referred to as the 21st amino acid (Böck et al., 1991), that is found in proteins from the three domains of life (Rother and Quitzke, 2018) Selenocysteine is synthesized via a complex mechanism involving a specific tRNA and a dedicated set of enzymes (Chambers et al., 1986), and the archaeal synthesis is more related with the eukaryotic than with the bacterial one (Rother and Quitzke, 2018). This amino acid plays a crucial role in the function of several selenoenzyme families, including glutathione peroxidases and thioredoxin reductases, which

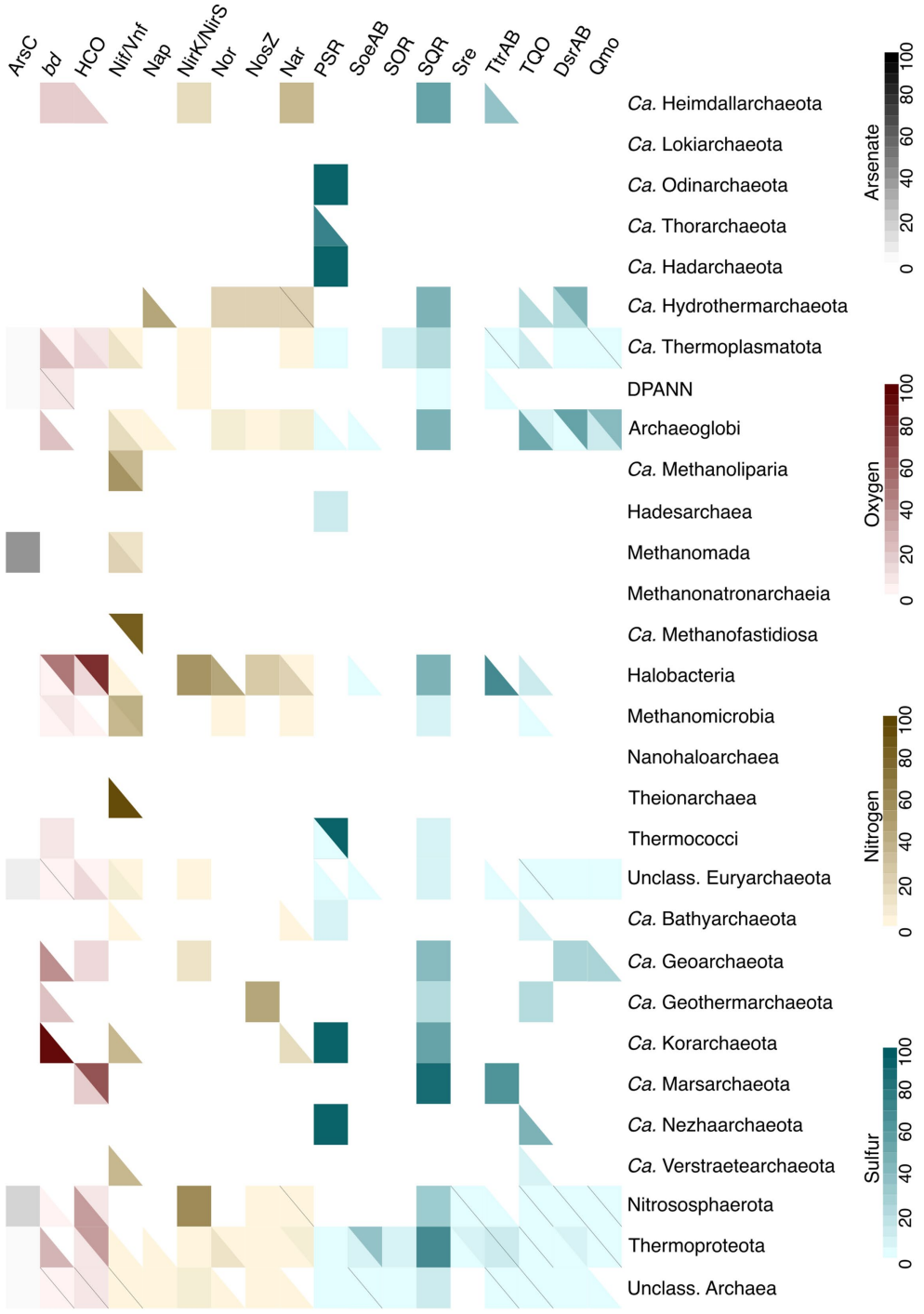

FIGURE 7
Presence of the marker proteins for different types of energy metabolism (KO and DiSCo annotations). Arsenate detoxification is marked in grey, oxygen metabolism in red, nitrogen metabolism in yellow, and sulfur metabolism in blue. ArsC: arsenate reductase, bd: bd oxidase, HCO: heme-copper oxidase, Nif/Vnf: nitrogenase, Nap: nitrate reductase, NirK/NirS: nitrite reductase, Nor: nitric oxide reductase, NosZ: nitrous oxide reductase, Nar: nitrate reductase, PSR: polysulfide reductase, SoeAB: sulfite:quinone dehydrogenase (subunits A and B), SOR: sulfur oxigenase/reductase, SQR: sulfide:quinone oxidoreductase, Sre: sulfur reductase, TtrAB: tetrathionate reductase (subunits A and B), TQO: thiosulfate:quinol oxidoreductase, DsrAB: dissimilatory sulfite reductase (subunits A and B), Qmo: quinone-modifiyng oxidoreductase. A full solid colored square indicates that, if a marker gene is a multisubunit complex, all of the subunits are present in a certain percent of the genomes. In cases where a square is split into two colors, the top part of the square indicates the percentage of genomes containing a full complex, while the bottom part shows the percent of genomes that have incomplete complexes.

are vital for oxidative stress management and redox reactions in archaeal cells (Rother and Quitzke, 2018). Another noteworthy cofactor found in a small number of methylamine-metabolizing archaea as well as a few bacteria (Tharp et al., 2018; Hao et al., 2002; Srinivasan et al., 2002; Borrel et al., 2014) is pyrrolysine, traditionally known as the 22nd amino acid. Pyrrolysine is encoded by the UAG

codon in some methanogenic Archaea (Hao et al., 2002; Srinivasan et al., 2002), and is integral to the activity of methyltransferase enzymes (Soares et al., 2005), which are involved in the final steps of methane production from methylated compounds (Rother and Quitzke, 2018). The existence of pyrrolysine highlights the diversity of genetic codes and implications for protein synthesis in archaea, further underlining their unique metabolic capabilities. Moreover, coenzymes such as coenzyme M (2-mercaptoethanesulfonic acid) and coenzyme F_{430} are central to the metabolic pathways of methanogens and other anaerobic archaea (Kaster et al., 2011), albeit also being found in some bacterial organisms. Coenzyme F_{430}, a nickel-containing porphyrin, plays an essential role in the enzymatic reaction catalyzed by methyl-coenzyme M reductase, where it participates in the final step of methane production, showcasing a highly specialized enzymatic system (Thauer et al., 2008). These unique biochemical components reveal how archaea have evolved distinct metabolic strategies that not only allow them to occupy a wide range of ecological niches but also highlight the evolutionary divergence between archaea and other life forms.

These differences, as well as the usage of non-archaeal sequences in the modules can lead to misassignments or false negatives in terms of functional predictions. For example, HCOs from *Ca.* Heimdallarchaeota fall short of the cutoffs for homology-based annotation. It is possible that this lineage's proteins diverge significantly from those in reference databases. However, this issue is not unique to *Ca.* Heimdallarchaeota; it also applies to halobacterial HCO proteins, indicating that not all divergence can be explained by this alone. Here we have shown 37.6% of the archaeal protein space remains uncharacterized, and that over 96% of archaeal metagenomes contain long stretches of genes, for which not even the protein domains (PFAM) are known. Also, within the uncharacterized proteins with PFAM annotations available, many contain cofactors and metal centers thought to have been playing a pivotal role since the origin of Life (Sanchez-Rocha et al., 2024; Weiss et al., 2016). For these uncharacterized cofactor-containing enzymes, the function is not yet known, and they may be a part of archaeal specific unexplored pathways, whose characterization would increase the diversity of microbial biology.

Enhancing current genomic classification databases and functional predictive models may involve refining them through additional analyses, like synteny analysis or integrating other omics data. This approach requires more sophisticated knowledge and operations rather than simple clicks to access and interpret this information. Some progress has been made regarding sulfur metabolism, where several dedicated tools, such as HMSS2 (Tanabe and Dahl, 2023) or DiSCo (Neukirchen and Sousa, 2021), were carefully built to identify specific types of metabolism, already integrating the current microbial diversity known. Progress has also been made in developing annotation-free strategies for identification of microbial dioxygen utilization from reads data, and in the last years, methods for TF identification from gene-expression data from quantitative phenotyping analysis (Darnell et al., 2017), approaches for a systematic inference of TF activity (Ma and Brent, 2021) and computational models for topological comparison of regulatory networks across the two domains of Life (Robinson and Schmid, 2018) have been developed. Another expanding area is phenomics, with several tools being developed in the last years, such as MicroPIE (Mao et al., 2016) to enable a fast extraction of phenotypic information from text records. Recently, the Functional Annotations of Prokaryotic Taxa (FAPROTAX) database (Louca et al., 2016) was tested for fast-functional screening of microbial metabolism, based on 16S RNA data (Sansupa et al., 2021) with promising results.

However, for an in-depth analysis of large datasets, better and faster tools need to be developed. Here is where statistical information theory (IT) plays an important role. Methods such as Mutual Information (MI; Vinga, 2014), Distance Correlation and its variants (Szé Kely and Rizzo, 2013; Monti et al., 2023) that already are useful to analyze, e.g., gene expression matrices, should be further developed to allow, e.g., comparisons of gene expression levels and inferences across independent samples. Moreover, in a recent study, MI was employed for pathway analysis, and, when applied to single-cell data, yielded robust and meaningful scores (Jeuken and Käll, 2024). For sequence data, IT can provide a broad range of inferences, from TF binding sites to gene mapping and phenotypic predictions, as comprehensively reviewed by Vinga (2014).

Artificial intelligence (AI), particularly machine learning algorithms, can also, in principle, provide valuable insights into archaeal metabolism by analyzing large genomic datasets and by filling in some of the gaps (Hoarfrost et al., 2022). Machine learning can aid in genome annotation (Chen et al., 2024; Khodabandelou et al., 2020), predict enzyme functions (Salas-Nuñez et al., 2024 and references within), and reconstruct metabolic pathways (Libbrecht and Noble, 2015). However, challenges and limitations persist in the field and the accuracy of metabolic reconstructions relies heavily on the quality of genomic and biochemical data available for archaeal species. Missing data and heterogenous datasets can lead to severe overfitting and other problems, as heavily discussed in the literature (Rodrigues, 2019; Xu and Jackson, 2019; Altman and Krzywinski, 2018). Customized models that consider the unique features of archaeal genomes and metabolic pathways can improve the accuracy and specificity of reconstructions. Integrating genomic, transcriptomic, proteomic, and metabolomic data can provide a comprehensive view of archaeal metabolism. AI and machine learning approaches that combine and analyze multi-omics data will facilitate more accurate reconstructions and deeper insights into the metabolic capabilities of Archaea, especially if coupled to statistical information theory. Perhaps this is the way to go in the future. But we must remember that computers only see zeros and ones (much better than we do), so we cannot forget that biology is more than math, and that without proper constraints and curated training data from experimental characterizations, distinguishing real results from artifacts is an almost impossible task. Moreover, as we have shown in this paper, without human manual curation, and extensive literature searches to get experimental archaeal characterized proteins to fill gaps in pathways, the distance between the vast amount of genomic information available and their analysis will only increase. In a perfect world, all data would be of high quality, with consistent information across different platforms. Additionally, linguistic and other barriers would be reduced to unite experimentalists, microbiologists, computational scientists, and mathematicians in the shared goal of closing this gap. Joining bottom-up with state of the art top-down predictive (ML) and inference (IT) approaches—merging the "*in silico*" and "*in vivo/vitro*" could increase the speed

at which we explore the archaeal world and disentangle its mysteries. This strategy would increase our understanding of archaeal metabolism, and life in general, offering new insights and opportunities for further research.

Data availability statement

The files with full annotation, all taxonomic levels for uncharacterized counts, and synteny tables, can be found at Figshare: 10.6084/m9.figshare.25782123. The pipeline (incl. examples of arCOG analysis scripts and plotting scripts) is available on GitHub: https://github.com/valkaravaeva/protein_classification_tool.

Author contributions

VK: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. FS: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 803768). FS thanks the Wiener Wissenschafts-, Forschungs-und Technologiefonds (grant agreement VRG15-007) and the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation program (grant agreement 803768).

Acknowledgments

The authors thank the Genome Evolution and Ecology group, Sinje Neukirchen, Maximilian Dreer, and Thomas Pribasnig for fruitful discussions. We also thank Brendan Mullins for advising on the pipeline development using Nextflow. Computational infrastructure and support were provided by the Life Science Computer Cluster (LiSC) at the University of Vienna.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2024.1433224/full#supplementary-material

References

Allers, T., and Ngo, H. P. (2003). Genetic analysis of homologous recombination in Archaea: *Haloferax volcanii* as a model organism. *Biochem. Soc. Trans.* 31, 706–710. doi: 10.1042/bst0310706

Al-Mailem, D. M., Sorkhoh, N. A., Marafie, M., Al-Awadhi, H., Eliyas, M., and Radwan, S. S. (2010). Oil phytoremediation potential of hypersaline coasts of the Arabian gulf using rhizosphere technology. *Bioresour. Technol.* 101, 5786–5792. doi: 10.1016/j.biortech.2010.02.082

Altman, N., and Krzywinski, M. (2018). The curse (s) of dimensionality. *Nat. Methods* 15, 399–400. doi: 10.1038/s41592-018-0019-x

Anantharaman, K., Hausmann, B., Jungbluth, S. P., Kantor, R. S., Lavy, A., Warren, L. A., et al. (2018). Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. *ISME J.* 12, 1715–1728. doi: 10.1038/s41396-018-0078-0

Angle, J. C., Morin, T. H., Solden, L. M., Narrowe, A. B., Smith, G. J., Borton, M. A., et al. (2017). Methanogenesis in oxygenated soils is a substantial fraction of wetland methane emissions. *Nat. Commun.* 8:1567. doi: 10.1038/s41467-017-01753-4

Aparici-Carratalá, D., Esclapez, J., Bautista, V., Bonete, M. J., and Camacho, M. (2023). Archaea: current and potential biotechnological applications. *Res. Microbiol.* 174:104080. doi: 10.1016/j.resmic.2023.104080

Arcus, V. L., McKenzie, J. L., Robson, J., and Cook, G. M. (2011). The PIN-domain ribonucleases and the prokaryotic Vap BC toxin-antitoxin array. *Protein Eng. Des. Sel.* 24, 33–40. doi: 10.1093/protein/gzq081

Attwood, T. K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P. B., Popov, I., et al. (2012). The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012. *Database (Oxford)* 2012:bas019. doi: 10.1093/database/bas019

Auernik, K. S., and Kelly, R. M. (2008). Identification of components of electron transport chains in the extremely thermoacidophilic crenarchaeon *Metallosphaera sedula* through iron and sulfur compound oxidation transcriptomes. *Appl. Environ. Microbiol.* 74, 7723–7732. doi: 10.1128/AEM.01545-08

Baker, B. J., De Anda, V., Seitz, K. W., Dombrowski, N., Santoro, A. E., and Lloyd, K. G. (2020). Diversity, ecology and evolution of Archaea. *Nat. Microbiol.* 5, 887–900. doi: 10.1038/s41564-020-0715-z Erratum in: Nat. Microbiol. 2020

Bandeiras, T. M., Pereira, M. M., Teixeira, M., Moenne-Loccoz, P., and Blackburn, N. J. (2005). Structure and coordination of CuB in the *Acidianus ambivalens* aa 3 quinol oxidase heme-copper center. *J. Biol. Inorg. Chem.* 10, 625–635. doi: 10.1007/s00775-005-0012-6

Bento, I., Martins, L. O., Gato Lopes, G., Arménia Carrondo, M., and Lindley, P. F. (2005). Dioxygen reduction by multi-copper oxidases; a structural perspective. *Dalton Trans.* 21, 3507–3513. doi: 10.1039/b504806k

Bevers, L. E., Bol, E., Hagedoorn, P. L., and Hagen, W. R. (2005). WOR5, a novel tungsten-containing aldehyde oxidoreductase from *Pyrococcus furiosus* with a broad substrate specificity. *J. Bacteriol.* 187, 7056–7061. doi: 10.1128/JB.187.20.7056-7061.2005

Bobay, L. M., and Ochman, H. (2017). The evolution of bacterial genome architecture. *Front. Genet.* 8:72. doi: 10.3389/fgene.2017.00072

Böck, A., Forchhammer, K., Heider, J., Leinfelder, W., Sawers, G., Veprek, B., et al. (1991). Selenocysteine: the 21st amino acid. *Mol. Microbiol.* 5, 515–520. doi: 10.1111/j.1365-2958.1991.tb00722.x

Boetius, A., Ravenschlag, K., Schubert, C. J., Rickert, D., Widdel, F., Gieseke, A., et al. (2000). A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* 407, 623–626. doi: 10.1038/35036572

Borrel, G., Gaci, N., Peyret, P., O'Toole, P. W., Gribaldo, S., and Brugère, J. F. (2014). Unique characteristics of the pyrrolysine system in the 7th order of methanogens:

implications for the evolution of a genetic code expansion cassette. *Archaea* 2015:374146. Erratum in: Archaea. 2015: 941836. doi: 10.1155/2015/941836

Brito, J. A., Sousa, F. L., Stelter, M., Bandeiras, T. M., Vonrhein, C., Teixeira, M., et al. (2009). Structural and functional insights into sulfide: quinone oxidoreductase. *Biochemistry* 48, 5613–5622. doi: 10.1021/bi9003827

Bulzu, P. A., Andrei, A. Ş., Salcher, M. M., Mehrshad, M., Inoue, K., Kandori, H., et al. (2019). Casting light on Asgardarchaeota metabolism in a sunlit microoxic niche. *Nat. Microbiol.* 4, 1129–1137. doi: 10.1038/s41564-019-0404-y

Bunker, R. D., McKenzie, J. L., Baker, E. N., and Arcus, V. L. (2008). Crystal structure of PAE0151 from *Pyrobaculum aerophilum*, a PIN-domain (Vap C) protein from a toxin-antitoxin operon. *Proteins* 72, 510–518. doi: 10.1002/prot.22048

Cavicchioli, R. (2010). Archaea--timeline of the third domain. *Nat. Rev. Microbiol.* 9, 51–61. doi: 10.1038/nrmicro2482

Chaban, B., Ng, S. Y., and Jarrell, K. F. (2006). Archaeal habitats--from the extreme to the ordinary. *Can. J. Microbiol.* 52, 73–116. doi: 10.1139/w05-147

Chambers, I., Frampton, J., Goldfarb, P., Affara, N., McBain, W., and Harrison, P. R. (1986). The structure of the mouse glutathione peroxidase gene: the selenocysteine in the active site is encoded by the 'termination' codon. *TGA. EMBO J.* 5, 1221–1227. doi: 10.1002/j.1460-2075.1986.tb04350.x

Charles-Orszag, A., Petek-Seoane, N. A., and Mullins, R. D. (2024). Archaeal actins and the origin of a multi-functional cytoskeleton. *J. Bacteriol.* 206:e0034823. doi: 10.1128/jb.00348-23

Chen, Z., Ain, N. U., Zhao, Q., and Zhang, X. (2024). From tradition to innovation: conventional and deep learning frameworks in genome annotation. *Brief. Bioinform.* 25:bbae138. doi: 10.1093/bib/bbae138

Colman, D. R., Lindsay, M. R., Amenabar, M. J., Fernandes-Martins, M. C., Roden, E. R., and Boyd, E. S. (2020). Phylogenomic analysis of novel Diaforarchaea is consistent with sulfite but not sulfate reduction in volcanic environments on early earth. *ISME J.* 14, 1316–1331. doi: 10.1038/s41396-020-0611-9

Costa, K. C., and Whitman, W. B. (2023). Model organisms to study Methanogenesis, a uniquely archaeal metabolism. *J. Bacteriol.* 205:e0011523. doi: 10.1128/jb.00115-23

Cui, H. L., and Dyall-Smith, M. L. (2021). Cultivation of halophilic archaea (class Halobacteria) from thalassohaline and athalassohaline environments. *Mar. Life Sci. Technol.* 3, 243–251. doi: 10.1007/s42995-020-00087-3

Dailey, H. A., Dailey, T. A., Gerdes, S., Jahn, D., Jahn, M., O'Brian, M. R., et al. (2017). Prokaryotic Heme biosynthesis: multiple pathways to a common essential product. *Microbiol. Mol. Biol. Rev.* 81, e00048–e00016. doi: 10.1128/MMBR.00048-16

Darmon, E., and Leach, D. R. (2014). Bacterial genome instability. *Microbiol. Mol. Biol. Rev.* 78, 1–39. doi: 10.1128/MMBR.00035-13

Darnell, C. L., Tonner, P. D., Gulli, J. G., Schmidler, S. C., and Schmid, A. K. (2017). Systematic discovery of archaeal transcription factor functions in regulatory networks through quantitative phenotyping analysis. *mSystems* 2, e00032–e00017. doi: 10.1128/mSystems.00032-17

De Lise, F., Iacono, R., Moracci, M., Strazzulli, A., and Cobucci-Ponzano, B. (2023). Archaea as a model system for molecular biology and biotechnology. *Biomol. Ther.* 13:114. doi: 10.3390/biom13010114

DiMarco, A. A., Bobik, T. A., and Wolfe, R. S. (1990). Unusual coenzymes of methanogenesis. *Annu. Rev. Biochem.* 59, 355–94. doi: 10.1146/annurev.bi.59.070190.002035

Dimonaco, N. J., Aubrey, W., Kenobi, K., Clare, A., and Creevey, C. J. (2022). No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study. *Bioinformatics.* 38, 1198–1207. doi: 10.1093/bioinformatics/btab827

Dombrowski, N., Lee, J.-H., Williams, T., Offre, P., and Spang, A. (2019). Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* 366:fnz008. doi: 10.1093/femsle/fnz008

Dong, Y., Shan, Y., Xia, K., and Shi, L. (2021). The proposed molecular mechanisms used by Archaea for Fe (III) reduction and Fe (II) oxidation. *Front. Microbiol.* 12:690918. doi: 10.3389/fmicb.2021.690918

Edwards, K. J., Bond, P. L., Gihring, T. M., and Banfield, J. F. (2000). An archaeal iron-oxidizing extreme acidophile important in acid mine drainage. *Science* 287, 1796–1799. doi: 10.1126/science.287.5459.1796

Elling, F. J., Becker, K. W., Könneke, M., Schröder, J. M., Kellermann, M. Y., Thomm, M., et al. (2016). Respiratory quinones in Archaea: phylogenetic distribution and application as biomarkers in the marine environment. *Environ. Microbiol.* 18, 692–707. doi: 10.1111/1462-2920.13086

Eme, L., Tamarit, D., Caceres, E. F., Stairs, C. W., De Anda, V., Schön, M. E., et al. (2023). Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* 618, 992–999. doi: 10.1038/s41586-023-06186-2

Falb, M., Müller, K., Königsmaier, L., Oberwinkler, T., Horn, P., von Gronau, S., et al. (2008). Metabolism of halophilic archaea. *Extremophiles* 12, 177–196. doi: 10.1007/s00792-008-0138-x

Falkowski, P. G., Fenchel, T., and Delong, E. F. (2008). The microbial engines that drive Earth's biogeochemical cycles. *Science* 320, 1034–1039. doi: 10.1126/science.1153213

Friedrich, M. W. (2005). Methyl-coenzyme M reductase genes: unique functional markers for methanogenic and anaerobic methane-oxidizing Archaea. *Methods Enzymol.* 397, 428–442. doi: 10.1016/S0076-6879(05)97026-2

Friedrich, C. G., Rother, D., Bardischewsky, F., Quentmeier, A., and Fischer, J. (2001). Oxidation of reduced inorganic sulfur compounds by bacteria: emergence of a common mechanism? *Appl. Environ. Microbiol.* 67, 2873–2882. doi: 10.1128/AEM.67.7.2873-2882.2001

Fuchs, G. (2011). Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Ann. Rev. Microbiol.* 65, 631–658. doi: 10.1146/annurev-micro-090110-102801

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2018). Phyletic distribution and lineage-specific domain architectures of archaeal two-component signal transduction systems. *J. Bacteriol.* 200, e00681–e00617. doi: 10.1128/JB.00681-17

Ghosh, W., and Dam, B. (2009). Biochemistry and molecular biology of lithotrophic sulfur oxidation by taxonomically and ecologically diverse bacteria and archaea. *FEMS Microbiol. Rev.* 33, 999–1043. doi: 10.1111/j.1574-6976.2009.00187.x

Goenrich, M., Mahlert, F., Duin, E. C., Bauer, C., Jaun, B., and Thauer, R. K. (2004). Probing the reactivity of Ni in the active site of methyl-coenzyme M reductase with substrate analogues. *J. Biol. Inorg. Chem.* 9, 691–705. doi: 10.1007/s00775-004-0552-1

Gough, J., and Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* 30, 268–272. doi: 10.1093/nar/30.1.268

Grant, W. D., and Ross, H. N. M. (1986). The ecology and taxonomy of halobacteria. *FEMS Microbiol. Rev.* 2, 9–15. doi: 10.1111/j.1574-6968.1986.tb01836.x

Gunsalus, R. P., Romesser, J. A., and Wolfe, R. S. (1978). Preparation of coenzyme M analogues and their activity in the methyl coenzyme M reductase system of Methanobacterium thermoautotrophicum. *Biochemistry* 17, 2374–2377. doi: 10.1021/bi00605a019

Hao, B., Gong, W., Ferguson, T. K., James, C. M., Krzycki, J. A., and Chan, M. K. (2002). A new UAG-encoded residue in the structure of a methanogen methyltransferase. *Science* 296, 1462–1466. doi: 10.1126/science.1069556

Heim, S., Künkel, A., Thauer, R. K., and Hedderich, R. (1998). Thiol: fumarate reductase (Tfr) from Methanobacterium thermoautotrophicum--identification of the catalytic sites for fumarate reduction and thiol oxidation. *Eur. J. Biochem.* 253, 292–299. doi: 10.1046/j.1432-1327.1998.2530292.x

Hoarfrost, A., Aptekmann, A., Farfañuk, G., and Bromberg, Y. (2022). Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nat. Commun.* 13:2606. doi: 10.1038/s41467-022-30070-8

Hohmann-Marriott, M. F., and Blankenship, R. E. (2011). Evolution of photosynthesis. *Annu. Rev. Plant Biol.* 62, 515–548. doi: 10.1146/annurev-arplant-042110-103811

Holmes, A. J., Costello, A., Lidstrom, M. E., and Murrell, J. C. (1995). Evidence that particulate methane monooxygenase and ammonia monooxygenase may be evolutionarily related. *FEMS Microbiol. Lett.* 132, 203–208. doi: 10.1016/0378-1097(95)00311-r

Hua, Z. S., Qu, Y. N., Zhu, Q., Zhou, E. M., Qi, Y. L., Yin, Y. R., et al. (2018). Genomic inference of the metabolism and evolution of the archaeal phylum Aigarchaeota. *Nat. Commun.* 9:2832. doi: 10.1038/s41467-018-05284-4

Huang, L., Chakrabarti, S., Cooper, J., Perez, A., John, S. M., Daroub, S. H., et al. (2021). Ammonia-oxidizing archaea are integral to nitrogen cycling in a highly fertile agricultural soil. *ISME Commun.* 1:19. doi: 10.1038/s43705-021-00020-4

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., et al. (2020). Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* 577, 519–525. doi: 10.1038/s41586-019-1916-6

Iorizzo, M., Paventi, G., and Di Martino, C. (2023). Biosynthesis of gamma-aminobutyric acid (GABA) by Lactiplantibacillus plantarum in fermented food production. *Curr. Issues Mol. Biol.* 46, 200–220. doi: 10.3390/cimb46010015

Jarrell, K. F., Walters, A. D., Bochiwal, C., Borgia, J. M., Dickinson, T., and Chong, J. P. J. (2011). Major players on the microbial stage: why archaea are important. *Microbiology* 157, 919–936. doi: 10.1099/mic.0.047837-0

Jeuken, G. S., and Käll, L. (2024). Pathway analysis through mutual information. *Bioinformatics* 40:btad 776. doi: 10.1093/bioinformatics/btad776

Jiao, J. Y., Liu, L., Hua, Z. S., Fang, B. Z., Zhou, E. M., Salam, N., et al. (2020). Microbial dark matter coming to light: challenges and opportunities. *Natl. Sci. Rev.* 8:nwaa280. doi: 10.1093/nsr/nwaa280

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). Inter pro scan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Jones, W. J., Leigh, J. A., Mayer, F., Woese, C. R., and Wolfe, R. S. (1983). *Methanococcus jannaschii* sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent. *Arch. Microbiol.* 136, 254–261. doi: 10.1007/BF00425213

Justice, N. B., Pan, C., Mueller, R., Spaulding, S. E., Shah, V., Sun, C. L., et al. (2012). Heterotrophic archaea contribute to carbon cycling in low-pH, suboxic

biofilm communities. *Appl. Environ. Microbiol.* 78, 8321–8330. doi: 10.1128/AEM.01938-12

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Karavaeva, V., and Sousa, F. L. (2023). Modular structure of complex II: an evolutionary perspective. *Biochim. Biophys. Acta Bioenerg.* 1864:148916. doi: 10.1016/j.bbabio.2022.148916

Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., et al. (2019). The bio Cyc collection of microbial genomes and metabolic pathways. *Brief. Bioinform.* 20, 1085–1093. doi: 10.1093/bib/bbx085

Kaster, A. K., Goenrich, M., Seedorf, H., Liesegang, H., Wollherr, A., Gottschalk, G., et al. (2011). More than 200 genes required for methane formation from H_2 and CO_2 and energy conservation are present in Methanothermobacter marburgensis and *Methanothermobacter thermautotrophicus*. *Archaea* 2011:973848. doi: 10.1155/2011/973848

Khodabandelou, G., Routhier, E., and Mozziconacci, J. (2020). Genome annotation across species using deep convolutional neural networks. *Peer J. Comput. Sci.* 6:e278. doi: 10.7717/peerj-cs.278

Kletzin, A. (2007) in General characteristics and important model organisms. ed. R. C. Archaea (Washington (DC): ASM Press).

Koonin, E. V., and Wolf, Y. I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* 36, 6688–6719. doi: 10.1093/nar/gkn668

Kutnowski, N., Shmulevich, F., Davidov, G., Shahar, A., Bar-Zvi, D., Eichler, J., et al. (2019). Specificity of protein-DNA interactions in hypersaline environment: structural studies on complexes of *Halobacterium salinarum* oxidative stress-dependent protein hs Ros R. *Nucleic Acids Res.* 47, 8860–8873. doi: 10.1093/nar/gkz604

Lancaster, C. R. (2002). Succinate: quinone oxidoreductases: an overview. *Biochim. Biophys. Acta* 1553, 1–6. doi: 10.1016/s0005-2728(01)00240-7

Lasken, R. S., and McLean, J. S. (2014). Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat. Rev. Genet.* 15, 577–584. doi: 10.1038/nrg3785

Laso-Pérez, R., Wegener, G., Knittel, K., Widdel, F., Harding, K. J., Krukenberg, V., et al. (2016). Thermophilic archaea activate butane via alkyl-coenzyme M formation. *Nature* 539, 396–401. doi: 10.1038/nature20152

Lee, S. J., Engelmann, A., Horlacher, R., Qu, Q., Vierke, G., Hebbeln, C., et al. (2003). Trm B, a sugar-specific transcriptional regulator of the trehalose/maltose ABC transporter from the hyperthermophilic archaeon *Thermococcus litoralis*. *J. Biol. Chem.* 278, 983–990. doi: 10.1074/jbc.M210236200

Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B. H., et al. (2012). Gene 3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Res.* 40, D465–D471. doi: 10.1093/nar/gkr1181. Erratum in: Nucleic Acids Res. 40 (10): 4725

Leigh, J. A. (2000). Nitrogen fixation in methanogens: the archaeal perspective. *Curr. Issues Mol. Biol.* 2, 125–131

Leigh, J. A., Albers, S. V., Atomi, H., and Allers, T. (2011). Model organisms for genetics in the domain Archaea: methanogens, halophiles, thermococcales and sulfolobales. *FEMS Microbiol. Rev.* 35, 577–608. doi: 10.1111/j.1574-6976.2011.00265.x

Lemmens, L., Maklad, H. R., Bervoets, I., and Peeters, E. (2019). Transcription regulators in Archaea: homologies and differences with bacterial regulators. *J. Mol. Biol.* 431, 4132–4146. doi: 10.1016/j.jmb.2019.05.045

Lencina, A. M., Ding, Z., Schurig-Briccio, L. A., and Gennis, R. B. (2013). Characterization of the type III sulfide: quinone oxidoreductase from Caldivirga maquilingensis and its membrane binding. *Biochim. Biophys. Acta* 1827, 266–275. doi: 10.1016/j.bbabio.2012.10.010

Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res.* 49, D458–D460. doi: 10.1093/nar/gkaa937

Li, W., O'Neill, K. R., Haft, D. H., DiCuccio, M., Chetvernin, V., Badretdin, A., et al. (2021). Ref Seq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Res.* 49, D1020–D1028. doi: 10.1093/nar/gkaa1105

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16, 321–332. doi: 10.1038/nrg3920

Liu, Y., Makarova, K. S., Huang, W. C., Wolf, Y. I., Nikolskaya, A. N., Zhang, X., et al. (2021). Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* 593, 553–557. doi: 10.1038/s41586-021-03494-3

López-García, P., and Moreira, D. (2020). The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nat. Microbiol.* 5, 655–667. doi: 10.1038/s41564-020-0710-4

Louca, S., Parfrey, L. W., and Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science* 353, 1272–1277. doi: 10.1126/science.aaf4507

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., et al. (2020). CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* 48, D265–D268. doi: 10.1093/nar/gkz991

Luo, Z. H., Li, Q., Xie, Y. G., Lv, A. P., Qi, Y. L., Li, M. M., et al. (2024). Temperature, pH, and oxygen availability contributed to the functional differentiation of ancient Nitrososphaeria. *ISME J.* 18:wrad 031. doi: 10.1093/ismejo/wrad031

Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. *Science* 252, 1162–1164. doi: 10.1126/science.252.5009.1162

Lyons, T. W., Tino, C. J., Fournier, G. P., Anderson, R. E., Leavitt, W. D., Konhauser, K. O., et al. (2024). Co-evolution of early earth environments and microbial life. *Nat. Rev. Microbiol.* 22, 572–586. doi: 10.1038/s41579-024-01044-y

Ma, C. Z., and Brent, M. R. (2021). Inferring TF activities and activity regulators from gene expression data with constraints from TF perturbation data. *Bioinformatics.* 37, 1234–1245. doi: 10.1093/bioinformatics/btaa947

Mac Leod, F., Kindler, G. S., Wong, H. L., Chen, R., and Burns, B. P. (2019). Asgard archaea: diversity, function, and evolutionary implications in a range of microbiomes. *AIMS Microbiol.* 5, 48–61. doi: 10.3934/microbiol.2019.1.48

Makarova, K. S., Tobiasson, V., Wolf, Y. I., Lu, Z., Liu, Y., Zhang, S., et al (2024). Diversity, origin, and evolution of the ESCRT systems. *mBio* 15:e0033524. doi: 10.1128/mbio.00335-24

Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2015). Archaeal clusters of orthologous genes (ar COGs): an update and application for analysis of shared features between Thermococcales, Methanococcales, and Methanobacteriales. *Life (Basel)* 5, 818–840. doi: 10.3390/life5010818

Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2019). Towards functional characterization of archaeal genomic dark matter. *Biochem. Soc. Trans.* 47, 389–398. doi: 10.1042/BST20180560

Malik, L., and Hedrich, S. (2022). Ferric Iron reduction in extreme Acidophiles. *Front. Microbiol.* 12:818414. doi: 10.3389/fmicb.2021.818414

Mao, J., Moore, L. R., Blank, C. E., Wu, E. H., Ackerman, M., Ranade, S., et al. (2016). Microbial phenomics information extractor (Micro PIE): a natural language processing tool for the automated acquisition of prokaryotic phenotypic characters from text sources. *BMC Bioinform.* 17:528. doi: 10.1186/s12859-016-1396-8

Mara, P., Geller-McGrath, D., Edgcomb, V., Beaudoin, D., Morono, Y., and Teske, A. (2023). Metagenomic profiles of archaea and bacteria within thermal and geochemical gradients of the Guaymas Basin deep subsurface. *Nat. Commun.* 14:7768. doi: 10.1038/s41467-023-43296-x

Marreiros, B. C., Batista, A. P., Duarte, A. M., and Pereira, M. M. (2013). A missing link between complex I and group 4 membrane-bound [NiFe] hydrogenases. *Biochim. Biophys. Acta* 1827, 198–209. doi: 10.1016/j.bbabio.2012.09.012

Marreiros, B. C., Calisto, F., Castro, P. J., Duarte, A. M., Sena, F. V., Silva, A. F., et al. (2016). Exploring membrane respiratory chains. *Biochim. Biophys. Acta* 1857, 1039–1067. doi: 10.1016/j.bbabio.2016.03.028

Martin, W., and Russell, M. J. (2007). On the origin of biochemistry at an alkaline hydrothermal vent. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 362, 1887–925. doi: 10.1098/rstb.2006.1881

Matelska, D., Steczkiewicz, K., and Ginalski, K. (2017). Comprehensive classification of the PIN domain-like superfamily. *Nucleic Acids Res.* 45, 6995–7020. doi: 10.1093/nar/gkx494

Matsumi, R., Atomi, H., Driessen, A. J., and van der Oost, J. (2011). Isoprenoid biosynthesis in Archaea--biochemical and evolutionary implications. *Res. Microbiol.* 162, 39–52. doi: 10.1016/j.resmic.2010.10.003

McKay, L. J., Dlakić, M., Fields, M. W., Delmont, T. O., Eren, A. M., Jay, Z. J., et al. (2019). Co-occurring genomic capacity for anaerobic methane and dissimilatory sulfur metabolisms discovered in the Korarchaeota. *Nat. Microbiol.* 4, 614–622. doi: 10.1038/s41564-019-0362-4

McMillan, D. G., Ferguson, S. A., Dey, D., Schröder, K., Aung, H. L., Carbone, V., et al. (2011). A1Ao-ATP synthase of *Methanobrevibacter ruminantium* couples sodium ions for ATP synthesis under physiological conditions. *J. Biol. Chem.* 286, 39882–39892. doi: 10.1074/jbc.M111.281675

Meng, K., Chung, C. Z., Söll, D., and Krahn, N. (2022). Unconventional genetic code systems in archaea. *Front. Microbiol.* 13:1007832. doi: 10.3389/fmicb.2022.1007832

Menon, S. K., and Lawrence, C. M. (2013). *Helix-Turn-Helix motif in Brenner's encyclopedia of genetics*. Second Edn, 142–145. Cambridge (MA): Academic Press.

Mistry, J., Bateman, A., and Finn, R. D. (2007). Predicting active site residue annotations in the Pfam database. *BMC Bioinformat.* 8:298. doi: 10.1186/1471-2105-8-298

Moissl, C., Rachel, R., Briegel, A., Engelhardt, H., and Huber, R. (2005). The unique structure of archaeal 'hami', highly complex cell appendages with nano-grappling hooks. *Mol. Microbiol.* 56, 361–370. doi: 10.1111/j.1365-2958.2005.04294.x

Moissl-Eichinger, C., Probst, A. J., Birarda, G., Auerbach, A., Koskinen, K., Wolf, P., et al. (2017). Human age and skin physiology shape diversity and abundance of Archaea on skin. *Sci. Rep.* 7:4039. doi: 10.1038/s41598-017-04197-4

Monti, F., Stewart, D., Surendra, A., Alecu, I., Nguyen-Tran, T., Bennett, S. A. L., et al. (2023). Signed distance correlation (SiDCo): an online implementation of distance correlation and partial distance correlation for data-driven network analysis. *Bioinformatics* 39:btad 210. doi: 10.1093/bioinformatics/btad210

Müller, F. H., Bandeiras, T. M., Urich, T., Teixeira, M., Gomes, C. M., and Kletzin, A. (2004). Coupling of the pathway of sulphur oxidation to dioxygen reduction: characterization of a novel membrane-bound thiosulphate:quinone oxidoreductase. *Mol. Microbiol.* 53, 1147–60. doi: 10.1111/j.1365-2958.2004.04193.x

Murali, R., Hemp, J., and Gennis, R. B. (2022). Evolution of quinol oxidation within the heme-copper oxidoreductase superfamily. *Biochim. Biophys. Acta Bioenerg.* 1863:148907. doi: 10.1016/j.bbabio.2022.148907

Musat, F., Kjeldsen, K. U., Rotaru, A. E., Chen, S. C., and Musat, N. (2024). Archaea oxidizing alkanes through alkyl-coenzyme M reductases. *Curr. Opin. Microbiol.* 79:102486. doi: 10.1016/j.mib.2024.102486

Necci, M., Piovesan, D., Clementel, D., Dosztányi, Z., and Tosatto, S. C. E. (2021). Mobi DB-lite 3.0: fast consensus annotation of intrinsic disorder flavors in proteins. *Bioinformatics* 36, 5533–5534. doi: 10.1093/bioinformatics/btaa1045

Nelson-Sathi, S., Dagan, T., Landan, G., Janssen, A., Steel, M., McInerney, J. O., et al. (2012). Acquisition of 1, 000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. USA* 109, 20537–20542. doi: 10.1073/pnas.1209119109

Nelson-Sathi, S., Sousa, F. L., Roettger, M., Lozada-Chávez, N., Thiergart, T., Janssen, A., et al. (2015). Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517, 77–80. doi: 10.1038/nature13805

Neukirchen, S., Pereira, I. A. C., and Sousa, F. L. (2023). Stepwise pathway for early evolutionary assembly of dissimilatory sulfite and sulfate reduction. *ISME J.* 17, 1680–1692. doi: 10.1038/s41396-023-01477-y

Neukirchen, S., and Sousa, F. L. (2021). DiSCo: a sequence-based type-specific predictor of Dsr-dependent dissimilatory Sulphur metabolism in microbial data. *Microb. Genom.* 7:000603. doi: 10.1099/mgen.0.000603

Nickell, S., Hegerl, R., Baumeister, W., and Rachel, R. (2003). Pyrodictium cannulae enter the periplasmic space but do not enter the cytoplasm, as revealed by cryo-electron tomography. *J. Struct. Biol.* 141, 34–42. doi: 10.1016/s1047-8477(02)00581-6

Offre, P., Spang, A., and Schleper, C. (2013). Archaea in biogeochemical cycles. *Ann. Rev. Microbiol.* 67, 437–457. doi: 10.1146/annurev-micro-092412-155614

Oren, A. (1994). The ecology of the extremely halophilic archaea. *FEMS Microbiol. Rev.* 13, 415–439. doi: 10.1111/j.1574-6976.1994.tb00060.x

Oren, A., and Litchfield, C. D. (1999). A procedure for the enrichment and isolation of Halobacterium. *FEMS Microbiol. Lett.* 173, 353–358. doi: 10.1111/j.1574-6968.1999.tb13525.x

Padalko, A., Nair, G., and Sousa, F. L. (2024). Fusion/fission protein family identification in Archaea. *mSystems* 9:e0094823. doi: 10.1128/msystems.00948-23

Pedruzzi, I., Rivoire, C., Auchincloss, A. H., Coudert, E., Keller, G., de Castro, E., et al. (2015). HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res.* 43, D1064–D1070. doi: 10.1093/nar/gku1002

Pereira, I. A., Ramos, A. R., Grein, F., Marques, M. C., da Silva, S. M., and Venceslau, S. S. (2011). A comparative genomic analysis of energy metabolism in sulfate reducing bacteria and archaea. *Front. Microbiol.* 2:69. doi: 10.3389/fmicb.2011.00069

Pereira, M. M., Santana, M., and Teixeira, M. (2001). A novel scenario for the evolution of haem-copper oxygen reductases. *Biochim. Biophys. Acta* 1505, 185–208. doi: 10.1016/s0005-2728(01)00169-4

Pester, M., Schleper, C., and Wagner, M. (2011). The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. *Curr. Opin. Microbiol.* 14, 300–306. doi: 10.1016/j.mib.2011.04.007

Pfeifer, K., Ergal, İ., Koller, M., Basen, M., Schuster, B., and Rittmann, S. K. R. (2021). Archaea biotechnology. *Biotechnol. Adv.* 47:107668. doi: 10.1016/j.biotechadv.2020.107668

Qi, Y. L., Chen, Y. T., Xie, Y. G., Li, Y. X., Rao, Y. Z., Li, M. M., et al. (2024). Analysis of nearly 3000 archaeal genomes from terrestrial geothermal springs sheds light on interconnected biogeochemical processes. *Nat. Commun.* 15:4066. doi: 10.1038/s41467-024-48498-5

Raux, E., Leech, H. K., Beck, R., Schubert, H. L., Santander, P. J., Roessner, C. A., et al. (2003). Identification and functional analysis of enzymes required for precorrin-2 dehydrogenation and metal ion insertion in the biosynthesis of sirohaem and cobalamin in *Bacillus megaterium*. *Biochem. J.* 370, 505–516. doi: 10.1042/BJ20021443

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437. doi: 10.1038/nature12352

Robinson, N. P., and Schmid, A. K. (2018). Conserved principles of transcriptional networks controlling metabolic flexibility in archaea. *Emerg. Top. Life Sci.* 2, 659–669. doi: 10.1042/ETLS20180036

Rodionov, D. A., Vitreshak, A. G., Mironov, A. A., and Gelfand, M. S. (2003). Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. *J. Biol. Chem.* 278, 41148–41159. doi: 10.1074/jbc.M305837200

Rodrigues, T. (2019). The good, the bad, and the ugly in chemical and biological data for machine learning. *Drug Discov. Today Technol.* 32-33, 3–8. doi: 10.1016/j.ddtec.2020.07.001

Rodrigues-Oliveira, T., Wollweber, F., Ponce-Toledo, R. I., Xu, J., Rittmann, S. K. R., Klingl, A., et al. (2023). Actin cytoskeleton and complex cell architecture in an Asgard archaeon. *Nature* 613, 332–339. doi: 10.1038/s41586-022-05550-y

Rother, M., and Quitzke, V. (2018). Selenoprotein synthesis and regulation in Archaea. *Biochim. Biophys. Acta. Gen. Subj.* 1862, 2451–2462. doi: 10.1016/j.bbagen.2018.04.008

Roy, R., and Adams, M. W. (2002). Characterization of a fourth tungsten-containing enzyme from the hyperthermophilic archaeon *Pyrococcus furiosus*. *J. Bacteriol.* 184, 6952–6956. doi: 10.1128/JB.184.24.6952-6956.2002

Salas-Nuñez, L. F., Barrera-Ocampo, A., Caicedo, P. A., Cortes, N., Osorio, E. H., Villegas-Torres, M. F., et al. (2024). Machine learning to predict enzyme-substrate interactions in elucidation of synthesis pathways: a review. *Meta* 14:154. doi: 10.3390/metabo14030154

Sanchez-Rocha, A. C., Makarov, M., Pravda, L., Novotný, M., and Hlouchová, K. (2024). Coenzyme-protein interactions since early life. *eLife* 13:RP94174. doi: 10.7554/eLife.94174.1

Sansupa, C., Wahdan, S. F. M., Hossen, S., Disayathanoowat, T., Wubet, T., and Purahong, W. (2021). Can we use functional annotation of prokaryotic taxa (FAPROTAX) to assign the ecological functions of soil Bacteria? *Appl. Sci.* 11:688. doi: 10.3390/app11020688

Schäfer, G., Anemüller, S., and Moll, R. (2002). Archaeal complex II: 'classical' and 'non-classical' succinate: quinone reductases with unusual features. *Biochim. Biophys. Acta* 1553, 57–73. doi: 10.1016/s0005-2728(01)00232-8

Scheller, S., Goenrich, M., Thauer, R. K., and Jaun, B. (2013). Methyl-coenzyme M reductase from methanogenic archaea: isotope effects on label exchange and ethane formation with the homologous substrate ethyl-coenzyme M. *J. Am. Chem. Soc.* 135, 14985–14995. doi: 10.1021/ja4064876

Schleper, C., and Nicol, G. W. (2010). Ammonia-oxidising archaea--physiology, ecology and evolution. *Adv. Microb. Physiol.* 57, 1–41. doi: 10.1016/B978-0-12-381045-8.00001-1

Sharma, K., Gillum, N., Boyd, J. L., and Schmid, A. (2012). The Ros R transcription factor is required for gene expression dynamics in response to extreme oxidative stress in a hypersaline-adapted archaeon. *BMC Genomics* 13:351. doi: 10.1186/1471-2164-13-351

Siebers, B., and Schönheit, P. (2005). Unusual pathways and enzymes of central carbohydrate metabolism in Archaea. *Curr. Opin. Microbiol.* 8, 695–705. doi: 10.1016/j.mib.2005.10.014

Sigrist, C. J., de Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., et al. (2013). New and continuing developments at PROSITE. *Nucleic Acids Res.* 41, D344–D347. doi: 10.1093/nar/gks1067

Sillitoe, I., Cuff, A. L., Dessailly, B. H., Dawson, N. L., Furnham, N., Lee, D., et al. (2013). New functional families (Fun Fams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.* 41, D490–D498. doi: 10.1093/nar/gks1211

Soares, J. A., Zhang, L., Pitsch, R. L., Kleinholz, N. M., Jones, R. B., Wolff, J. J., et al. (2005). The residue mass of L-pyrrolysine in three distinct methylamine methyltransferases. *J. Biol. Chem.* 280, 36962–36969. doi: 10.1074/jbc.M506402200

Solomon, E. I., Sundaram, U. M., and Machonkin, T. E. (1996). Multicopper oxidases and Oxygenases. *Chem. Rev.* 96, 2563–2606. doi: 10.1021/cr950046o

Song, G. C., Im, H., Jung, J., Lee, S., Jung, M. Y., Rhee, S. K., et al. (2019). Plant growth-promoting archaea trigger induced systemic resistance in *Arabidopsis thaliana* against Pectobacterium carotovorum and *Pseudomonas syringae*. *Environ. Microbiol.* 21, 940–948. doi: 10.1111/1462-2920.14486

Soppa, J. (2006). From genomes to function: haloarchaea as model organisms. *Microbiology* 152, 585–590. doi: 10.1099/mic.0.28504-0

Spang, A. (2023). Is an archaeon the ancestor of eukaryotes? *Environ. Microbiol.* 25, 775–779. doi: 10.1111/1462-2920.16323

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., et al. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 521, 173–179. doi: 10.1038/nature14447

Spang, A., Stairs, C. W., Dombrowski, N., Eme, L., Lombard, J., Caceres, E. F., et al. (2019). Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat. Microbiol.* 4, 1138–1148. doi: 10.1038/s41564-019-0406-9

Srinivasan, G., James, C. M., and Krzycki, J. A. (2002). Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. *Science* 296, 1459–1462. doi: 10.1126/science.1069588

Stadtman, T. C. (1974). Selenium biochemistry. *Science* 183, 915–922. doi: 10.1126/science.183.4128.915

Szé Kely, G. J., and Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *J. Multivar. Anal.* 117, 193–213. doi: 10.1016/j.jmva.2013.02.012

Taffner, J., Cernava, T., Erlacher, A., and Berg, G. (2019). Novel insights into plant-associated archaea and their functioning in arugula (*Eruca sativa* Mill.). *J. Adv. Res.* 19, 39–48. doi: 10.1016/j.jare.2019.04.008

Taffner, J., Erlacher, A., Bragina, A., Berg, C., Moissl-Eichinger, C., and Berg, G. (2018). What is the role of Archaea in plants? New insights from the vegetation of alpine bogs. *mSphere* 3:e0012200118. doi: 10.1128/msphere.00122-18

Tanabe, T. S., and Dahl, C. (2023). HMSS2: an advanced tool for the analysis of sulphur metabolism, including organosulphur compound transformation, in genome and metagenome assemblies. *Mol. Ecol. Resour.* 23, 1930–1945. doi: 10.1111/1755-0998.13848

Teske, A. (2018). Aerobic Archaea in iron-rich springs. *Nat. Microbiol.* 3, 646–647. doi: 10.1038/s41564-018-0168-9

Tharp, J. M., Ehnbom, A., and Liu, W. R. (2018). tRNAPyl: structure, function, and applications. *RNA Biol.* 15, 441–452. doi: 10.1080/15476286.2017.1356561

Thauer, R. K., Jungermann, K., and Decker, K. (1977). Energy conservation in chemotrophic anaerobic bacteria. *Bacteriol. Rev.* 41, 100–180. doi: 10.1128/br.41.1.100-180.1977

Thauer, R. K., Kaster, A. K., Seedorf, H., Buckel, W., and Hedderich, R. (2008). Methanogenic archaea: ecologically relevant differences in energy conservation. *Nat. Rev. Microbiol.* 6, 579–591. doi: 10.1038/nrmicro1931

Thomas, C. M., Desmond-Le Quéméner, E., Gribaldo, S., and Borrel, G. (2022). Factors shaping the abundance and diversity of the gut archaeome across the animal kingdom. *Nat. Commun.* 13:3358. doi: 10.1038/s41467-022-31038-4

Thomas, P. D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L. P., and Mi, H. (2022). PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci.* 31, 8–22. doi: 10.1002/pro.4218

Thurl, S., Buhrow, I., and Schäfer, W. (1985). Quinones from Archaebacteria, I. New types of menaquinones from the thermophilic archaebacterium Thermoproteus tenax. *Biol. Chem. Hoppe Seyler.* 366, 1079–83. doi: 10.1515/bchm3

Tillier, E., and Collins, R. (2000). Genome rearrangement by replication-directed translocation. *Nat. Genet.* 26, 195–197. doi: 10.1038/79918

Tomita, H., Yokooji, Y., Ishibashi, T., Imanaka, T., and Atomia, H. (2014). An archaeal glutamate decarboxylase homolog functions as an aspartate decarboxylase and is involved in β-alanine and coenzyme a biosynthesis. *J. Bacteriol.* 196, 1222–1230. doi: 10.1128/JB.01327-13

Urich, T., Bandeiras, T. M., Leal, S. S., Rachel, R., Albrecht, T., Zimmermann, P., et al. (2004). The Sulphur oxygenase reductase from *Acidianus ambivalens* is a multimeric protein containing a low-potential mononuclear non-haem iron Centre. *Biochem. J.* 381, 137–146. doi: 10.1042/BJ20040003

Urich, T., Gomes, C. M., Kletzin, A., and Frazão, C. (2006). X-ray structure of a self-compartmentalizing sulfur cycle metalloenzyme. *Science* 311, 996–1000. doi: 10.1126/science.1120306

Van Lis, R., Nitschke, W., Duval, S., and Schoepp-Cothenet, B. (2013). Arsenics as bioenergetic substrates. *Biochim. Biophys. Acta* 1827, 176–188. doi: 10.1016/j.bbabio.2012.08.007

Van Wolferen, M., Pulschen, A. A., Baum, B., Gribaldo, S., and Albers, S. V. (2022). The cell biology of archaea. *Nat. Microbiol.* 7, 1744–1755. doi: 10.1038/s41564-022-01215-8

Ver Eecke, H. C., Butterfield, D. A., Huber, J. A., Lilley, M. D., Olson, E. J., Roe, K. K., et al. (2012). Hydrogen-limited growth of hyperthermophilic methanogens at deep-sea hydrothermal vents. *Proc. Natl. Acad. Sci. USA* 109, 13674–13679. doi: 10.1073/pnas.1206632109

Verhees, C. H., Kengen, S. W., Tuininga, J. E., Schut, G. J., Adams, M. W., De Vos, W. M., et al. (2003). The unique features of glycolytic pathways in Archaea. *Biochem. J.* 375, 231–246. doi: 10.1042/BJ20021472. Erratum in: Biochem. J. 2004 Feb 1; 377 (Pt 3): 819-22

Vinga, S. (2014). Information theory applications for biological sequence analysis. *Brief. Bioinform.* 15, 376–389. doi: 10.1093/bib/bbt068

Walsby, A. E. (2005). Archaea with square cells. *Trends Microbiol.* 13, 193–195. doi: 10.1016/j.tim.2005.03.002

Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., et al. (2016). The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* 1:16116. doi: 10.1038/nmicrobiol.2016.116

Weiss, D. S., and Thauer, R. K. (1993). Methanogenesis and the unity of biochemistry. *Cell* 72, 819–822. doi: 10.1016/0092-8674(93)90570-g

Wells, M., Kanmanii, N. J., Al Zadjali, A. M., Janecka, J. E., Basu, P., Oremland, R. S., et al. (2020). Methane, arsenic, selenium and the origins of the DMSO reductase family. *Sci. Rep.* 10:10946. doi: 10.1038/s41598-020-67892-9

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., et al. (2009). SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* 37, D380–D386. doi: 10.1093/nar/gkn762

Wu, C. H., Nikolskaya, A., Huang, H., Yeh, L. S., Natale, D. A., Vinayaka, C. R., et al. (2004). PIRSF: family classification system at the protein information resource. *Nucleic Acids Res.* 32, 112D–1114D. doi: 10.1093/nar/gkh097

Xu, C., and Jackson, S. A. (2019). Machine learning and complex biological data. *Genome Biol.* 20:76. doi: 10.1186/s13059-019-1689-0

Yadav, A. N., Sachan, S. G., Verma, P., and Saxena, A. K. (2015). Prospecting cold deserts of north western Himalayas for microbial diversity and plant growth promoting attributes. *J. Biosci. Bioeng.* 119, 683–693. doi: 10.1016/j.jbiosc.2014.11.006

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, D., Juzokaite, L., Vancaester, E., et al. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 541, 353–358. doi: 10.1038/nature21031

Zhang, I. H., Borer, B., Zhao, R., Wilbert, S., Newman, D. K., and Babbin, A. R. (2024). Uncultivated DPANN archaea are ubiquitous inhabitants of global oxygen-deficient zones with diverse metabolic potential. *MBio* 15, e02918–e02923. doi: 10.1128/mbio.02918-23

Zhang, Y., and Lin, K. (2012). A phylogenomic analysis of *Escherichia coli*/Shigella group: implications of genomic features associated with pathogenicity and ecological adaptation. *BMC Evol. Biol.* 12:174. doi: 10.1186/1471-2148-12-174

Zhang, C., Liu, X., Shi, L. D., Li, J., Xiao, X., Shao, Z., et al. (2023). Unexpected genetic and microbial diversity for arsenic cycling in deep sea cold seep sediments. *NPJ Biofilms Microbio.* 9:13. doi: 10.1038/s41522-023-00382-8

Zhang, R. Y., Wang, Y. R., Liu, R. L., Rhee, S. K., Zhao, G. P., and Quan, Z. X. (2024). Metagenomic characterization of a novel non-ammonia-oxidizing Thaumarchaeota from hadal sediment. *Microbiome* 12:7. doi: 10.1186/s40168-023-01728-2

Zhang, X., Zhang, C., Liu, Y., Zhang, R., and Li, M. (2023). Non-negligible roles of archaea in coastal carbon biogeochemical cycling. *Trends Microbiol.* 31, 586–600. doi: 10.1016/j.tim.2022.11.008

Zillig, W., Tu, J., and Holz, I. (1981). Thermoproteales—a third order of thermoacidophilic archaebacteria. *Nature* 293, 85–86. doi: 10.1038/293085a0