OPEN ACCESS

Microbiology ARTICLE

The Innovation Life

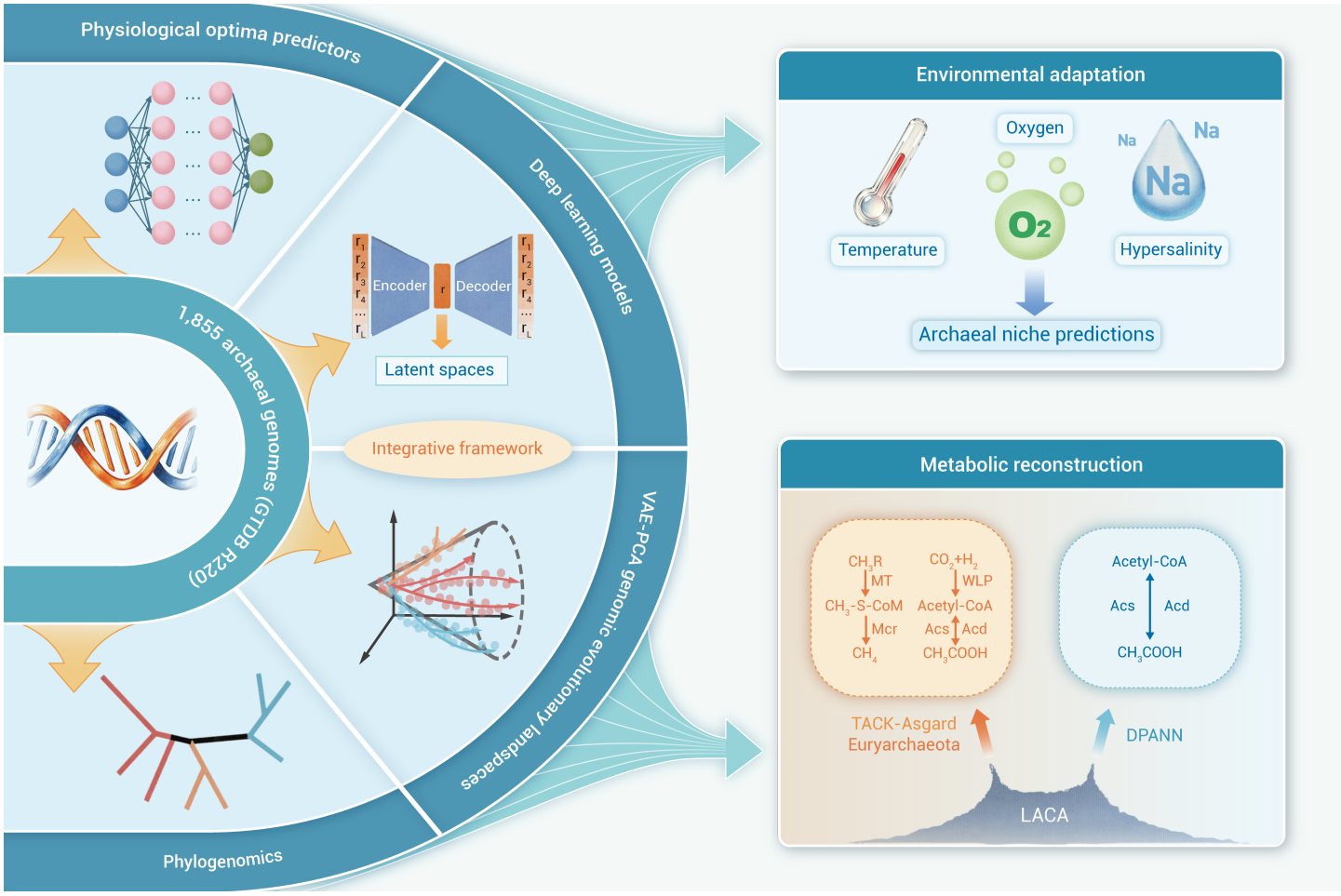
Radiating pattern revealed by a deep learning model traces the evolutionary dynamics of the Archaea domain

Zhenbo Lv,[1,2] Fengping Wang,[3] and Yinzhao Wang[1,2,*]

*Correspondence: wyz@sjtu.edu.cn

Received: March 25, 2026; Accepted: June 29, 2026; Published Online: July 6, 2026; https://doi.org/10.59717/j.xinn-life.2026.100235

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Early diversification of archaea remains unsettled due to uncertainty from conventional phylogenomics.
- By integrating deep learning and traditional phylogenomics, we revealed archaeal evolutionary patterns.
- Early diversification of DPANN potentially linked to acetate metabolism in lower temperature environments.
- The other archaea might share an ancestor with methyl-reducing methanogenesis and carbon fixation.
- From domain-wide radiating patterns to genus-level adaptation, our framework offers comprehensive insights.

ARTICLE Microbiology

OPEN ACCESS

Radiating pattern revealed by a deep learning model traces the evolutionary dynamics of the Archaea domain

Zhenbo Lv,[1,2] Fengping Wang,[3] and Yinzhao Wang[1,2,*]

[1]State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

[2]Yazhou Bay Institute of Deepsea Science and Technology, Shanghai Jiao Tong University, Sanya 572024, China

[3]School of Oceanography, Shanghai Jiao Tong University, Shanghai 200240, China

*Correspondence: wyz@sjtu.edu.cn

Received: March 25, 2026; Accepted: June 29, 2026; Published Online: July 6, 2026; https://doi.org/10.59717/j.xinn-life.2026.100235

Citation: Lv Z., Wang F. and Wang Y. (2026). Radiating pattern revealed by a deep learning model traces the evolutionary dynamics of the Archaea domain. *The Innovation Life* **4**:100235.

Archaea, one of the primary domains of life, have coevolved with Earth for billions of years and play critical roles in biogeochemical cycles across diverse ecosystems. However, resolving deep archaeal phylogeny is confounded by among-site heterogeneity and compositional biases, which together exacerbate long-branch attraction in lineages like DPANN. Here, using variational autoencoders and other deep learning models along with phylogenomics, we re-evaluated archaeal evolution and identified a radiating evolutionary landscape of three major archaeal superphyla, DPANN, TACK-Asgard and Euryarchaeota. Our integrated results also support the scenario in which DPANN evolved as an early monophyletic lineage, potentially relying on acetate-related metabolism under lower temperature conditions, whereas TACK-Asgard and Euryarchaeota possibly share a hydrogen-dependent methyl-reducing methanogen ancestor with the complete Wood–Ljungdahl pathway in higher temperature habitats. Environmental factors, including temperature, oxygen, and salinity, have kept influencing subsequent archaeal evolution in distinct patterns, highlighting the interplay between environmental factors and evolutionary trajectories of this ancient domain. By demonstrating that deep learning can be systematically integrated with phylogenomics to generate testable hypotheses on ancient metabolisms and environmental adaptation, this work provides a new analytical insight for archaeal evolutionary studies.

INTRODUCTION

Archaea are one of the fundamental domains of life alongside bacteria[1-3] and have coevolved with Earth environments for approximately four billion years.[4-6] The origin and evolution of archaea throughout geological history have long been discussed, debated and renewed because of the rapid discovery of new archaeal species[7-12] and novel evolutionary tools.[13-19] Archaea taxonomy has been expanded from the Euryarchaeota and Crenarchaeota phyla to at least 20 phyla, approximately four superphyla, Euryarchaeota (comprising eight phyla in GTDB R220),[12] TACK (Thermoproteota),[20] Asgard (Asgardarchaeota), and DPANN (comprising eleven phyla in GTDB R220).[21] Estimating the metabolic repertoire of the last archaeal common ancestor (LACA) and resolving the deep phylogenetic relationships among these superphyla remain interdependent and contentious challenges. One phylogenomic study[22] inferred the evolutionary history of each gene in four archaeal superphyla through gene duplication, transfer, and loss analysis. They placed the archaeal tree root between DPANN and other archaea and predicted that the LACA was an anaerobe that may have been able to fix CO_2 to acetyl-coenzyme A (acetyl-CoA) via the Wood–Ljungdahl pathway (WLP) and subsequently generate acetate and ATP using an acetyl-CoA ligase (Acd, ADP-forming).[22] However, another study argued that this rooting may reflect the fast evolutionary rate of DPANN and the artifacts of long-branch attraction (LBA).[23] They proposed that the monophyletic DPANN archaea probably evolved from a free-living, Euryarchaeota-like ancestor.[23] Therefore, based on their species trees, LACA corresponds to the last common ancestor of Euryarchaeota and TACK-Asgard. Moreover, an integrated phylogenetic investigation of methanogenesis-related proteins revealed that LACA was a versatile methanogen capable of both reducing CO_2 and methylated compounds and that the inheritance/loss/gain/innovation of methanogenesis drove early archaeal diversification.[24] These competing reconstructions illustrate that the placement of DPANN and the metabolic identity of LACA are tightly coupled, yet conventional phylogenomic approaches, reliant on site-independent substitution models, yield conflicting topologies that are highly sensitive to sampling and alignment strategies.

The abovementioned studies suggested that methanogenesis, complete WLP, and acetate-related metabolism might have existed in the LACA. For methanogenesis, it is primarily categorized into four types based on different substrates: methyl-reducing hydrogenotrophic (H_2/CH_3R), CO_2-reducing hydrogenotrophic (H_2/CO_2), methyl-dismutating, and acetoclastic.[25] The methyl-coenzyme M reductase complex (Mcr) is the key enzyme for all these methanogenesis processes, as Mcr catalyzes the last step of methanogenesis, reducing methyl-coenzyme M to CH_4. Traditionally, methanogenesis is linked to the tetrahydromethanopterin (H_4MPT) methyl branch of the WLP, which reduces CO_2 to the methyl group of CH_3-H_4MPT,[26-28] serving as a substrate for H_2/CO_2 methanogenesis.[25,29,30] In the H_4MPT branch, H_2 acts as an electron donor, and CO_2 is progressively reduced to a methyl group through catalysis by following enzymes: the tungsten or molybdenum formylmethanofuran dehydrogenase complex (Fwd), the formylmethanofuran: H_4MPT formyltransferase (Ftr), the methenyl-H4MPT cyclohydrolase (Mch), F420-dependent methylene (CH_2)-H_4MPT dehydrogenase (Mtd) or the bacterial methylene-H4MPT dehydrogenase (bMtd) or H_2-forming methylene-H_4MPT dehydrogenase (Hmd) and the methylene-H_4MPT reductase (Mer).[26-28] The carbon monoxide dehydrogenase/acetyl-CoA synthase (Cdh) complex catalyzes reactions in the carbonyl branch of the WLP, and its four main subunits (CdhACDE) are widely present in both archaea and bacteria.[27,28,31] In archaea, CdhA/B reduces CO_2 to CO, and CdhCDE converts CO and the methyl group of CH_3-H_4MPT into an acetyl group to synthesize acetyl-CoA.[27,28,31] One study reported that the complete WLP might be used as an electron sink in acetate-related metabolism in the deeply branching *Korarchaeia*[32] and that these archaea can convert acetyl-CoA to acetate to generate ATP via Acd or acetate kinase (Ack) and phosphate acetyltransferase (Pta).[32-36] Moreover, another study focused on acetate-related metabolism, including Acd and acetyl-CoA synthase (Acs, AMP-forming), in methane-oxidizing archaea,[37] and cultivation experiments have demonstrated that some methane-oxidizing archaea can utilize acetate for acetoclastic methanogenesis through Acs.[37] Collectively, there are intricate connections between methanogenesis, WLP, and acetate-related metabolism, and these metabolic processes may drive the early evolution of the archaeal domain. While these metabolic capacities have been extensively investigated, their ancestral presence and the evolutionary order of their emergence remain actively debated, motivating the development of complementary analytical frameworks.

Recent expansions in publicly available protein sequence databases, coupled with advances in unsupervised generative modeling,[17-19] have enabled us to probe life evolution at unprecedented depth using variational autoencoders (VAEs) and other algorithms, offering methods to address the limitations of conventional phylogenomics. By feeding a full-length multiple sequence alignment (MSA) of a gene or protein into a VAE model, the encoder of the VAE compresses high-dimensional sequence embeddings into a set of latent variables governed by an approximate Gaussian prior, while the decoder of the VAE reconstructs the original MSA within a Bayesian framework.[17-19,38,39] This captures higher-order sequence dependencies, including epistatic interactions among residues, that site-independent models treat as noise and yields a continuous, low-dimensional latent space. In the resulting VAE protein evolutionary landscape, sequences arrange in a star-like pattern

around the coordinate origin (inferred ancestral state) and radiate outward along continuous evolutionary trajectories, providing a heuristic proxy for relative divergence that is orthogonal to tree-based branch lengths.[17-19] Overlaying this landscape onto conventional phylogenomic trees refines branching orders and pinpoints residues under selective pressure.[17,18] Based on the evolutionary landscapes of conserved proteins, VAE-PCA genomic evolutionary landscapes proposed in this study can generate conical radiating patterns, and crucially, this framework does not assume site independence. Instead, it models each MSA as a joint distribution, potentially mitigating LBA artifacts when rapidly evolving lineages are involved. Beyond life evolution, other deep learning models are trained on whole-genome and proteome features to estimate species-level physiological optima, such as growth temperature,[40-44] pH,[40,44,45] salinity[40,44,46] and oxygen tolerance,[40,44,47,48] and project these traits onto the species tree and the latent manifold. The resulting multiscale framework links shifts in molecular ancestry to paleoenvironmental variables, illuminating how proteome and metabolic networks adapt to fluctuating ancient habitats and offering concrete hypotheses about the ecological niche of ancestral organisms.[11,22,49,50] Importantly, this multiscale framework generates computationally testable hypotheses that are intended to complement, rather than replace, conventional phylogenetic inference.

In this study, archaeal evolutionary relationships were explored using genomes from the Genome Taxonomy Database (GTDB R220), the potential metabolic traits of archaeal ancestors were inferred, and their key evolutionary processes were depicted, by integrating a VAE-related framework with phylogenomics to generate complementary, data-driven evolutionary landscapes. Our results are consistent with the potential evolutionary scenario in which DPANN forms a monophyletic group that diverges early, relying on acetate-related metabolism in lower temperature environments, in contrast to the ancestor of Euryarchaeota and TACK-Asgard possibly living in higher temperature habitats. Metabolic pathway reconstructions suggest that the H_4MPT branch of the WLP might have emerged from the common ancestor of Euryarchaeota and TACK-Asgard. H_2/CH_3R methanogenesis represents a plausible earliest form, and the origin and duplication of the Mtr complex in Euryarchaeota potentially contributed to multiple Mtr-dependent methanogenesis diversification events. Other deep learning models were integrated to predict the environmental adaptations of each archaeal genus, revealing that hypersalinity, temperature and oxygen are major variables associated with archaeal evolution. This work thereby provides a reusable analytical framework that integrates deep learning-based methods and phylogenomics to offer a complementary angle and supportive evidence for understanding archaeal origin, metabolism, and adaptation, including insights into the placement of DPANN and the nature of early archaeal metabolisms and environmental adaptation.

MATERIALS AND METHODS

Collection of archaeal genomes

A total of 1847 archaeal genera were verified based on the Genome Taxonomy Database (GTDB R220, https://data.gtdb.ecogenomic.org/releases/release220/220.0/). For each archaeal genus, one genome was chosen as a representative for each genus and their genome information was collected. Additionally, eight genomes from methanogens were added, making the 1,855 archaeal genomes set (dataset1: arc1855). Moreover, all genomes of *Nezhaarchaeales*, *Methanonatronarchaeia* and *Halarchaeoplasmatales* from GTDB R220 were collected for further research. Whole-genome protein sequences of each genome were predicted by Prodigal (Version 2.6.3).[51] To further reduce the influence of genome quality on metabolic pathway reconstruction results, 603 archaeal families were additionally verified based on the GTDB database R226. For each archaeal family, one genome with the highest quality score (completeness - 5 × contamination) was selected as a representative, to test the completeness and effects of the selected pathway.

Phylogenomic and phylogenetic analyses

Conserved genes and key metabolism-related proteins were identified by Blastp (version 2.12.0+)[52] with "-p 40 -e 1e-5 --query-cover 50 -f 6" and in a genome, each conserved gene with the best Blast hit was preserved. Protein sequences were aligned by the rapid multiple sequence alignment program MAFFT (version 7.526)[53] with "--ep 0 --genafpair --maxiterate 1000 --thread -20". Their alignments were manually checked to remove false positives and particularly incomplete sequences, and then filtered by trimAl software (version 1.5.rev0)[54] with "-automated1". For the construction of species trees, the MSAs of a set of conserved genes were concatenated into a single concatenated alignment and the species trees were constructed using FastTree software (version 2.1)[55] with "-wag". A total of 169 genomes for all archaeal orders and 343 genomes for all non-DPANN archaeal families and DPANN orders were randomly selected from the 1855 genomes. These two species trees and key metabolism-related proteins trees were constructed using either FastTree with "-wag" or IQ-TREE (version 1.6.2)[56] with "-alrt 1000 -bb 1000 -nt 40" for model selection.[57] For IQ-TREE analyses, the "C60" mixture model was incorporated into the best-fit model to construct the final tree (details in Table S1). iTOL (http://itol.embl.de/release220/220.0/) was used to modify the phylogenomic trees.

Construction of VAE protein evolutionary landscapes and VAE-PCA genomic evolutionary landscapes

Based on the full-length MSA of a gene, VAE protein evolutionary landscapes were constructed with default parameters (the random seed is 123).[17] The Euclidean distance of each protein sequence to the coordinate origin in the VAE landscape was calculated. The VAE protein evolutionary score of the protein sequence farthest from the coordinate origin was set to 1, and normalization was performed to obtain the VAE protein evolutionary scores of all protein sequences. The VAE protein evolutionary scores of a set of conserved genes in a genome were used as embeddings to describe the relative divergent distance of this genome. If a conserved gene was not present in some genomes, the average VAE protein evolutionary score of that conserved gene from other genomes was used to fill the missing value in these genomes. Principal Component Analysis (PCA) was employed to extract the three principal components of the embeddings of all genomes. These components were then visualized to construct the VAE-PCA genomic evolutionary landscape. The Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) and parallel analysis of PCA were tested using Python scripts. If the VAE-PCA evolutionary landscape presented a conical shape, the genome closest to the apex of the cone in the evolutionary landscape would be set as the starting point, and the Euclidean distance of other genomes to the starting point could be calculated. The VAE-PCA genomic evolutionary score of the genome farthest from the starting point was set to 1, and normalization was performed to obtain the VAE-PCA evolutionary scores of all genomes. Four additional random seeds (42, 710, 1011, 1115) were selected to validate the robustness of the VAE protein evolutionary landscape and the VAE-PCA genomic evolutionary landscapes by calculating the coefficient of variation of VAE scores for protein sequences and VAE-PCA scores for genomes across different random seeds. Python scripts for the above VAE-PCA framework are available at https://doi.org/10.6084/m9.figshare.31849165.

Protein structure prediction and alignment

Protein structures were predicted using ColabFold[58] with "--num-recycle 5", and the top-ranked result was used as the final prediction. Protein structures were aligned using USalign[59] with default parameters. The alignment results were visualized using PyMOL (http://www.pymol.org/pymol).

Prediction of optimal conditions for archaea

The oxygen tolerance for archaea was predicted by Aerobicity[47] and Aerobot.[48] For a given archaeal genome, if both software provided consistent prediction results, the archaea was labeled "oxygen-tolerant" or "oxygen-intolerant", while if the prediction results were inconsistent, the archaea was labeled "inconsistent". The optimum growth temperature for archaea was predicted by Tome.[42] The possibility of a protein sequence to be halophilic was predicted by HPClas.[46] If the possibility was greater than 0.9, the protein was considered a predicted halophilic protein, and if it was less than 0.1, the protein was considered a predicted nonhalophilic protein. If the proportion of predicted halophilic proteins in a genome was more than 38%, it was inferred that the archaea could tolerate hypersalinity environments.

RESULTS AND DISCUSSIONS

Phylogenomic tree and VAE-PCA evolutionary landscape of archaea

The VAE evolutionary landscape method[17-19] coupled with phylogenomic

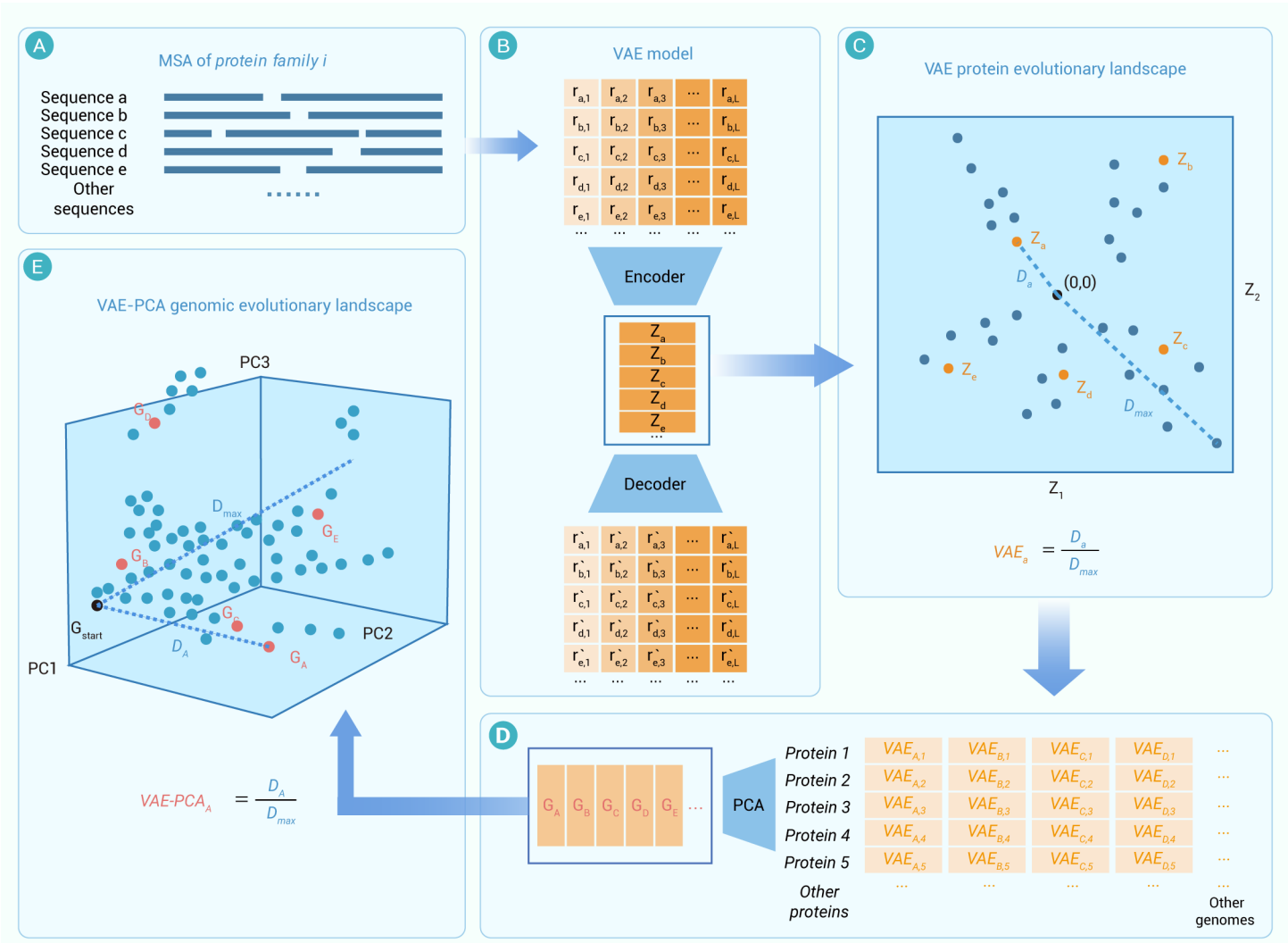


Figure 1. Workflow for constructing a VAE-PCA evolutionary landscape (A) The multiple sequence alignment (MSA) of a conserved gene is used as an input for the VAE model. (B) One-hot encoding is applied to all amino acids in this MSA, converting the sequences into sequence embeddings of the same length. The VAE model is trained by compressing the high-dimensional sequence embeddings into 2-dimensional embeddings (Z_i) and reconstructing the MSA based on Z_i repetitively. (C) The Z_i of sequences in MSA from the well-trained VAE are projected to construct a VAE protein evolutionary landscape. The VAE evolutionary score of a protein sequence is calculated based on the Euclidean distance of its projection point to the coordinate origin. (D) VAE landscapes of a set of conserved genes from genomes are constructed, and the VAE evolutionary score for each conserved gene in each genome is calculated. The VAE evolutionary scores of a set of conserved genes in each genome are used as embeddings to explore patterns potentially related to the relative divergent distance of these genomes. Principal component analysis (PCA) is performed on the embeddings of all genomes and three principal components (PCs) are extracted. (E) The three PCs of each genome embedding are projected to construct a VAE-PCA evolutionary landscape, which typically appears as a conical shape. The VAE-PCA evolutionary scores are calculated based on the Euclidean distance from projection points of genomes to the projection point of the genome closest to the conical vertex.

analysis was used to investigate the origin and evolution of archaea following a custom procedure (Figures 1A-E). Through the VAE model, based on the MSAs of a set of conserved genes, their VAE evolutionary landscapes were constructed (Figures 1A-C). The VAE evolutionary score of a protein sequence, calculated based on the Euclidean distance from it to the coordinate origin in the VAE landscape, provides a quantitative descriptor in latent space that we interpret as potentially reflecting relative divergent distance (Figure 1C). The closer the sequence is to the coordinate origin, the earlier it is likely to have diverged under this model assumption.[17-19] Therefore, the relative divergent distance of a set of conserved genes in a genome could be used as an embedding to characterize the relative divergent distance of this genome (Figure 1D). However, this representation is used as a heuristic framework for exploring evolutionary relationships, not as a direct measure of absolute divergent distance. PCA was performed on these embeddings of all the genomes, and the results were visualized to construct a VAE-PCA evolutionary landscape (Figures 1D-E). Other dimensionality reduction methods, including Uniform Manifold Approximation and Projection (UMAP) and T-distributed Stochastic Neighbor Embedding (*t*-SNE), were also attempted for visualizing these embeddings. Nevertheless, because the linear combination of relative divergent distances of conserved genes may provide a better description of the relative divergent distances of genomes, the main PCs represent the major axes of variation of these embeddings and VAE-PCA landscapes may capture conical evolutionary patterns, with genomes closer to the vertex of the cone potentially representing earlier-diverging lineages. The VAE-PCA evolutionary score of a genome, calculated based on the Euclidean distance from it to the genome closest to the vertex of the cone in the VAE-PCA landscape, is used here as an exploratory proxy that may reflect the relative divergent distance of this genome (Figure 1E).

To investigate the origin, evolution, metabolism, and environmental adaptation of all archaea, representative genomes with high completeness and low contamination (Figure 2A) for all 1,847 archaeal genera from GTDB R220[12] as well as eight additional archaeal genera were collected (Table S2). The 1,855 archaeal genomes set (dataset1: arc1855) included all the genera from the four archaeal superphyla, and their genome size, GC content, and amino acid utilization were extensive and evenly distributed (Figures 2A-B), allowing unbiased analysis of the origin, evolution, and metabolism of major archaeal groups. These genomic features are associated with evolutionary trajectories and served as indicators of metabolic diversity and environmental

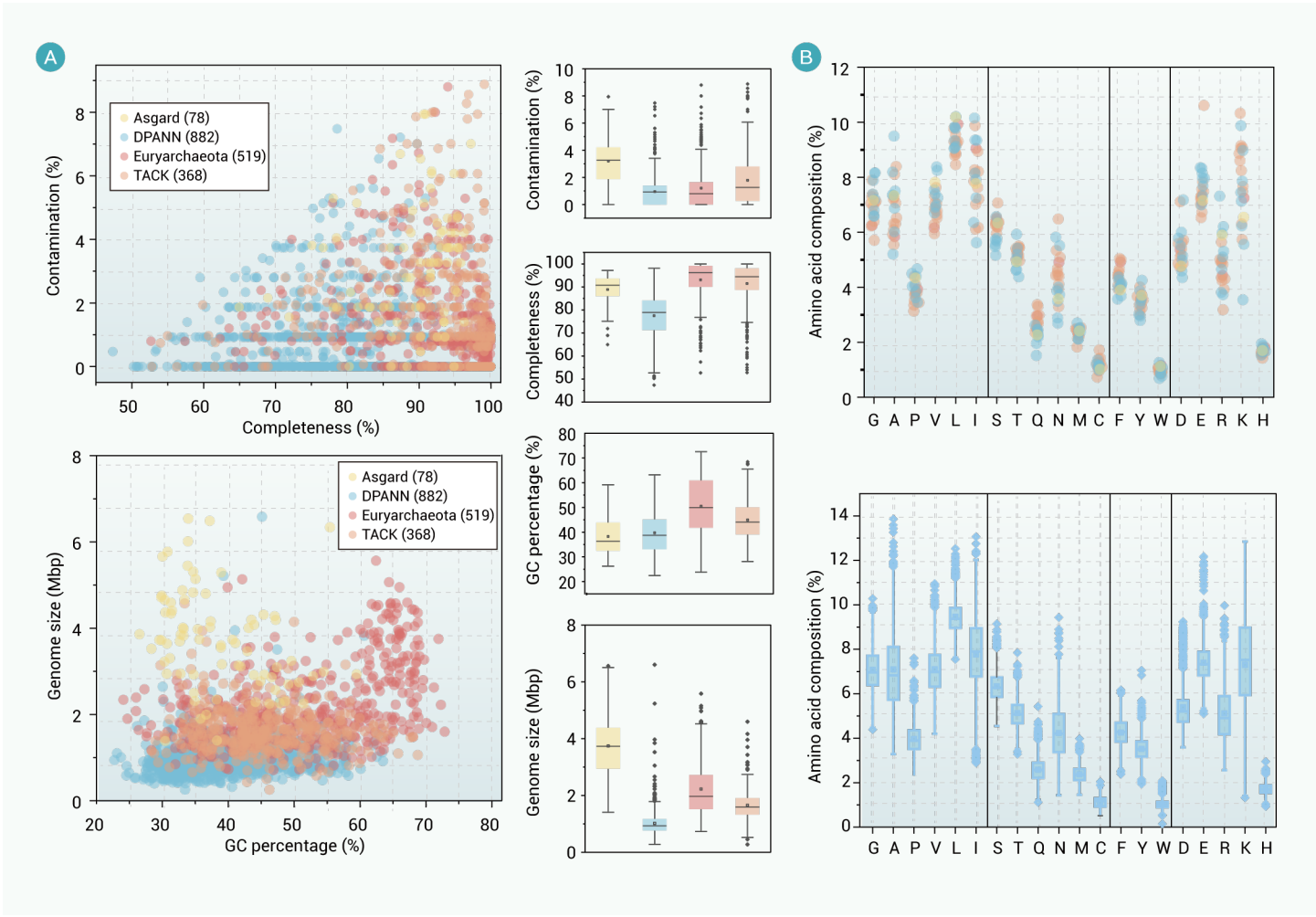


Figure 2. Genome information of the 1855 archaeal genera (A) Information of contamination, completeness, genome size and GC content, for representative genomes of 1847 archaeal genera from GTDB R220. Blue represents genomes from DPANN, red represents Euryarchaeota, orange represents TACK, and yellow represents Asgard (with the number of genomes in parentheses). The four box plots on the right show the genome contamination, genome completeness, GC content, and genome size for four archaeal superphyla. (B) The upper panel shows the average utilization of amino acids in archaeal genomes from different phyla, with blue for DPANN, red for Euryarchaeota, orange for TACK, and yellow for Asgard. The lower panel shows box plots of amino acid utilization in each archaeal genome.

adaptation, since differences in genome size, GC content, and amino acid utilization among archaeal groups are associated with significant variations in their metabolic strategies and ecological niches (Figures 2A-B), consistent with adaptation to diverse and even extreme environments over evolutionary time.

Traditional phylogenomic tree of archaea. Species trees for arc1855 based on concatenated sets of 16, 27 and 37 conserved genes (Table S3)[60-62] were constructed (Figures 3A & S1A). Consistent with previous phylogenomic analyses,[22,23] all three trees indicated that DPANN formed a monophyletic group, and arc1855 can be divided into three major supergroups, DPANN, TACK-Asgard, and Euryarchaeota, suggesting that the Asgard and TACK lineages are closely related early-diverging archaeal clades (Figures 3A & S1A). After species trees for 169 representative genomes from 169 archaeal orders (dataset2: arc169), as well as 343 representative genomes from 53 DPANN orders and 290 other archaeal families (dataset3: arc343) based on 37 conserved genes, were constructed, similar results were obtained (Figure S1B). In both trees, DPANN consistently formed a monophyletic group, whereas Asgard formed a monophyletic group but originated within TACK (Figure S1B). Interestingly, in the recent phylogenomic tree constructed by Baker et al.,[23] the root of the archaeal tree was placed between TACK-Asgard superclade and other archaeal superphyla, with DPANN forming a monophyletic group that originated within Euryarchaeota. If the same out-group rooting method was used, this topology shows similarities to the trees of arc169 based on 37 conserved genes and arc1855 based on 27 conserved genes in our study (Figures S1B-C), suggesting a certain consistency in the placement of DPANN and its evolutionary relationship with Euryarchaeota. More recently, phylogenetic reconciliation analyses have likewise recovered DPANN as a monophyletic sister group to TACK and Asgard, converging on a root near the base of the Euryarchaeota and supporting a complex, methanogenic LACA.[63] These convergent inferences, despite differing root placements, provide an alternative view on the nesting of DPANN within Euryarchaeota.

These recent studies highlight the importance of rooting the archaeal tree and provide a framework for understanding the evolutionary relationships among major archaeal superphyla.[23] These findings also underscore the challenges associated with traditional phylogenomic tree construction, which is highly sensitive to factors such as genome sampling and alignment strategies, leading to inconsistencies in resolving relationships between distantly related archaeal groups. Moreover, traditional phylogenomic methods assume that all sites in a MSA evolve independently, an assumption that fails to account for epistatic interactions between sites, which are now recognized as critical drivers of protein evolution.[64-67] This limitation becomes particularly problematic when rapidly evolving lineages such as DPANN, where conserved genes may also accumulate saturated mutations, eroding phylogenomic signals, are studied.[68-71] Such saturation can lead to long-branch attraction,[72-74] where unrelated lineages cluster together due to convergent evolutionary rates rather than shared ancestors. The placement of DPANN within Euryarchaeota raises important questions about whether this topology reflects true evolutionary relationships or methodological biases.[23] These challenges highlight the need for other phylogeny-independent approaches or more advanced phylogenomic methods that incorporate site interdependencies and account for the complexities of molecular evolu-

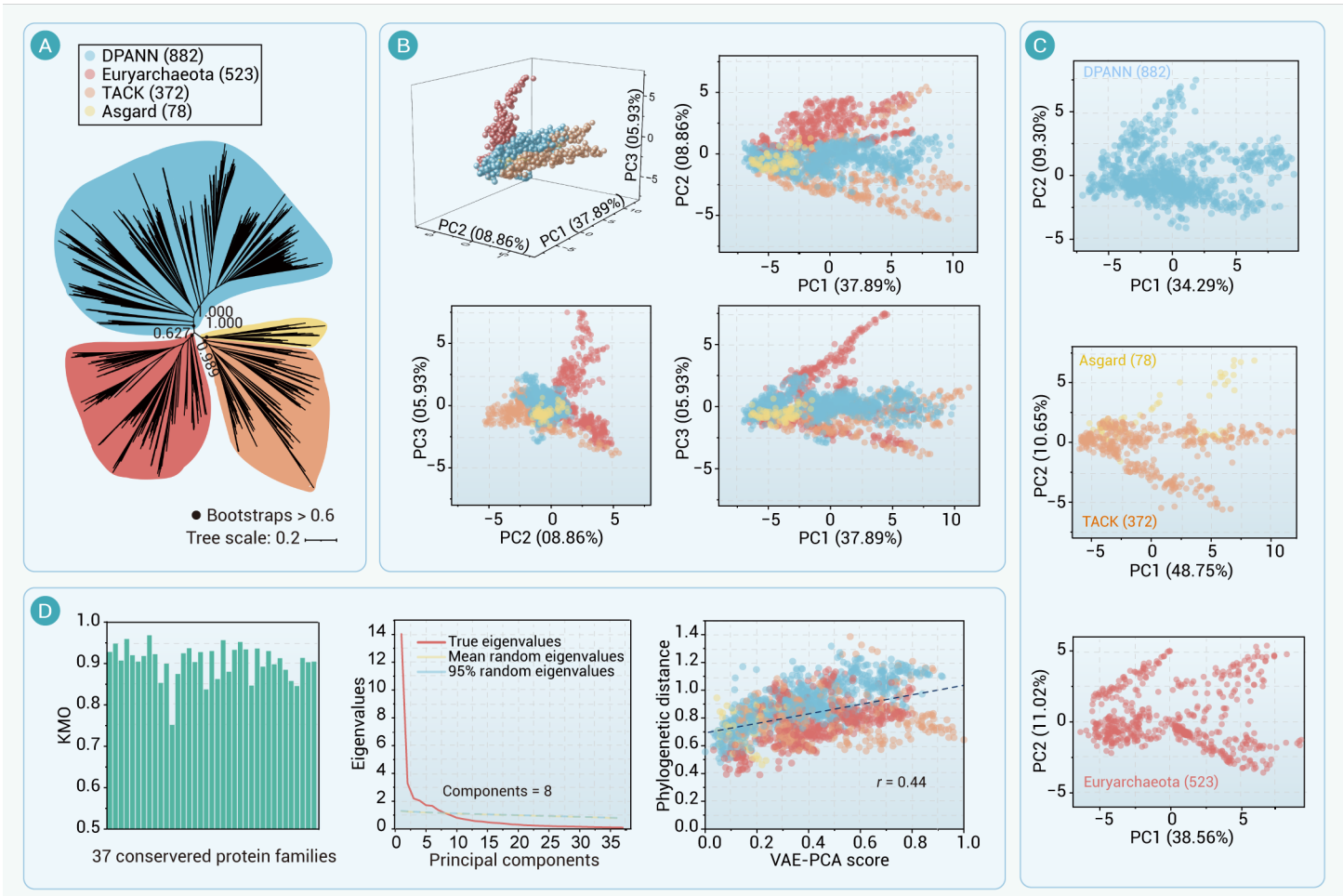


Figure 3. Phylogenomic tree and VAE-PCA evolutionary landscapes of 1855 archaeal genomes (A) The species tree of 1855 archaeal genomes based on the concatenated MSA of 37 conserved genes, is rooted between DPANN and other archaeal phyla. Blue represents DPANN (882 genomes), yellow represents Asgard (78 genomes), orange represents TACK (372 genomes), and red represents Euryarchaeota (523 genomes). (B) The VAE-PCA evolutionary landscape of 1855 archaeal genomes was constructed based on the VAE evolutionary scores of 37 conserved genes. And three-view drawing of this VAE-PCA evolutionary landscape is displayed. (C) VAE-PCA evolutionary landscapes were constructed for genomes only from DPANN, TACK-Asgard, and Euryarchaeota, respectively. (D) The results of the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) for the VAE protein evolutionary scores of 37 conserved genes in 1,855 genomes show that all the conserved genes are suitable for PCA, with COG0091 having the smallest KMO value of 0.7516. Additionally, 1,000 parallel analyses were performed to determine the number of principal components to be retained. The correlation between the VAE-PCA evolutionary scores and the branch length distances from the root on the species tree was calculated, and their Pearson's correlation coefficient is 0.44.

tion, particularly when reconstructing the deep evolutionary history of archaeal superphyla.

The evolutionary radiating pattern of archaea via VAE-PCA. To address these issues, we attempted to introduce VAE to assist in the analysis of the origin and evolution of archaea.[17-19] VAE evolutionary landscapes were constructed for each conserved gene in the above phylogenomic analysis, and the VAE evolutionary scores for protein sequences were calculated (Figures 1A-C). The VAE evolutionary scores of 16, 27 or 37 conserved genes (Table S3) in genomes were used as embeddings to describe the relative divergent distance of arc1855 by constructing the VAE-PCA evolutionary landscapes, and the evolutionary trajectories in each VAE-PCA landscape exhibited a conical-radiating shape (Figures 3B & S1A). Similar evolutionary trajectories were obtained using other sets of conserved genes, such as sets of 48 and 126 conserved genes in the 321 genomes from the study of Baker et al. (Figure S1C).[23] The construction of VAE-PCA landscapes using 37 conserved genes from only DPANN, TACK-Asgard, or Euryarchaeota yielded similar evolutionary landscapes that exhibited a conical-radiating shape (Figures 3C & S1D).

In the VAE-PCA evolutionary landscape based on the 37 conserved genes of arc1855, the genomes at the base of the cone are distributed along eight evolutionary trajectories extending at different angles (Figure 3B), and the genomes on each of them belong to the same phylum, including *Thermoplasmatota*, *Halobacteria* (splitting into two evolutionary trajectories), and *Methanomicrobia* of Euryarchaeota, *Nanoarchaeia* of DPANN, and *Bathyarchaeia*, *Thermoprotei*, and *Nitrososphaeria* of TACK, suggesting that these archaea may diverge later in evolution. In contrast, the genomes at the top of the cone are relatively densely clustered, including archaea from four superphyla, and are difficult to distinguish in terms of evolutionary trajectories (Figure 3B), suggesting that these lineages possibly represent earlier-diverging taxa. Therefore, the genome closest to the apex of the cone in the VAE-PCA landscape was tentatively designated as the possible evolutionary starting point (G_{start}), and the distances of all the other genomes from G_{start} were calculated and normalized to serve as the VAE-PCA evolutionary scores for each genome, as indicated in Figure 1E. To evaluate the reproducibility of this analytical framework, we conducted repeated computations employing four additional random seeds (see Materials and Methods), and subsequently calculated the coefficients of variation of VAE scores for protein sequences and VAE-PCA scores for genomes across five seeds. Across the 1,855 genomes analyzed using 37 conserved genes, the median CV for VAE-PCA scores was 6.1%, with the vast majority of genomes exhibiting CVs under 30%. These results indicate that the landscape construction remains stable despite variations in initial randomization parameters. Within a cluster on the species tree of arc1855 based on 37 conserved genes, genomes with smaller VAE-PCA scores are typically in deeper branches closer to the root, whereas those with larger VAE-PCA scores are often found in branches away from the root. The Pearson correlation coefficient between their VAE-PCA scores and branch lengths to the root in the species tree is 0.44, indicating a certain degree of consistency in the inferred evolutionary relationships of archaea

using these two methods (Figure 3D). There are 212 archaeal genomes closest to G_{start} (with VAE-PCA values less than 0.15) from 11 phyla of all three superclades (Figure 3B), suggesting that DPANN, TACK-Asgard, and Euryarchaeota might have originated rapidly after LACA. This pattern is consistent with the hypothesis that DPANN is a deeply branching, early-diverging lineage, supported by the DPANN-root hypothesis advanced by Williams et al.[22] and the recent reconciliation analyses recovering DPANN as the sister to TACK and Asgard.[63] However, these deep learning-based results do not resolve the placement of DPANN; rather, they are presented here as providing an additional complementary perspective that extends beyond traditional phylogenomics.

It is important to clarify what the VAE-PCA landscape adds beyond traditional phylogenomic approaches. Conventional species trees rely on concatenated MSAs analyzed under site-independent models; they yield discrete branching orders highly sensitive to outgroup choice, alignment trimming, and LBA. In contrast, the VAE-PCA landscape treats each genome as a point in a continuous latent space derived from per-gene VAE evolutionary scores. The moderate correlation between VAE-PCA scores and root-to-tip branch lengths (r = 0.44) indicates that the two methods share an underlying evolutionary signal (Figure 3D), yet the discrepancies, particularly for deep-branching lineages, highlight regions where site-independent models and latent-space embeddings diverge. These discrepancies likely arise because VAEs implicitly model residue covariation including epistatic interactions between sites and capture the information of back-mutations from MSAs, whereas phylogenomic likelihoods treat recurrent substitutions as homoplasy. A distinctive advantage of the VAE-PCA genomic landscape is its capacity to infer the relative divergent distance of genomic differentiation from a continuous, data-driven perspective beyond discrete tree topology. In conventional phylogenomics, the relative order of lineage divergence is inferred indirectly through root-to-tip branch lengths under specific substitution models and molecular-clock assumptions, which can be distorted by rate heterogeneity or long-branch attraction. By contrast, the VAE-PCA evolutionary score provides a direct, genome-wide quantitative measure of latent-space distance from the inferred ancestral origin (the apex of the cone). This allows genomes to be ranked along a trajectory of relative divergence, offering an independent heuristic to distinguish earlier-diverging lineages (those closer to the origin) from later-diverging ones (those farther along the radiating trajectories). Consequently, the VAE-PCA landscape should be viewed as a complementary heuristic that identifies broad radiating trends and flags lineages warranting closer scrutiny, rather than as a replacement for phylogenetic inference.

Evolution of potential LACA and its descendants

As previously mentioned, Williams et al. hypothesized that the complete Wood–Ljungdahl pathway (WLP) and acetyl-CoA ligase (Acd, ADP-forming) of acetate-related metabolism were potentially present in LACA, whereas the methyl-coenzyme M reductase complex (Mcr), as the key enzyme in methanogenesis, could only be traced to the common ancestor of Euryarchaeota and TACK.[22,30] However, a study by Mei et al. based on phylogenetic analysis of enzymes related to methanogenesis and the results of autocatalytic experiments using the cofactor F_{430} proposed that methanogenesis may have begun from primordial protein-free F_{430}-driven reactions and that the origin of methanogenesis coincided with LACA and that LACA was capable of reducing both CO_2 and CH_3OH to CH_4.[24] Therefore, we identified key proteins involved in acetate-related enzymes, WLP, methanogenesis, as well as other key metabolisms from arc1855 to explore the evolution of potential LACA and its descendants (Figures 4A-C). In addition, to mitigate the potential effects of genome completeness, we further surveyed these pathways across a different dataset constructed from the highest-quality genomes available in GTDB R226, selecting one representative per archaeal family based on quality scores. Among the 603 families thus selected, the integrity of the relevant complexes and metabolic pathways was evaluated, and the resulting patterns were consistent with those observed in arc1855 (Table S4).

Acetate-related metabolism. Through the key proteins of acetate-related metabolism, including Acd, Acs, and the combination of Ack and Pta,[32-37,75] archaea can consume ATP to convert acetate into acetyl-CoA, which serves as a substrate for subsequent biomass synthesis; alternatively, they can convert excess acetyl-CoA into acetate to obtain ATP, fulfilling their energy metabolism needs.[33-37,76] As in a previous study,[34,35,76] structural comparisons revealed that Acd is composed of α and β subunits (Acd α and Acd β) and that they are not necessarily in a fixed order (Figure S2). Acd α and Acd β can also exist independently; therefore, sequences of independent Acd α and β, as well as fusion Acd with both orders (Acd α+β and Acd β+α), were separately identified in arc1855. Because Acd α is necessary for catalytic activity, whereas the β subunit may aid in specific binding to acetate,[35,36] if a genome contains Acd α, it is inferred that the archaeal genome has the potential for archaeal-type acetate-related metabolism (Table S5). The sequences of Acs, Ack, and Pta were also separately identified in arc1855. If an archaeal genome possesses both Ack and Pta, it is inferred to have the potential for bacterial-type acetate-related metabolism (Table S5); similarly, if an archaeal genome contains Acs, it is considered to have universal acetate-related metabolism (Table S5).

Acd is distributed mainly in archaea,[37,76] and 841 genomes of arc1855 have independent Acd α sequences, 17 of which contain more than 10 copies with the most sequence copies at 25 (GCA_019058315), whereas 585 genomes have fused Acd α+β sequences with the most sequence copies at four, and only 91 genomes contain fused Acd β+α sequences (Table S5). The widespread presence of Acd across different archaeal superphyla is consistent with the hypothesis that Acd was possibly present in LACA.[22] Among genomes with Acd, genomes with only independent (not fused) Acd α were dominant in DPANN (376/434) and TACK (137/321), whereas genomes with only fused Acd α+β were dominant in Euryarchaeota (195/420, Table S5). Notably, genomes with both independent Acd α and fused Acd α+β are dominant in Asgard with Acd (59/77, Table S5). Moreover, the genomes at the base of TACK and Euryarchaeota almost only have independent Acd α, including *JANJXX01*, *Korarchaeia*, and *Methanomethylicia* of TACK and *Thermococci* and *Hadarchaeia* of Euryarchaeota (Table S5), suggesting that the independent Acd α may represent the ancestral state of all Acd types. The protein tree of all the Acd α sequences could be divided into three clades, one of which is dominated by independent Acd α, whereas in the other two clades, independent Acd α and Acd α in fused Acd α+β are sister groups (Figures 4A & S3). In the VAE landscape, although the sequences of Acd α in Acd α+β are close to the coordinate origin, those of independent Acd α are closer (Figure S3). Similar results were obtained from the protein tree and VAE landscape of all the Acd β (Figures 4A & S3). The above results are consistent with a scenario in which Acd α and β may have existed independently and in multiple copies in early archaea and some of them may have fused into Acd α+β multiple times during the early evolution of archaea. Finally, Acd β+α is densely distributed in *Nitrososphaerales* and *Bathyarchaeales* of TACK and *Methanococcales*, *Methanobacteriales*, *Halobacteria*, and *Thermoplasmata* of Euryarchaeota, and their sequences are almost always within the sequences of independent Acd subunits in either protein tree of the Acd α or β, which again suggested that the two independent subunits could fuse into a single protein (Figures 4A & S3). However, the number and distribution range of the fused Acd β+α are much smaller than those of the fused Acd α+β (Table S5), raising the possibility that an unknown selective pressure favored the fusion of the Acd subunits to occur in the order of α+β.

Acs is widely present in all three domains of life and is a universal acetate-related metabolism enzyme that may be present in the last universal common ancestor[77,78] and potentially inherited by early archaea. However, the protein tree of Acs is divided into mainly two large clades, TACK and Euryarchaeota, and sequences from DPANN and Asgard are distributed in multiple clusters within these two clades (Figures 4A & S3), consistent with a scenario in which the ancestors of Asgard and DPANN may have lost Acs and then regained it multiple times independently through HGT from TACK and Euryarchaeota. In the VAE landscape of Acs, sequences of Euryarchaeota and TACK are near the coordinate origin and extend into the corresponding two sets of evolutionary trajectories, and sequences of Acs from Asgard and DPANN are distributed in these two trajectories (Figures 4A & S3), yielding results similar to those of the phylogenetic analysis. Acs and Acd can coexist or complement each other in archaea, especially in Asgard, as all 78 genomes of Asgard in arc1855 have Acs and/or Acd (Table S5). Characterization experiments based on purified enzymes revealed that both enzymes

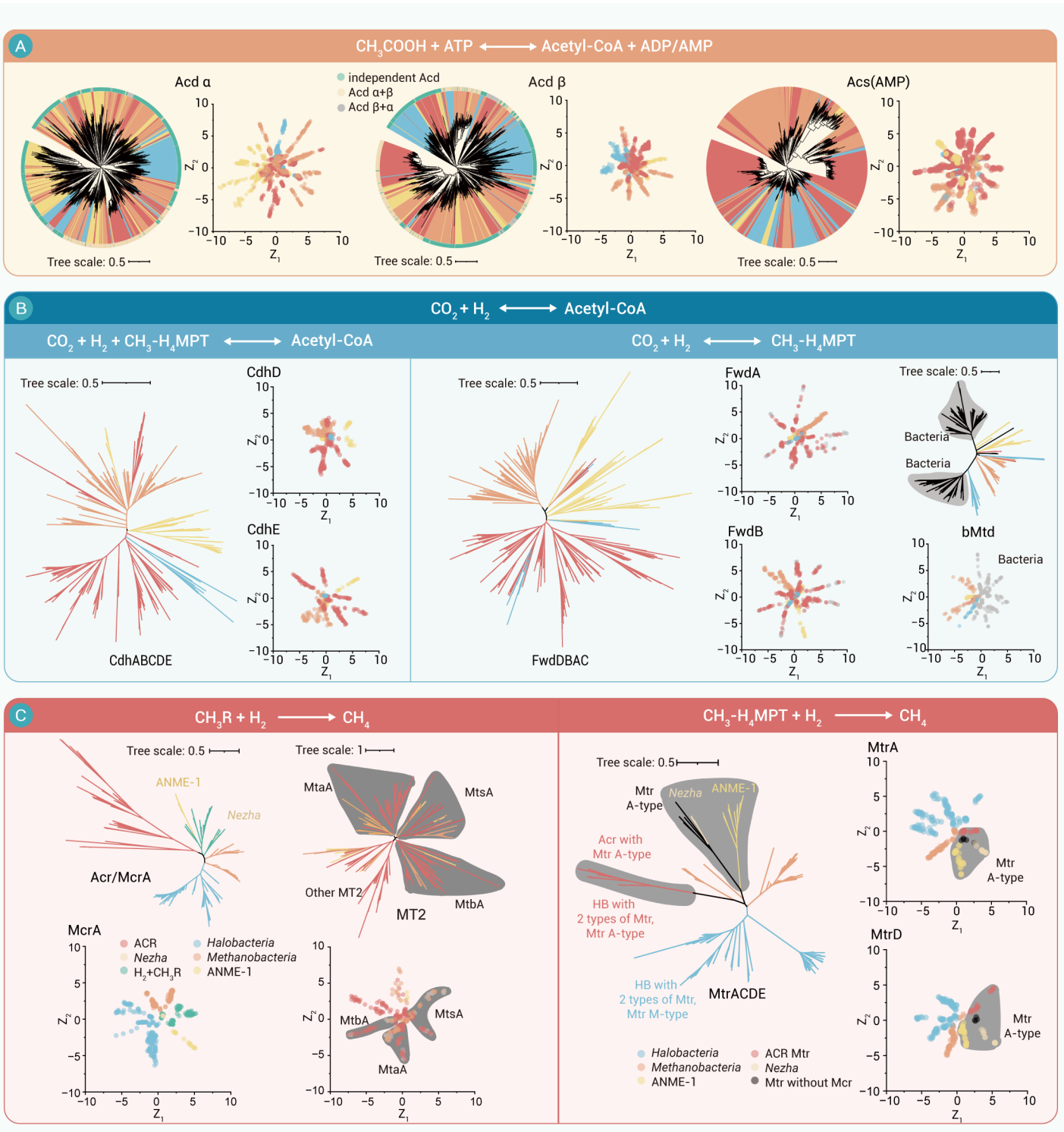


Figure 4. Protein trees and VAE evolutionary landscapes of enzymes related to acetate-related metabolism, the WLP and methanogenesis (A) Protein trees and VAE evolutionary landscapes of enzymes related to acetate-related metabolism are displayed, including α and β subunits of Acd and Acs (AMP). Blue represents protein sequences from DPANN, orange represents TACK, yellow represents Asgard, and red represents Euryarchaeota. And in both subunits of Acd, green represents sequences from independent Acd subunits, beige represents sequences from fusion Acd α+β, and gray represents Acd β+α. (B) Protein trees and VAE evolutionary landscapes of proteins related to the WLP in archaea are displayed, including CdhACDEB in the carbonyl branch, and FwdABCD, bmtd in the H4MPT branch, and gray fill indicates protein sequences from bacteria. (C) Protein trees and VAE evolutionary landscapes of McrABG, MT2, and MtrACDE related to methanogenesis are displayed. In Mcr, red indicates Acr, green indicates Mcr from H_2/CH_3R methanogens, yellow indicates Mcr from *Syntropharchaeia*, beige indicates Mcr from *Nezhaarchaeales*, orange indicates Mcr from *Methanobacteriota*, and blue indicates Mcr from *Halobacteriota*. In MT2, orange indicates protein sequences from TACK, yellow indicates those from Asgard, red indicates those from Euryarchaeota, and gray filled indicates the ranges of MtaA, MtbA, and MtsA. In Mtr, orange indicates Mtr from *Methanobacteriota*, blue indicates Mtr from *Halobacteriota*, yellow indicates Mtr from *Syntropharchaeia* and *Methanoliparia*, beige indicates Mcr from *Nezhaarchaeales*, black indicates Mtr from genomes without Mcr, and red indicates Mtr from genomes with Acr. Some *Methanosarcinia* have two subtypes of Mtr (Mtr A- and M-type), and Mtr A-type clusters with the red-labeled Mtr while Mtr M-type clusters with the blue-labeled Mtr. The gray filled indicates range of the Mtr A-type. CdhABCDE, FwdDBAC, and MtrABCD denote the concatenation orders of MSAs of these key protein.

could be reversed, but Acs is more active in the direction of acetyl-CoA, whereas Acd is more active in the direction of acetate.[37,79] The coexistence of the two enzymes may allow archaea to utilize acetyl-CoA as a vehicle for energy storage, whereas their complementarity suggests that acetate

concentrations vary in different space-time. Therefore, one plausible interpretation for the potential loss of Acs in the DPANN and Asgard ancestors is that these lineages might have experienced temporarily limited environmental acetate availability, reducing the selective pressure to maintain this energetically costly enzyme that normally converts acetate to acetyl-CoA at the expense of ATP.[37,79]

Ack and Pta originated from bacteria, and in their protein trees, archaeal sequences are distributed within different bacterial lineages (Figure S3), confirming that they were transferred into archaea multiple times independently through HGT from bacteria.[33,80] For example, *Methanosarcina* (GCA_004099695) evolved to utilize acetate for methanogenesis after acquiring Ack and Pta.[81] Moreover, the large-scale replacement of Acd and/or Acs by Ack and Pta occurred mainly in the *Woesearchaeales* of DPANN (Table S5). Therefore, a plausible scenario is that some ancestors of *Woesearchaeales* may have lost Acd and Acs during their process of reductive evolution and subsequently acquired Ack and Pta through ancient HGT from bacteria, potentially in response to changes in environmental acetate concentrations. Ack and Pta are sporadically distributed in other archaea in the form of replacing Acd or coexisting with Acd, potentially reflecting recent gene loss and/or gain events. Moreover, their quantity and distribution range are far smaller than those of Acs, and nearly none of them are distributed in Asgard (Table S5).

In summary, archaea possibly have acquired the above three different kinds of acetate-related metabolism through vertical inheritance (Acs), innovation (Acd), or horizontal gene transfer (Ack and Pta); moreover, the multiple gene duplications, losses, and gains of those enzymes in archaeal evolutionary history suggest that acetate-related metabolism may be an important factor associated with the evolution of archaea, particularly in Asgard and DPANN.

Wood–Ljungdahl pathway. The WLP in archaea consists of a carbonyl branch and a H_4MPT branch. The reactions in the carbonyl branch are catalyzed by the CdhACDEB complex, with CdhB regarded as a later addition that is functionally redundant to CdhA.[27,28,31] Following established practice,[82] we adopted a pathway-calling threshold of 75% of expected components to account for potential genome incompleteness; accordingly, if a genome encodes at least three of the subunits of CdhACDE, the archaeon is considered to possess a carbonyl branch (Table S6). The reactions in the H_4MPT branch are catalyzed sequentially by five enzymes, the FwdDBAC complex, Ftr, Mch, Mtd or bMtd or Hmd, and Mer.[26-28] If a genome has at least three genes encoding subunits of FwdDBAC, the archaea is considered to have an Fwd complex. Mtd, bMtd, and Hmd were separately identified in arc1855, and if one of them is identified in a genome, the archaea is considered to catalyze the fourth reaction in the H_4MPT branch. Nevertheless, as Mer is a large family of luciferases[26] and its sequences identified through alignment are taxonomically challenging to classify, it is not included in this study. For the whole pathway, we assume that if a genome has at least three genes encoding enzymes of the Fwd complex, Ftr, Mch, and Mtd or bMtd or Hmd, the archaea is considered to have the H_4MPT branch WLP (Table S6). Moreover, if a genome has both a carbonyl branch and a H_4MPT branch WLP, it is considered to have the complete WLP (Table S6).

The complete WLP is widely distributed in TACK (77/372), Asgard (38/78) and Euryarchaeota (119/523). However, only 10 genera near the base of DPANN have the near-complete WLP and 9 of them are from *Altiarchaeota* (Table S6). Based on this observation, there are two possible scenarios for the distribution of the WLP in *Altiarchaeota* based on the hypothesis that the complete WLP was potentially present in early archaea.[22,26,31] In Scenario one, the common ancestor of DPANN lost the WLP, and the ancestor of *Altiarchaeota* reacquired the complete WLP through HGT from other archaeal superphyla. In Scenario two, after the origin of DPANN, most DPANN ancestral nodes lost the WLP, but the ancestor of *Altiarchaeota* at the base of DPANN retained this pathway, which was ultimately vertically inherited by their descendants. However, by constructing protein trees and VAE protein evolution landscapes for the major protein families of the WLP, it appeared that the evolutionary scenarios for the H_4MPT branch and the carbonyl branch of the WLP in *Altiarchaeota* may differ.

CdhACDE of *Altiarchaeota* cluster together in their protein trees and are close to the coordinate origin in their VAE landscapes (Figures 4B & S4), consistent with Scenario two, under which the common ancestor of *Altiarchaeota* may have preserved the carbonyl branch and that other DPANN lineages may have lost it. Notably, in DPANN, only one *Altiarchaeota* genus and one *SpSt-1190* genus encode CdhB (Figure S4 & Table S6); thus, CdhB is nearly absent in *Altiarchaeota* and DPANN. Because CdhB is thought to have been added to the archaeal CdhACDE complex later,[31,83] under the hypothesis that DPANN represents an early-diverging lineage, the absence of CdhB in DPANN could reflect the addition of CdhB occurring in the common ancestor of Euryarchaeota and TACK. However, if DPANN diverged within, for example, Euryarchaeota, *Altiarchaeota* might be expected to have the complete CdhABCDE complex inherited from their possible Euryarchaeota-like ancestor, which could indirectly support the hypothesis that DPANN originated from LACA rather than Euryarchaeota. Nevertheless, CdhB is functionally redundant with CdhA,[83] and *Altiarchaeota* may have lost CdhB during its reductive evolution, as in most DPANN. However, these *Altiarchaeota* genera with complete WLP may be free-living, as inferred in some studies,[84,85] and maintain their genome sizes with an average genome size of 1.65 Mb, which is larger than the average genome size of DPANN in arc1855 at 1.01 Mb (Table S2). Therefore, more genomes and in-depth research are needed to confirm the absence of CdhB in *Altiarchaeota* and its underlying mechanism.

The evolutionary scenarios of the protein families in the H_4MPT branch of DPANN are more complex than those in the carbonyl branch, and seven other *Altiarchaeota* genera encode the H_4MPT branch without the Cdh complex (Table S6). *Altiarchaeota* may have gained the subunits of FwdDBAC and Ftr from other archaeal superphyla by HGT, as the related protein sequences do not cluster together in their protein trees and VAE landscapes, consistent with Scenario one (Figures 4B & S5). In contrast, the Mch sequences of *Altiarchaeota* cluster together in the protein tree similar to the subunits of Cdh, although these sequences are not close to the coordinate origin in the VAE landscape (Figure S5), consistent with the possibility that *Altiarchaeota* may have gained Mch via an ancient HGT. The reduction of methenyl-H_4MPT to methylene-H_4MPT is catalyzed by three isozymes, Mtd, bMtd, and Hmd. Among them, Mtd is widely distributed in archaeal superphyla other than DPANN, and only two Mtd sequences are identified in a genome in phylum *SpSt-1190* of DPANN (Figure S5 & Table S6). *Altiarchaeota* in DPANN only use bMtd, which is also distributed in TACK-Asgard; however, only 8 bMtd sequences are identified in Euryarchaeota (Figures 4B & S5 & Table S6). Moreover, there may be two subtypes of Hmd with sequence and structural differences, and Hmd is found only in Euryarchaeota (Figure S6 & Table S6). Because some bMtd sequences of *Altiarchaeota* are very close to the coordinate origin in the VAE landscape (Figure 4B) and other superphyla already have functionally equivalent Mtd or Hmd (Figure S5 & Table S6), it is hypothesized that bMtd may have originated independently within *Altiarchaeota* of DPANN and subsequently been transferred into TACK, Asgard, and bacteria by ancient HGT (Figure 4B). However, it cannot be completely ruled out that bMtd may have originated in the common ancestor of TACK and Asgard, as inferred by another study,[26] and that *Altiarchaeota* gained bMtd via HGT (Figure 4B), though it remains speculative. Almost all the above-mentioned enzymes in the H_4MPT branch appear to have been horizontally transferred to *Altiarchaeota*. Therefore, the common ancestor of DPANN might have lost the H_4MPT branch, and the *Altiarchaeota* ancestor at the base of DPANN in the species tree possibly reacquired it later. However, this does not conform to the principle of evolutionary parsimony. Thus, this study does not provide support for the hypothesis that the H_4MPT branch of the WLP was present in the common ancestor of DPANN under the current evolutionary framework, which represents one of several competing hypotheses regarding archaeal evolution.

If adopting the findings of Baker et al.,[23] which places DPANN as a subgroup within Euryarchaeota, it becomes difficult to explain the aforementioned observations: the absence of the CdhB subunit and the near-exclusive reliance on bMtd in DPANN (Euryarchaeota mainly rely on Mtd and the exclusive Hmd), and the loss of the H_4MPT branch in *Altiarchaeota* at the base of DPANN followed by subsequent reacquisition via HGT. Alternatively, under the results of Williams et al.,[22] which places DPANN at the root of archaea, these issues are more readily reconciled with a scenario in which LACA potentially only possessed CdhACDE functioning in catabolism, as in the last universal common ancestor,[31,86] and the H_4MPT branch and CdhB might have

originated from the common ancestor of TACK-Asgard and Euryarchaeota after the divergence of DPANN. We note that the strength of this interpretation depends on the phylogenetic placement of DPANN, and it remains one of several possible evolutionary scenarios that warrant further testing with additional data and refined models.

In summary, a hypothesis consistent with our analysis is that the H_4MPT branch may have originated in the common ancestors of TACK-Asgard and Euryarchaeota, and with the carbonyl branch, they could fix CO_2 to acetyl-CoA. This metabolic difference is hypothesized to be associated with the divergence of DPANN from other archaea, and most of DPANN may have become dependent on ectosymbiosis with other species for carbon sources. While the presence of the complete WLP in *Altiarchaeota* has been noted,[26,31] the differential evolutionary histories of its carbonyl and H_4MPT branches—suggested here by the discordance in both protein trees and VAE landscapes—represent a nuanced extension of previous hypotheses. In our hypothesis, *Altiarchaeota* appears to retain the carbonyl branch of the WLP from LACA and to have acquired enzymes in the H_4MPT branch through HGT or even independent innovation (bMtd), thereby potentially restoring the complete WLP and maintaining the potential for free living.

Methanogenesis. Methanogenesis-related proteins include the methyl-coenzyme M reductase complex (Mcr), tetrahydromethanopterin S-methyltransferase complex (Mtr) and methyl-substrate:coenzyme-M methyltransferase complex (MT).[25,29,30,87] If a genome has any gene encoding subunits of McrABG, the archaea is considered potentially methane related, and if a genome has at least three genes encoding subunits of MtrACDE, it is considered to have a Mtr complex (Table S7). MT includes substrate-specific MT1 (including MtaB, MtbB, MttB, MtmB, and MtsB) and MT2 (including MtaA, MtbA, and MtsA), as well as a corrinoid protein, which were separately identified in arc1855.[30,87] MT is widely distributed in archaea and does not strongly correlate with methanogens, and this study focused on MT2. The methanogenesis type of a genome was determined by combining the presence of Mcr, Mtr, and the H_4MPT branch of WLP, as well as their taxonomic information (Table S7).

As the key enzyme for methanogenesis, Mcr has been proposed to be traced back before the divergence of TACK-Asgard and Euryarchaeota.[22,24,29,30] In H_2/CH_3R methanogenesis, MT (including Mta/Mtb/Mts)) directly transfers the methyl group to coenzyme M, after which Mcr reduces methyl-coenzyme M to methane.[25,30] An analysis of the consistency between the species tree of methanogens and the protein tree of MtaA revealed that MtaA may be vertically inherited in TACK and Euryarchaeota and may have existed in their common ancestor.[24,30,87] Therefore, H_2/CH_3R methanogenesis represents one plausible earliest form of methanogenesis that might have been present in the common ancestor of TACK and Euryarchaeota. The other types of methanogenesis all require Mtr to link Mcr and the H_4MPT branch of the WLP to complete methanogenesis.[25,29,30] Methyl-dismutating and acetoclastic methanogenesis are found only in *Methanosarcinia* and may be metabolic innovations of this lineage.[81,88] However, some studies suggest that because Mtr in *Nezhaarchaeales* and other TACK archaea may be vertically inherited from extinct lineages of H_2/CO_2 methanogens within TACK, Mtr and H_2/CO_2 methanogenesis can also be traced back to the divergence of TACK and Euryarchaeota.[24,29] Therefore, whether H_2/CH_3R methanogenesis or H_2/CO_2 methanogenesis is the earliest type is still debated.

In this study, the McrABG subunits of H_2/CH_3R methanogens in the VAE protein evolutionary landscapes are slightly closer to the coordinate origin, consistent with the possibility that H_2/CH_3R methanogens represent an earliest form (Figures 4C & S7). The McrABG subunits of ANME-1 in *Syntropharchaeia* and H_2/CH_3R methanogens cluster together in both protein trees and the VAE landscapes, as ANME-1 probably gained McrABG from *Methanofastidiosales* by HGT and McrABG replaced its original alkyl–coenzyme M reductase complex (Acr).[30,89] H_2/CH_3R methanogenesis requires the MT to transfer a methyl group from CH_3R to CoM,[24,25,30] and based on structural alignment, although utilizing different methyl substrates, the methyltransferases that transfer methyl groups from corrinoid proteins to CoM (MT2, including MtaA, MtbA, and MtsA) have high structural similarity and catalyze the same reaction, indicating that MtaA, MtbA, MtsA and other MT2 are homologous (Figure S8). Therefore, the VAE landscape and the protein tree of all kinds of MT2 identified in arc1855 were constructed. MT2 is a large protein family widely distributed across the three archaeal superphyla other than DPANN, and its distribution does not show a strong association with methanogens in arc1855 (Figure 4C & Table S7). In the VAE landscape, the sequences of MtaA and MtsA near the coordinate origin are from both TACK and Euryarchaeota. In the protein tree, both MtaA and MtsA are split into two clades at the base of the tree; the sequences at the base of one clade are from TACK, while those at the base of another clade are from Euryarchaeota (Figure 4C). In the VAE landscape, the sequences of MtbA and other MT2 near the coordinate origin are only from Euryarchaeota, and in the protein tree, MtbA has a clade only from Euryarchaeota, whereas the other MT2 are split into two clades, and the sequences at the base of them are mainly from Euryarchaeota (Figure 4C). Therefore, MT2 might have originated and duplicated in the common ancestor of Euryarchaeota and TACK-Asgard before these lineages diverged. Taken together, these findings raise the possibility that H_2/CH_3R methanogenesis dependent on Mcr and MT could have been present in the common ancestor of TACK and Euryarchaeota, although alternative interpretations cannot be excluded.

Because all other types of methanogenesis and anaerobic oxidation of methane (AOM) in *Methanobacteriota* and *Halobacteria* require the Mtr complex to link the H_4MPT branch of the WLP with Mcr,[29,30] these Mtr-dependent types of methanogenesis would logically emerge after the origin of the Mtr complex. Mtr is predominantly found in Euryarchaeota, with a few occurrences in *Nezhaarchaeales* or other archaea within TACK (Figures 4C & S9, Table S7). Based on protein trees and VAE landscapes, the Mtr complex may have originated within the *Methanobacteriota* of Euryarchaeota and may have been vertically transferred along with the Mcr complex from *Methanobacteriota* to *Halobacteria*. However, some methanogens of *Methanobacteriota* and *Halobacteria* have entered host-related systems, such as *Methanosphaera* of *Methanobacteriota* isolated from bovine rumen digesta or human stool[90] and *Methanimicrococcus* of *Halobacteria* from a cockroach (*Periplaneta americana*).[91,92] In these host-related systems, where CH_3OH and H_2 are abundant,[90-92] these originally H_2/CO_2 methanogens with complete Mtr complexes can also perform H_2/CH_3OH methanogenesis (Table S7), consistent with the hypothesis that if the environment of the common ancestor of methanogens was rich in H_2 and methyl substrates,[93-95] H_2/CH_3R methanogenesis would be the earliest type of methanogenesis, regardless of the presence of the Mtr complex.

Moreover, two subtypes of Mtr were identified from five genera within *Methanosarcinia* in *Halobacteria*, namely, *UBA204* (GCA_030621155), *Methanoperedens A* (GCA_014859785), *Kmv04* (GCA_014859725), *QBUR01* (GCA_029210685), and *DQIP01* (GCA_030611265) (Figures 4C & S9, Table S7). This is consistent with the hypothesis that gene duplication of Mtr may have occurred in the common ancestor of *Halobacteria* or before the divergence of *Methanoliparia*, *Syntropharchaeia*, and *Methanosarcinia*, potentially leading to the emergence of two Mtr subtypes, which were temporarily designated Methane(M)- and Abnormal(A)-type (Figures 4C & S9, Table S7). In *Halobacteriota*, ANME-1 (AOM, Mcr)[30,89] in *Syntropharchaeia*, *Methanoliparia* (non-syntrophic methanogenic hydrocarbon degradation, Acr and Mcr),[96] and *EX4572-44* (anaerobic oxidation of alkane, Acr)[97] in *Methanosarcinia* have the Mtr A-type, whereas other *Halobacteriota* (Mcr), including *Methanosarcinia*, *Bog-38*, *Methanocellia*, and *Methanomicrobia*, have the Mtr M-type (Table S7). The sequences of the Mtr M-type do not significantly change from the preduplication state, whereas the Mtr A-type may have undergone rapid evolution, as the sequences of each subunit of the Mtr A-type complex are represented as longer branches with variable positions in the protein trees and form separate clusters radiating directly from the coordinate origin in the VAE evolutionary landscapes (Figures 4C & S9). *Nezhaarchaeales* may have gained the Mtr A-type from *Halobacteriota* through HGT, shifting from H_2/CH_3R methanogenesis to Mtr-dependent methanogenesis. Notably, according to a recent report on the cultivation of *Ca.* Methanonezhaarchaeum fastidiosum YNP3N, it cannot grow in the absence of methylamines, suggesting that it actually performs methyl-related methanogenesis rather than the H_2/CO_2 methanogenesis inferred from MAG,[98] thus, H_2/CO_2 methanogenesis is still only distributed in Euryarchaeota currently. Moreover, some Mtr A-type genes have been horizontally transferred to archaea without Mcr in TACK such as in *Hecatellaceae*, and their function remains unclear (Figures 4C & S9, Table S7). Interestingly, archaea

with the Mtr A-type appear to have periods when Mtr was present but inactive; for example, *EX4572-44* only has Acr,[97] whose Mtr would be inactive, and in ANME-1, before its original Acr was replaced by Mcr,[30,89] Mtr would be inactivated (Table S7). This may explain the rapid evolution of the Mtr A-type, which clearly differs from the Mtr M-type and may have new functions not related to methanogenesis, but may have still maintained its original structure and function (Figures 4C & S10). However, further research is needed to determine whether the Mtr A-type originated through gene duplication and underwent rapid evolution. Notably, *Methanosarcinia* within *Halobacteria* have a diverse array of methanogenesis, including H_2/CH_3OH, methylotrophic and acetoclastic methanogenesis, AOM, and even alkane oxidation; the archaea with two Mtr subtypes also belong to *Methanosarcinia*. Therefore, the diversification of methanogenesis may be potentially associated with Mtr duplication and rapid evolution; however, further in-depth research is needed.

In summary, though several evolutionary hypotheses exist and remain hotly debated, our results are consistent with the hypothesis that Mcr and MT may have originated before the divergence of Euryarchaeota and TACK-Asgard, and H_2/CH_3R methanogenesis represents a plausible earliest form (Figure 5A). Within Euryarchaeota, Mtr may have originated before the divergence of *Methanobacteriota*, and then Mcr and Mtr may have been co-inherited within Euryarchaeota, potentially giving rise to various Mtr-dependent methanogenesis, especially in *Methanosarcinia* (Figure 5A). *Methanobacteriota* and *Halobacteria* primarily utilize Mtr for methanogenesis, but some of its descendants shift to H_2/CH_3OH methanogenesis in methanol-rich host-related environments (Figure 5A). In the *Halobacteriota*, before the divergence of *Methanoliparia*, *Syntropharchaeia*, and *Methanosarcinia*, Mtr may have undergone gene duplication, resulting in two Mtr subtypes (A- and M-type, Figure 5A). Most methanogens in *Halobacteriota* are the Mtr M-type, whereas the Mtr A-type is mainly found in ANME-1, *Methanoliparia*, and *EX4572-44* of *Methanosarcinia*, which either currently or previously may not utilize Mtr for methanogenesis, consistent with the possibility that the Mtr A-type may have rapidly evolved, resulting in significant differences in sequence from the other Mtr types (Figure 5A & Table S7). To our knowledge, the divergence of Mtr into two subtypes (M-type and A-type) linked to different methanogenesis has not been previously highlighted at this phylogenomic scale. Within TACK, some descendants retained H_2/CH_3R methanogenesis, and some descendants, such as *Nezhaarchaeales*, may have acquired the A type Mtr through HGT, thereby shifting to methyl-dismutating methanogenesis (Figure 5A).

Perspective on archaeal origin and metabolism. The distribution and evolutionary relationships of the above proteins among arc1855 are consistent with the hypothesis that acetate-related metabolism, the carbonyl branch of the WLP, and H_2/CH_3R methanogenesis may have existed in LACA. Afterward, DPANN, TACK-Asgard, and Euryarchaeota possibly diverged rapidly after LACA. The H_4MPT branch then potentially originated in the common ancestor of TACK-Asgard and Euryarchaeota, in line with the hypothesis that the ancestor most likely was a generalist capable of ensuring at least one energy pathway and one biosynthetic pathway in a variety of environments (Figure 5B).[22,30,99] If H_2 and CO_2 are abundant, it can fix CO_2 into acetyl-CoA via the WLP and generate acetate from excess acetyl-CoA via Acd to obtain ATP. If CO_2 is scarce but methyl substrates, H_2, and acetate are abundant, it can obtain energy through H_2/CH_3R methanogenesis while consuming ATP to convert environmental acetate into acetyl-CoA via Acs. The common ancestor of TACK and Asgard probably retained the generalist state (Figure 5B), as exemplified by modern archaea in TACK, including *Bathyarchaeota-3300028193_21*, *LGG-1_spades_metabat_bin213*, and two genera of *JACAEJ01*, which possess WLP, acetate-related metabolism, and H_2/CH_3R methanogenesis (Figure 5B & Table S8). The common ancestor of Euryarchaeota was probably also in the generalist state (Figure 5B), but modern Euryarchaeota with all three metabolic pathways have not been found. After the WLP was lost during their further evolution, TACK and Euryarchaeota with H_2/CH_3R methanogenesis gave rise to most modern H_2/CH_3R methanogens (Figure 5B). Mtr may have originated within Euryarchaeota, linking the H_4MPT branch of the WLP to Mcr to achieve H_2/CO_2 methanogenesis, potentially adapting to environments rich in H_2 and CO_2 (Figure 5B) and diversifying into other kinds of Mtr-dependent methanogenesis consistent with environmental changes. Most TACK and Asgard eventually lost methanogenesis and retained only the WLP and acetate metabolism, and potentially used the WLP as an electron acceptor during the fermentation of organic substrates for heterotrophic growth (Figure 5B).[32] The origin of the H_4MPT branch in the common ancestor of TACK and Euryarchaeota, and the potential loss of methanogenesis (or never has) in the common ancestor of DPANN may have been associated with the divergence of DPANN from other archaea. However, the *Altiarchaeota* at the base of DPANN possibly retained the carbonyl branch and gained the H_4MPT branch through HGT, thus acquiring the complete WLP (Figure 5B). These observations suggest that acetate-related metabolism, the WLP, and methanogenesis, along with environmental changes, were associated with the early archaeal diversification (Figure 5B) and may have contributed to the foundation for the development of diverse metabolisms in modern archaeal species. We acknowledge that these metabolic scenarios remain hypothetical; their resolution will require more advanced AI tools, phylogenomic reconciliation analyses, biochemical validation, and expanded genome sampling.

To further explore the potential metabolic evolutionary history of early archaea, we additionally surveyed key genes involved in a range of metabolic pathways across arc1855, including hydrogen metabolism, pyruvate synthesis, phosphoenolpyruvate (PEP) metabolism, the non-oxidative pentose phosphate pathway, the TCA cycle, glycolysis/gluconeogenesis, and nitrogen and sulfur metabolisms (Figures S11-S17 & Table S9). Although the genes associated with these pathways are broadly distributed across the archaeal domain (Table S9), their evolutionary histories are obscured by pervasive horizontal gene transfer, lineage-specific gene gains and losses, and the potential extinction of intermediate lineages, which together render ancestral reconstructions inherently uncertain. Despite these challenges, the majority of these genes are widely present in both TACK-Asgard and Euryarchaeota, with the notable exception of 2,3-bisphosphoglycerate-independent phosphoglycerate mutase (Pgm) in glycolysis, which is absent from TACK-Asgard. Furthermore, in several phylogenetic trees, sequences from TACK-Asgard and Euryarchaeota each tend to form monophyletic clades (Figures S11-S17 & Table S9). These patterns are consistent with the hypothesis that the last common ancestor of TACK-Asgard and Euryarchaeota possessed a complex metabolic repertoire and a generalist lifestyle adaptable to diverse environments.[22,30,99] By contrast, DPANN sequences cluster together only in a subset of gene trees, including those for hydrogenase subunits (HydABG), pyruvate kinase (Pyk), and several key enzymes in pyruvate synthesis, the non-oxidative pentose phosphate pathway, and glycolysis/gluconeogenesis (Figures S11-S14 & S16). This pattern is compatible with the hypothesis that DPANN lineages underwent extensive genome reduction and reductive evolution associated with a transition from a free-living to an ectosymbiotic lifestyle.[23,100,101] Under the phylogenetic framework in which DPANN is placed within Euryarchaeota, LACA would likely have possessed complete core carbon metabolism, hydrogen metabolism, and sulfur and nitrogen metabolism, with subsequent differential loss shaping the modern distributions observed in the arc1855 dataset. Alternatively, under the early-diverging-DPANN hypothesis,[22] with which our results are consistent, LACA may still have harbored a comparably complex metabolism, as gene loss in DPANN could have occurred after its divergence from LACA. Collectively, these analyses outline a set of testable hypotheses regarding the metabolic capacity of LACA; more comprehensive phylogenomic reconciliation analyses, complemented by structural and biochemical validation, will be essential to resolve these ancestral states with greater confidence.

Coevolution of archaea and earth environments

As described above, in the VAE-PCA evolutionary landscape of arc1855 based on the set of 37 conserved genes, the genomes at the top of the cone are densely distributed, whereas at the base of the cone, the genomes are distributed along eight trajectories (Figure 3B). The genomes on each trajectory belong to the same phylum and cluster together in the corresponding species tree (Figure 3B). Notably, the archaea located on some trajectories may be strongly associated with specific environments (Figure 6E & Table S8). The trajectory of *Thermoprotei* of TACK primarily includes members from extremely high-temperature environments,[102] the trajectory of *Bathyarchaeia* of TACK mostly includes members from high-temperature

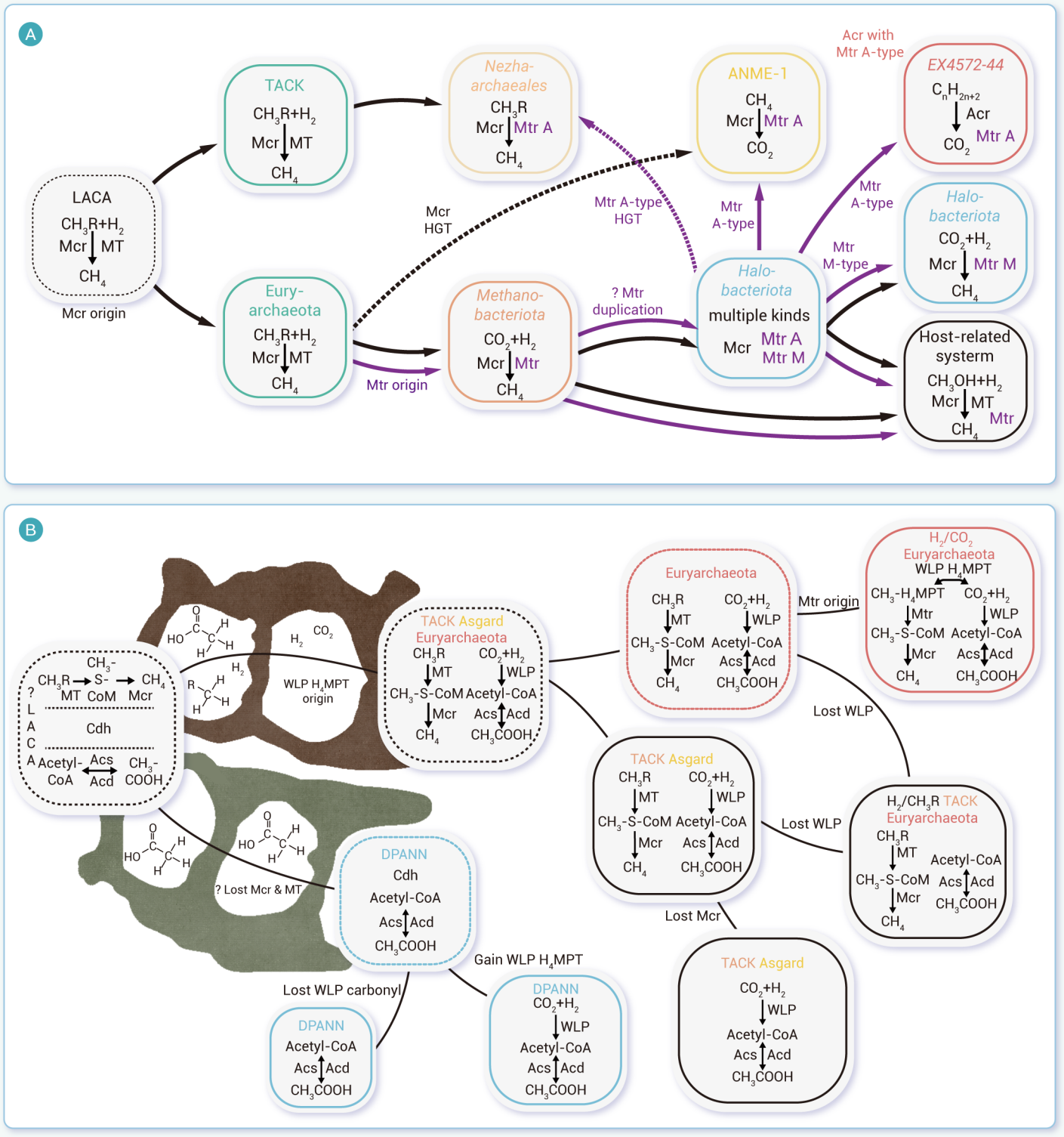


Figure 5. Acetate-related metabolism, the WLP, and methanogenesis in LACA and its descendants (A) Possible evolutionary history of methanogens was inferred. Black arrows indicate the transfer process of Mcr and purple arrows indicate that of Mtr, and dashed arrows indicate possible horizontal gene transfer while solid arrows indicate almost vertical inheritance. (B) Metabolic pathways consistent with the hypothesis that the carbonyl branch of WLP, acetate-related metabolism, and H_2/CH_3R methanogenesis may have been present in LACA are shown, as well as pathways potentially present in its various descendants. The dashed boxes indicate the inferred transitional states of archaea that may have existed in geological history but have not been found in modern archaea.

environments,[103,104] the trajectory of *Nitrososphaeria* of TACK largely consists of members from aerobic environments,[105-107] and the trajectory of *Halobacteriota* of Euryarchaeota predominantly includes members from aerobic and hypersalinity environments (Figure 6E).[108-110] These observations are consistent with the possibility that archaea may have expanded into different types of environments during their later stages of evolution. Therefore, it is interesting to study how archaea adapt to various extreme environmental conditions and the coevolution of archaea with Earth environments. To this end, in this study, a variety of deep learning models based on single proteins or entire genomes were integrated to predict the optimal growth temperatures (pOGTs), hypersalinity tolerance, and oxygen tolerance of arc1855 (Figure 6A & Table S8). As these predictions provide a genome-scale, computational

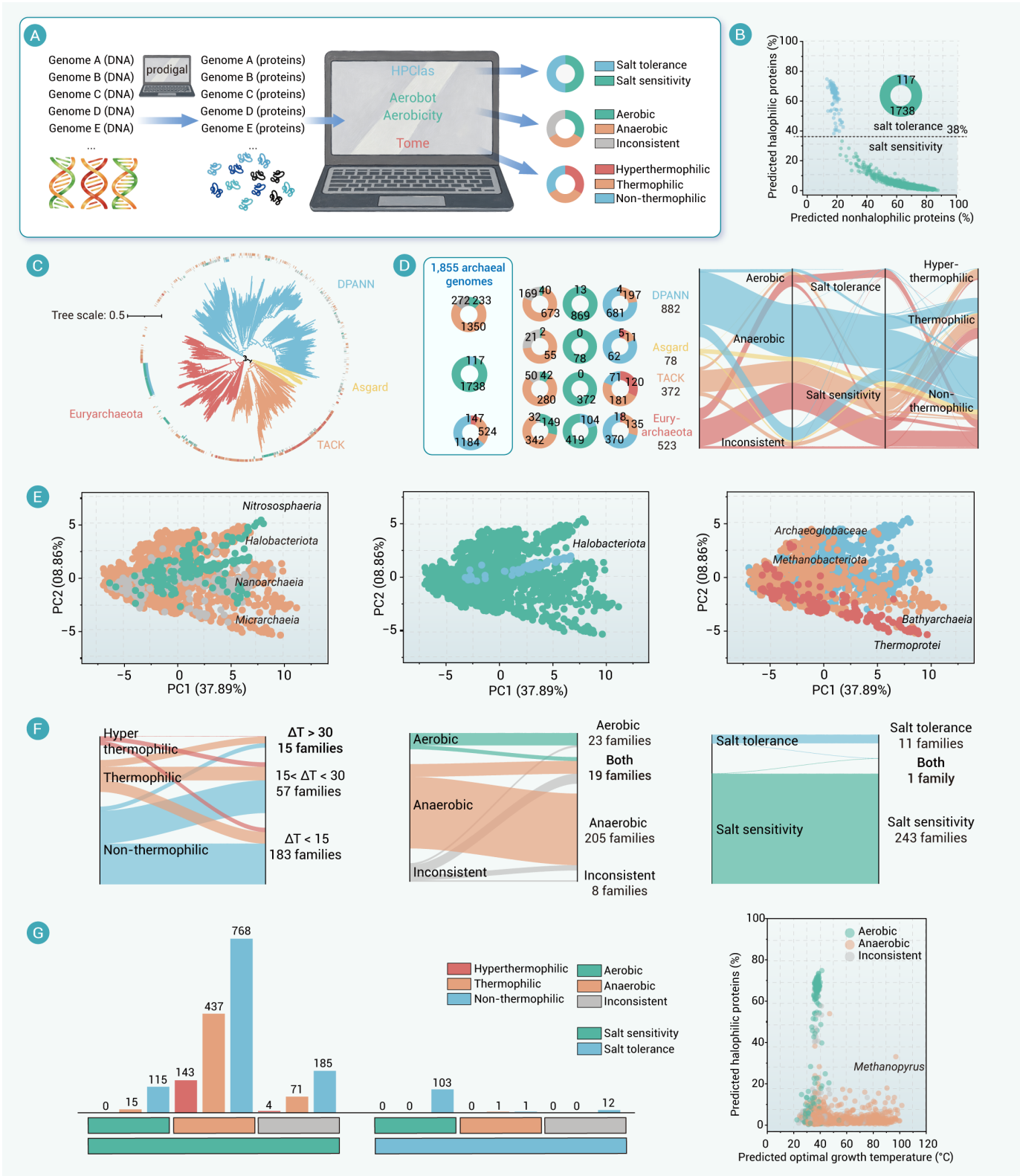


Figure 6. Predicted optimum physiological conditions for 1855 archaeal genomes (A) Based on the whole-genome protein sequences, predictions of whether those archaea can tolerate hypersalinity environments (HPClas) or oxygen (Aerobicity and Aerobot), and their optimal growth temperatures (pOGTs, by Tome). For oxygen-tolerance, if Aerobicity and Aerobot provides consistent results, the results are reliable, while if the predictions are inconsistent, the corresponding archaea are labeled as "inconsistent". For OGT, archaea with a pOGT less than 45°C are classified as non-thermophilic, those with a pOGT between 45°C and 80°C are classified as thermophilic, and those with a pOGT more than 80°C are classified as hyperthermophilic. (B) For hypersalinity tolerance, if the proportion of predicted halophilic proteins (pHPP) by HPClas in a genome is more than 38%, the archaea is predicted to be able to tolerate hypersalinity environments. (C) Distribution of 1855 archaea with predicted physiological conditions on the species tree based on 37 conserved genes is displayed. (D) Proportions of archaeal genomes with predicted physiological conditions among all archaeal genera and within each superphylum are displayed. (E) Distribution of archaeal genera with predicted physiological conditions in the VAE-PCA evolutionary landscape based on 37 conserved genes is displayed. (F) The 1855 archaeal genomes belong to 564 families, and 255 families contain more than one genus. Among them, only one family contains members that are predicted to either tolerate or not tolerate hypersalinity environments, 19 families contain genera that are predicted to either tolerate or not tolerate oxygen, and pOGT ranges of 15 families exceed 30°C. (G) The prediction of adaptation to three environmental factors for 1,855 archaeal genomes was statistically analyzed and displayed. In the right-hand scatter plot, the X-axis represents the pOGT (°C) of archaea, while the Y-axis represents the pHPP (%).

perspective on the ecological preferences of these archaea, they represent testable hypotheses whose confirmation will benefit from targeted culture-based experimental validation under the inferred conditions.

Temperature. The pOGTs of arc1855 were predicted by Tome,[42] and among them, 1,184 genera with pOGTs less than 45°C were classified as nonthermophilic, 524 genera with pOGTs between 45°C and 80°C were classified as thermophilic, and 147 genera with pOGTs greater than 80°C were classified as hyperthermophilic (Figure 6D & Table S10A). In the VAE-PCA landscape, the genera at the top of the cone are predicted to adapt to a variety of temperatures, whereas those on the trajectories at the base of the cone adapt to consistent temperatures (Figure 6E). These findings are consistent with the hypothesis that archaea may have rapidly adapted to a broad range of temperatures during the early stages of evolution. Among the 372 genera from TACK, approximately one-third (120 genera) are hyperthermophilic, and nearly half (181 genera) are thermophilic (Figure 6D). Among the 523 genera from Euryarchaeota, only 18 are hyperthermophilic, and 135 are thermophilic; however, these hyperthermophilic and thermophilic archaea are distributed at the base of the Euryarchaeota and TACK lineages (Figures 6C-E, Tables S2 & S10A). This is consistent with the possibility of a thermophilic or even hyperthermophilic origin for the common ancestors of Euryarchaeota and TACK[22,111] and suggested that different temperature environments are associated with diversity among Euryarchaeota (Figures 6C-E). In DPANN and Asgard, thermophilic and hyperthermophilic members together account for less than a quarter of the genera, consistent with the possibility that nonhigh-temperature environments are associated with the divergence of Asgard from TACK and DPANN from other archaea (Figures 6C-E, Tables S2 & S10A).[22,49]

Oxygen tolerance. Two deep learning models, Aerobot[48] and Aerobicity,[47] were used to predict the oxygen tolerance of arc1855 (Figure 6A). If they provide consistent results, the results are considered to be reliable; however, if the predictions are inconsistent, the corresponding archaeal genera are labeled "inconsistent", and these archaeal genera may be in a transitional state between oxygen tolerance and intolerance. Among arc1855, 233 genera are oxygen tolerant, whereas 1,350 genera are oxygen intolerant, and 272 genera are labeled "inconsistent" (Figure 6D). In the VAE-PCA landscape, both oxygen-intolerant and oxygen-tolerant genera are distributed at the top of the cone, whereas those on the trajectory at the base of the cone are either oxygen-intolerant or oxygen-tolerant (Figure 6E). Interestingly, *WAKA01* (GCA_015523565) and *Kariarchaeum* (GCA_002728275) of *Heimdallarchaeales* are very close to the vertex of the cone, consistent with the hypothesis that oxygen may be associated with the evolution of early archaeal lineages, potentially prior to the Great Oxidation Event, as some studies inferred.[112-114] Moreover, oxygen-tolerant archaea are concentrated in *Nitrososphaerales* of TACK and *Thermoplasmataceae*, *SW-10-69-26*, *JACPAO01 Poseidoniales* and *Halobacteriales* of Euryarchaeota (Figures 6C-E, Tables S2 & S11A), and the other oxygen-tolerant archaeal species are distributed sporadically across the four archaeal superphyla (Figures 6C-E, Tables S2 & S11A), consistent with the association between oxygen and the diversification and niche expansion of the archaeal domain, especially in Euryarchaeota (149/523).

As a basis for predicting the optimal physiological conditions of organisms, there is a debate over whether the adaptation of organisms to different environmental factors is a result of adaptation across the entire proteome or even the cellular system, or whether it can be achieved through a set of specific genes. To predict the optimal growth temperature (OGT), researchers initially used reverse gyrase as a marker to identify hyperthermophilic organisms;[115-117] however, this enzyme was later discovered in thermophilic organisms.[118,119] Currently, predictions of OGT by deep learning models are almost entirely based on whole-genome-scale feature extraction,[42,44] as environmental temperature affects the whole proteome and the whole cell. In contrast, predictions of oxygen tolerance still involve two approaches. One is based on annotations focusing on a set of specific genes (as in Aerobicity),[47,120] and the other focuses on the whole-genome scale (as in Aerobot).[48] Researchers of Aerobicity have identified a series of specific protein families that are closely related to oxygen tolerance, and their predictions are based on the types and copy numbers of these genes in a genome.[47] This correlation was undeniable, but it may lack causality—that is, whether these genes helped organisms tolerate oxygen or whether organisms gained these protein families for oxygen utilization after they had been oxygen tolerant. Moreover, if oxygen tolerance is due to the gain of specific genes, the number of archaeal families with both oxygen-tolerant and oxygen-intolerant members should far exceed 19 families (Figure 6F), as the gain and loss of genes can be accomplished easily and are common in nature.[121-123] Therefore, the presence of certain oxygen-related genes in oxygen-tolerant archaea is more likely a result of their ability to tolerate oxygen. Additionally, the researchers of Aerobicity further developed a faster and more convenient model for predicting oxygen tolerance based on the concatenated MSA of a set of conserved genes in the genomes.[47] They reported that among the 148,765 sites of this MSA, only 69 had nonzero weights for predicting oxygen tolerance and concluded that differences in amino acid usage between aerobes and anaerobes were site specific rather than generalizable.[47] However, this may also be just a correlation, because if only 69 specific sites out of 148,765 determine whether an organism is aerobic or anaerobic, the number of archaeal families with both aerobic and anaerobic members should be far more than 19 (Figure 6F), as conserved genes and conserved acid amino sites also have certain mutation rates.[124-126] Therefore, this study proposes that adaptation across the entire proteome makes organisms adapt to temperature and oxygen, indicating that whole-genome-scale sequence features are predictive of whether an organism is aerobic or anaerobic and its OGT.

Hypersalinity. Extremely halophilic organisms typically employ a salt-in strategy, in which potassium and sodium ions accumulate within the cell, leading to a cytoplasmic salt concentration that is comparable to that of the external hypersaline environment.[127-129] Consequently, a high proportion of acidified halophilic proteins are present in the proteome, with exceptions including proteins that can be protected from the hypersaline cytoplasm (such as internal ribosome subunits and membrane proteins) or those under selective pressure to avoid acidification (such as DNA-binding proteins).[129] Therefore, like adaptation to temperature, adaptation to hypersalinity is also at the whole-genome scale. By utilizing HPClas,[46] it is convenient and fast to predict whether a protein is halophilic, and the proportion of predicted halophilic proteins (pHPP) among the whole-genome proteins of each genome in arc1855 can be determined. Finally, based on the pHPP (38% in this study) and sampling site information of a genome, whether the archaea are extremely halophilic can be predicted (Figures 6A-B).[46]

Traditionally, there are four groups of extremely halophilic archaea, namely, *Halobacteria*,[108,110] *Nanosalinia*,[130-132] *Methanonatronarchaeales*,[133] and *Halarchaeoplasmatales*.[134] In arc1855, 101 of the 103 genera in *Halobacteria* are extremely halophilic, whose pHPPs range from 43–75%, and 13 of the 17 genera in *Nanosalinia* are extremely halophilic, whose pHPPs range from 38–58% (Table S12B). Therefore, most members of *Halobacteria* and *Nanosalinia* are extremely halophilic archaea. However, in *Methanonatronarchaeia*, only one genus, *JAANXF01* (GCA_018263395),[135] has a pHPP of 54%, and none of the members in *Halarchaeoplasmatales* has a pHPP exceeding 38% (Table S12B). To eliminate the sampling bias of arc1855, all the genomes of *Methanonatronarchaeia* and *Halarchaeoplasmatales* were collected from GTDB R220; however, except GCA_018263395, the pHPPs in the other five genomes of *Methanonatronarchaeia* range from 9–15%, and those in the 17 genomes of *Halarchaeoplasmatales* range from 6–24% (Table S12C). Therefore, *Methanonatronarchaeia* and *Halarchaeoplasmatales* may be not extremely halophilic archaea, or they do not employ the salt-in strategy. Moreover, a study expanded the scope of extremely halophilic archaea, and new extremely halophilic archaea, such as *Asbonarchaeaceae* and *Afararchaeaceae*, were found.[109] *Asbonarchaeaceae* is a sister group to *Halobacteria*,[109] and pHPPs of its four MAGs range from 44–66%, whereas *Afararchaeaceae* belongs to *Nanosalinia*,[109] and among its 9 MAGs, except for GCA_034190605 (pHPPs at 29%), the pHPPs in the other 8 MAGs range from 49–63% (Table S12C). The above results validated the accuracy of the method used in this study for predicting extremely halophilic archaea.

Beyond those extremely halophilic archaea, the two genera of *SW-10-69-26* in *Thermoplasmatota*, *SW-10-69-26* (GCA_003009755) sampled from the salt crust in the Atacama Desert in Chile and *PWVY01* (GCA_003561665) sampled from hypersaline soda lake sediment in Russia, have proportions of predicted halophilic proteins (pHPPs) of 40% and 69%, respectively, suggesting that they may represent a new potentially independently evolved

halophilic archaeal lineage (Tables S2 & S12A). In summary, most proteins in the genomes of *Methanonatronarchaeia* and *Halarchaeoplasmatales* are not halophilic, and even though these organisms could tolerate hypersalinity, they may not mainly employ the salt-in strategy.[109,127-129,136] The extremely halophilic archaea defined in this study in arc1855 are only distributed in *Nanosalinia* of DPANN, *SW-10-69-26* of *Thermoplasmatota*, and *Halobacteria* of Euryarchaeota (Tables S2 & S12A), suggesting that hypersalinity may only have a limited impact on the diversity of the archaeal domain.

Combined effects of environmental factors. To quantify the adaptability of archaea to temperature, oxygen and hypersalinity, their adaptation to these three environmental factors was analyzed within each archaeal family. Arc1855 belongs to 564 families, and among them, 255 families contain more than one genus (Table S2). With respect to temperature, 15 families present a range of predicted optimal growth temperatures (pOGTs) exceeding 30°C, with *UBA233* having the highest pOGT range (more than 46°C). Within this family, *Termiticorpusculum* (GCA_009781675)[137] has a pOGT of 35.43°C, whereas *DSZD01* (GCA_018396935)[104] has a pOGT of 82.05°C (Figure 6F & Table S10B). Different genera within the same family can adapt to a wide range of temperatures, suggesting a strong adaptability of archaea to temperature. With respect to oxygen, 19 families contain both predicted oxygen-tolerant and oxygen-intolerant archaeal genera (Figure 6F & Table S11B). Among these 19 families, 11 families have more than 5 archaeal genera, and in 6 of these 11 families, half or more of the genera are labeled "inconsistent", suggesting that these families may be in a transitional state between oxygen tolerance and intolerance (Table S11C). Different genera within the same family can adapt to either aerobic or anaerobic environments, suggesting the strong adaptability of archaea to oxygen. However, for hypersalinity, almost no archaeal family has members that are either extremely halophilic or not, except for *Nanoanaerosalinaceae* (Figure 6F). This family comprises three genera. *JALIDO01* (GCA_022865265, proportions of predicted nonhalophilic proteins (pNHPP) at 18% and proportions of predicted halophilic proteins (pHPP) at 49%), which was sampled from a high-salinity lake in China (PRJNA820349, Table S12A), is predicted to be aerobic and extremely halophilic. In contrast, *JALIDQ01* (GCA_022865275, pNHPP at 23%, pHPP at 28%) from the same sampling site as *JALIDO01* and *Nanoanaerosalina* (GCA_022257775, pNHPP at 27%, pHPP at 20%), sampled from a soda lake in China (PRJNA797678), both are predicted to be neither aerobic nor extremely halophilic (Tables S11A & S12A). Therefore, almost all the members of an archaeal family are either extremely halophilic or not, suggesting that the difficulty for archaea to adapt to hypersaline environments is far greater than the difficulties associated with adapting to differences in temperature and oxygen.

Extremely halophilic archaea, such as most members of *Halobacteria* and *Nanosalinia*, are always oxygen tolerant; thus, it is interesting to investigate the adaptation of archaea to combinations of the above three environmental factors (Figure 6G). With respect to hypersalinity and the other two factors, almost all the extremely halophilic archaea are also oxygen tolerant and nonthermophilic, and there are only two exceptions in arc1855, as follows (Figure 6G). The extremely halophilic *JAANXF01* (GCA_018263395, pHPP at 54%) of *Methanonatronarchaeia* is thermophilic and anaerobic, and *SW-4-43-9* (GCA_003009795, pHPP at 38%) of *Nanosalinia* is anaerobic (Tables S11A & S12A). With respect to temperature and oxygen tolerance, only 15 thermophilic archaeal genera are predicted to be able to tolerate oxygen, and the oxygen tolerance of four hyperthermophilic archaeal genera and 71 thermophilic archaeal genera is inconsistent (Figure 6G & Table S8). Therefore, in arc1855, almost no hyperthermophilic archaea and only a few thermophilic archaea can tolerate oxygen. Collectively, these findings indicate that, in archaea, extremely halophilic activity is closely linked to oxygen tolerance, whereas a mutual exclusivity exists between oxygen tolerance and thermophilicity. Several possible explanations may account for this observed pattern, as outlined below. On the one hand, the introduction of oxygen into the cell may increase the ability of archaea to tolerate hypersaline environments, as extremely halophilic archaea require a substantial amount of energy to synthesize high amounts of intracellular osmotic molecules to reduce the difference in osmotic pressure across the cell membrane (salt-out strategy) or to actively expel inorganic ions into the cytoplasm for accumulation (salt-in strategy),[127-129,136] which may be difficult to achieve solely through anaerobic metabolism.[138-140] On the other hand, this finding may rule out the possibility of hyperthermophilic archaea tolerating oxygen, as in extreme high-temperature environments, free radicals are more likely to be activated within cells, which can damage cellular structures and genetic information.[141,142] Thus, the combination of hyperthermophily and extremely halophily is rare among archaea, consistent with the proposed constraints of oxygen (Figure 6G).

Coevolutionary history of Halobacteria and Nanosalinia. Evolutionary history of *Halobacteria* and *Nanosalinia* could serve as an excellent example of coevolution between species and their environments, since most of them are both extremely halophilic and oxygen tolerant and engage in ectosymbiotic relationships. Based on the predicted optimal physiological conditions and the sampling environments of *Halobacteria* and *Nanosalinia* in arc1855, this study proposed a speculative evolutionary scenario as follows. First, the last common ancestor of *Nanosalinia* may be similar to that of *JAFCCC01* (GCA_017610385),[143] with a predicted nonhalophilic protein proportion (pNHPP) of 75% and a predicted halophilic protein proportion (pHPP) of 3%, and this MAG was sampled from the Guaymas Basin, an anaerobic marine environment (PRJNA692327, Figures 7A-C & Tables S11A & S12A). Because no *Halobacteria* or methanogens were found in MAGs from PRJNA692327,[143] *JAFCCC01* and the last common ancestor of *Nanosalinia* were likely free-living organisms. Subsequently, some descendants of *Halobacteria* and *Nanosalinia* probably diverged into organisms similar to *JAHENH01* (GCA_018609935) and *JAHENU01* (GCA_018609855), respectively (Figures 7A-C). These two MAGs were both sampled from Bonneville Salt Flats Sediment (PRJNA522308), and both of *JAHENH01* (pNHPP 38% and pHPP 15%) and *JAHENU01* (pNHPP 42% and pHPP 16%) can tolerate oxygen and certain salt concentrations but not hypersalinity (Figures 7A-C & Tables S11A & S12A). The exclusive exosymbiosis between *Halobacteria* and *Nanosalinia* is consistent with the hypothesis that their ancestors may have independently acquired tolerance to oxygen and certain salt concentrations before engaging in ectosymbiosis in environments similar to salt flat sediments from which *JAHENU01* and *JAHENH01* were sampled. Finally, similar to most of the remaining *Halobacteria* and *Nanosalinia* genera, *Nanosalinia* appears to require ectosymbiosis with *Halobacteria*, and both may have tolerated oxygen and hypersalinity (Figures 7A-B & Table S8).[109,132] In *Nanosalinia*, there are 5 families and 17 genera, with 3 families (*Nanohalalkaliarchaeacea*, *Nanoanaerosalinaceae*, and *Nanosalinaceae*) being extremely halophilic (Figure 7B). *Nanohalalkaliarchaeacea* include two genera with pHPPs of 54% and 58% inhabiting deep sediment layers in hypersaline environments (Figures 7A-C & Table S12A).[132] *Nanoanaerosalinaceae* include three genera, with only one genus (*JALIDO01*, GCA_022865265, pHPP 49%) being extremely halophilic and inhabiting surface sediment layers.[132] *Nanosalinaceae* include 10 genera whose pHPPs range from 38–58%, inhabiting the brine layers (Figures 7A-C & Table S12A).[132] Interestingly, the pHPP in these *Nanosalinia* families decreased slightly from deep sediments to surface sediments and brine layers (Figure 7C), a pattern potentially explained by the observation that hypersaline lakes often have higher evaporation rates than inflows do, where salts accumulate in deep sediments and diffuse from there to surface sediments and brine layers.[144] In *Halobacteria*, as discussed above, 101 of the 103 genera (excluding *JAHENH01* and *UBA12382*) are extremely halophilic (Figure 7C). Notably, *UBA12382* (GCA_027249345, pNHPP 54%, pHPP 6%), sampled from the Sargasso Sea (PRJNA767318), a marine environment, is oxygen tolerant but may have lower salt tolerance than *JAHENH01* (pNHPP 38%, pHPP 15%), and similar to *JAFCCC01* (pNHPP 75%, pHPP 3%) of *Nanosalinia* in the marine environment (Figures 7A & C). Some studies have suggested that *UBA12382* (*Hikarchaea*) may evolve from extremely halophilic ancestors and have adapted to nutrient-poor deep-sea environments through gene loss.[109,145,146] In summary, the descendants of *Halobacteria* and *Nanosalinia* may adapt to extremely saline environments together, and their coevolution with their environments highlights the difficulty of adapting to extreme salinity.

Limitations of phylogenomic- and deep learning-based evolutionary studies. The VAE-PCA evolutionary landscape yields a representation that can be viewed as reflecting the relative divergence of genomes based on a conserved gene set, whereas the phylogenomic tree is reconstructed from the concatenated multiple sequence alignment of the same set of conserved

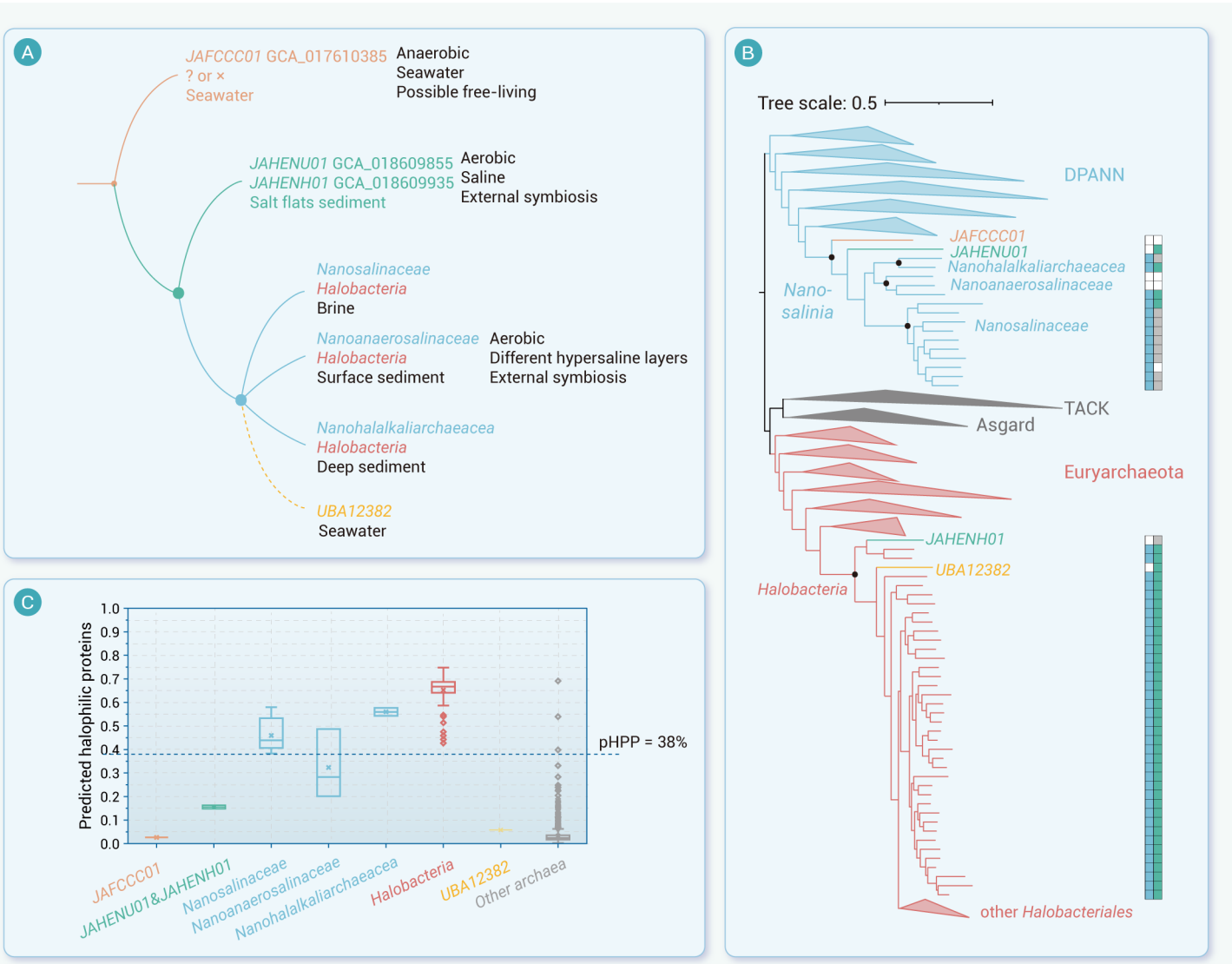


Figure 7. Coevolutionary history of Halobacteria and Nanosalinia (A) The coevolutionary history of *Halobacteria* and *Nanosalinia* is proposed based on their predicted optimum physiological conditions and the environmental contexts where their MAGs were sampled. The ancestor of *Nanosalinia* probably lived independently in seawater, and "? or ×" indicates that extant archaea analogous to the ancestor of *Halobacteria* (non-halophilic and anaerobic) may remain undiscovered or may not exist (orange). The exclusive exosymbiosis between *Halobacteria* and *Nanosalinia* is consistent with the hypothesis that their ancestors may have acquired tolerance to oxygen and a certain level of salinity, before engaging in ectosymbiosis in environments similar to where *JAHENU01* and *JAHENH01* were sampled from (green). Eventually, their descendants together have adapted to hypersalinity environments (blue), and *Nanosalinia* diversified into different ecological niches of hypersalinity environments (blue), and one genus (UBA12328) may have returned to marine (yellow). (B) The species tree of 1855 archaea based on 37 conserved genes focuses on *Nanosalinia* and some *Halobacteria* along with their predicted physiological conditions. In the first column, blue indicates that the archaea are predicted to tolerate hypersalinity environments, and in the second column, green indicates that the archaea are predicted to tolerate oxygen, and gray indicates inconsistent results from Aerobicity and Aerobot. (C) The proportions of predicted halophilic proteins in the above-mentioned groups in arc1855 are shown in the box plot.

genes. Both methods face the same challenge: which set of conserved genes should be selected to better describe the evolutionary relationships among genomes? Additionally, is it reasonable to describe the evolutionary relationships among genomes solely based on protein sequences? More essentially for deep learning-based evolutionary studies, how can protein sequences be represented with meaningful embeddings? In the preliminary work of this study, the evolutionary landscapes for 37 conserved genes were constructed using the ESM language model,[15,16] but these landscapes could not effectively distinguish proteins of different phylogenomic origins. The ESM model extracts the embeddings of each amino acid in a protein sequence, and the protein sequence is represented by averaging the embeddings of amino acids,[15] which does not provide a good global representation of the protein sequence.[17] The VAE model has been proven to describe the evolution of genes relatively well; however, it uses one-hot encoding for amino acids, treating different kinds of amino acids as completely independent entities.[17] Therefore, this may not be the best way to represent amino acids. Using amino acid embeddings obtained from language models such as ESM to encode the amino acids in multiple sequence alignments, instead of one-hot encoding, and then applying the VAE model for dimensionality reduction to construct the evolutionary landscapes of genes may be a more meaningful approach. This could aid future protein evolutionary studies and provide more reliable estimates of the evolutionary degrees of protein sequences. Moreover, how can the evolution of a set of conserved genes better describe the evolution of the whole genome? With the emergence of language models trained on the whole-genome-scale DNA sequence, such as Evo2,[147] researchers may directly infer the evolutionary relationships among genomes. These are still challenges worth exploring and addressing in future evolutionary studies. Finally, we reiterate that this study is intended as a methodological proof-of-concept rather than a definitive phylogenomic revision. The VAE-PCA landscape is a heuristic exploratory tool; although it recapitulates and expands broad phylogenomic patterns, its evolutionary scores should not be interpreted as absolute divergence times or branch lengths.

If a gene is present and vertically inherited across most archaeal lineages, it may be inferred as potentially part of the LACA proteome. Similarly, if most

genes in a metabolic pathway are present in the LACA proteome, it is plausible that LACA may have possessed that metabolic pathway. Given that archaea are believed to have originated at least four billion years ago,[4–6] homologous proteins between distantly related archaeal lineages may exhibit significant sequence divergence, potentially falling below the detection threshold of sequence alignment methods.[148] This poses a challenge to the study of the LACA proteome. Fortunately, the accuracy of protein structure prediction models such as AlphaFold is continually improving,[148–150] and structural alignment and search tools such as USalign and FoldSeek are becoming increasingly sophisticated.[59,151] The emergence of these deep learning-based tools may be able to meet this challenge.

The field of physiological condition prediction for organisms still faces technical bottlenecks.[152] Deep learning models for predicting the temperature, salinity, and oxygen tolerance of organisms still need improvement. Additionally, developing models to predict other physiological conditions, such as pH tolerance, light requirements, heavy metal ion concentration tolerance, and nutrient needs, is worth exploring.[152] Although these predictions are computationally derived and remain to be confirmed by culture-based studies, they provide a useful framework for exploring how life adapts to individual and combined environmental factors, and could inform the selection and domestication of bioengineering strains based on naturally evolved strategies. In this study, almost all extremely halophilic archaea are oxygen tolerant, and almost all hyperthermophilic archaea are oxygen intolerant. Therefore, it is advisable to domesticate oxygen-tolerant organisms to adapt to hypersaline environments and oxygen-intolerant organisms to adapt to extremely high-temperature environments. Because very few thermophilic archaea can tolerate extreme salinity, adapting to extreme salinity environments with high temperatures for microbes is very challenging. However, selecting a strain such as *JAANXF01* (GCA_018263395) with a pOGT of 47°C and a predicted halophilic protein proportion of 54% or *Methanopyrus* (GCA_000007185) with a pOGT of 97°C and a predicted halophilic protein proportion of 33% (Figure 6G), as the chassis cell for modification and domestication might yield an engineered strain that can tolerate both hypersalinity and extremely high temperatures.

After the evolution of key metabolisms and physiological conditions of archaea and even all other organisms are fully understood, a geological history model of the coevolution of life and the environment of the Earth will be constructed. This will provide a more solid theoretical foundation for studying the origin and evolution of life and their adaptive strategies to environmental changes.

CONCLUSIONS

In this study, we present an integrative VAE–PCA/phylogenomics framework combining deep learning-based evolutionary landscape inference to investigate patterns of early archaeal evolutionary history, using 1,855 representative archaeal genomes from the GTDB R220. The inferred evolutionary landscape revealed distinct diversification trajectories corresponding to DPANN, TACK-Asgard, and Euryarchaeota, offering a complementary data-driven perspective on their early divergence. Consistent with previous hypotheses, metabolic pathway reconstructions indicate that the H_4MPT branch of the Wood–Ljungdahl pathway, together with methyl-reducing methanogenesis, likely emerged in the common ancestor of Euryarchaeota and TACK, and the potential emergence and subsequent duplication of the Mtr complex within Euryarchaeota identified in this study may have contributed to multiple diversification events in methanogenesis. Additionally, comprehensive deep learning models were employed to predict environmental adaptation patterns across the archaeal tree, identifying temperature, oxygen availability, and hypersalinity as major variables associated with niche specialization. Our analyses suggest that temperature may have acted as a persistent selective pressure throughout archaeal evolution, whereas oxygen may have primarily influenced later-branching lineages, and hypersalinity may impose family-specific constraints. Collectively, these findings are consistent with a scenario among several competing hypotheses in which DPANN forms an early-diverging monophyletic group potentially associated with acetate-related metabolism in low-temperature environments, while the last common ancestor of TACK-Asgard and Euryarchaeota may have been a methyl-reducing methanogen possessing the complete Wood–Ljungdahl pathway under high-temperature conditions. Methodologically, this work provides a reusable analytical framework that integrates deep learning with phylogenomics to generate testable hypotheses; biologically, these inferences provide a foundation for future experimental and phylogenomic testing.

REFERENCES

1. Woese C. R. and Fox G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. USA* **74**:5088–5090. DOI:10.1073/pnas.74.11.5088
2. Woese C. R., Magrum L. J. and Fox G. E. (1978). Archaebacteria. *J. Mol. Evol.* **11**:245–251. DOI:10.1007/BF01734485
3. Woese C. R., Kandler O. and Wheelis M. L. (1990). Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87**:4576–4579. DOI:10.1073/pnas.87.12.4576
4. Martin W. F. and Sousa F. L. (2015). Early microbial evolution: The age of anaerobes. *Cold Spring Harb. Perspect. Biol.* **8**:a018127. DOI:10.1101/cshperspect.a018127
5. Ueno Y., Yamada K., Yoshida N., et al. (2006). Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. *Nature* **440**:516–519. DOI:10.1038/nature04584
6. Schopf J. W., Kitajima K., Spicuzza M. J., et al. (2018). SIMS analyses of the oldest known assemblage of microfossils document their taxon-correlated carbon isotope compositions. *Proc. Natl. Acad. Sci. USA* **115**:53–58. DOI:10.1073/pnas.1718063115
7. Baker B. J., De Anda V., Seitz K. W., et al. (2020). Diversity, ecology and evolution of archaea. *Nat. Microbiol.* **5**:887–900. DOI:10.1038/s41564-020-0715-z
8. Tahon G., Geesink P. and Ettema T. J. G. (2021). Expanding archaeal diversity and phylogeny: Past, present, and future. *Annu. Rev. Microbiol.* **75**:359–381. DOI:10.1146/annurev-micro-040921-050212
9. Berghuis B. A., Yu F. B., Schulz F., et al. (2019). Hydrogenotrophic methanogenesis in archaeal phylum Verstraetearchaeota reveals the shared ancestry of all methanogens. *Proc. Natl. Acad. Sci. USA* **116**:5037–5044. DOI:10.1073/pnas.1815631116
10. Borrel G., Adam P. S., McKay L. J., et al. (2019). Wide diversity of methane and short-chain alkane metabolisms in uncultured archaea. *Nat. Microbiol.* **4**:603–613. DOI:10.1038/s41564-019-0363-3
11. Spang A., Caceres E. F. and Ettema T. J. G. (2017). Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science* **357**:eaaf3883. DOI:10.1126/science.aaf3883
12. Rinke C., Chuvochina M., Mussig A. J., et al. (2021). A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat. Microbiol.* **6**:946–959. DOI:10.1038/s41564-021-00918-8
13. Moi D., Bernard C., Steinegger M., et al. (2025). Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. *Nat. Struct. & Mol. Biol.* **32**:2492–2502. DOI:10.1038/s41594-025-01649-8
14. Greener J. G., Moffat L. and Jones D. T. (2018). Design of metalloproteins and novel protein folds using variational autoencoders. *Sci. Rep.* **8**:16189. DOI:10.1038/s41598-018-34533-1
15. Hie B. L., Yang K. K. and Kim P. S. (2022). Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Syst.* **13**:274–285,e276. DOI:10.1016/j.cels.2022.01.003
16. Lu A. X., Yan W., Yang K. K., et al. (2024). Tokenized and continuous embedding compressions of protein sequence and structure. *bioRxiv*:2024.2008.2006.606920. DOI:10.1101/2024.08.06.606920.
17. Detlefsen N. S., Hauberg S. and Boomsma W. (2022). Learning meaningful representations of protein sequences. *Nat. Commun.* **13**:1914. DOI:10.1038/s41467-022-29443-w
18. Ding X., Zou Z. and Brooks C. L. III (2019). Deciphering protein evolution and fitness landscapes with latent space models. *Nat. Commun.* **10**:5644. DOI:10.1038/s41467-019-13633-0
19. Ziegler C., Martin J., Sinner C., et al. (2023). Latent generative landscapes as maps of functional diversity in protein sequence space. *Nat. Commun.* **14**:2222. DOI:10.1038/s41467-023-37958-z
20. Guy L. and Ettema T. J. (2011). The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends Microbiol.* **19**:580–587. DOI:10.1016/j.tim.2011.09.002
21. Rinke C., Schwientek P., Sczyrba A., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**:431–437. DOI:10.1038/nature12352
22. Williams T. A., Szollosi G. J., Spang A., et al. (2017). Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. USA* **114**:E4602–E4611. DOI:10.1073/pnas.1618463114
23. Baker B. A., McCarthy C. G. P., Lopez-Garcia P., et al. (2025). Phylogenomic analyses indicate the archaeal superphylum DPANN originated from free-living euryarchaeal-like ancestors. *Nat. Microbiol.* **10**:1593–1604. DOI:10.1038/s41564-025-02024-5
24. Mei R., Kaneko M., Imachi H., et al. (2023). The origin and evolution of methanogenesis and archaea are intertwined. *PNAS Nexus* **2**:pgad023. DOI:10.1093/pnasnexus/pgad023
25. Garcia P. S., Gribaldo S. and Borrel G. (2022). Diversity and evolution of methane-related pathways in archaea. *Annu. Rev. Microbiol.* **76**:727–755. DOI:10.1146/

annurev-micro-041020-024935
26. Adam P. S., Borrel G. and Gribaldo S. (2019). An archaeal origin of the Wood-Ljungdahl H(4)MPT branch and the emergence of bacterial methylotrophy. *Nat. Microbiol.* **4**:2155–2163. DOI:10.1038/s41564-019-0534-2
27. Berg I. A., Kockelkorn D., Ramos-Vera W. H., et al. (2010). Autotrophic carbon fixation in archaea. *Nat. Rev. Microbiol.* **8**:447–460. DOI:10.1038/nrmicro2365
28. Weiss M. C., Sousa F. L., Mrnjavac N., et al. (2016). The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1**:16116. DOI:10.1038/nmicrobiol.2016.116
29. Adam P. S., Kolyfetis G. E., Bornemann T. L. V., et al. (2022). Genomic remnants of ancestral methanogenesis and hydrogenotrophy in archaea drive anaerobic carbon cycling. *Sci. Adv.* **8**:eabm9651. DOI:10.1126/sciadv.abm9651
30. Wang Y., Wegener G., Williams T. A., et al. (2021). A methylotrophic origin of methanogenesis and early divergence of anaerobic multicarbon alkane metabolism. *Sci. Adv.* **7**:eabj1453. DOI:10.1126/sciadv.abj1453
31. Adam P. S., Borrel G. and Gribaldo S. (2018). Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. *Proc. Natl. Acad. Sci. USA* **115**:E1166–E1173. DOI:10.1073/pnas.1716667115
32. Vulcano F., Hribovsek P., Denny E. O., et al. (2023). Potential for homoacetogenesis via the Wood-Ljungdahl pathway in Korarchaeia lineages from marine hydrothermal vents. *Environ. Microbiol. Rep.* **15**:698–707. DOI:10.1111/1758-2229.13168
33. He Y., Li M., Perumal V., et al. (2016). Genomic and enzymatic evidence for acetogenesis among multiple lineages of the archaeal phylum Bathyarchaeota widespread in marine sediments. *Nat. Microbiol.* **1**:16035. DOI:10.1038/nmicrobiol.2016.35
34. Schmidt M. and Schonheit P. (2013). Acetate formation in the photoheterotrophic bacterium Chloroflexus aurantiacus involves an archaeal type ADP-forming acetyl-CoA synthetase isoenzyme I. *FEMS Microbiol. Lett.* **349**:171–179. DOI:10.1111/1574-6968.12312
35. Jones C. P., Khan K. and Ingram-Smith C. (2017). Investigating the mechanism of ADP-forming acetyl-CoA synthetase from the protozoan parasite Entamoeba histolytica. *FEBS Lett.* **591**:603–612. DOI:10.1002/1873-3468.12573
36. Brasen C., Schmidt M., Grotzinger J., et al. (2008). Reaction mechanism and structural model of ADP-forming acetyl-CoA synthetase from the hyperthermophilic archaeon Pyrococcus furiosus: Evidence for a second active site histidine residue. *J. Biol. Chem.* **283**:15409–15418. DOI:10.1074/jbc.M710218200
37. Ouboter H. T., Arshad A., Berger S., et al. (2023). Acetate and acetyl-CoA metabolism of ANME-2 anaerobic archaeal methanotrophs. *Appl. Environ. Microbiol.* **89**:e0036723. DOI:10.1128/aem.00367-23
38. Kingma D. P. and Welling M. (2013). Auto-encoding variational Bayes. *arXiv*:1312.6114. DOI:10.48550/arXiv.1312.6114
39. Kingma D. P. and Welling M. (2019). An introduction to variational autoencoders. *Found. Trends Mach. Learn.* **12**:4–89. DOI:10.1561/2200000056
40. Zhu M., Song Y., Yuan Q., et al. (2024). Accurately predicting optimal conditions for microorganism proteins through geometric graph learning and language model. *Commun. Biol.* **7**:1709. DOI:10.1038/s42003-024-07436-3
41. Tan P., Li M., Zhang L., et al. (2023). TemPL: A novel deep learning model for zero-shot prediction of protein stability and activity based on temperature-guided language modeling. *arXiv:2304.03780.* DOI:10.48550/arXiv.2304.03780
42. Li G., Rabe K. S., Nielsen J., et al. (2019). Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. *ACS Synth. Biol.* **8**:1411–1420. DOI:10.1021/acssynbio.9b00099
43. Gado J. E., Beckham G. T. and Payne C. M. (2020). Improving enzyme optimum temperature prediction with resampling strategies and ensemble learning. *J. Chem. Inf. Model.* **60**:4098–4107. DOI:10.1021/acs.jcim.0c00489
44. Barnum T. P., Crits-Christoph A., Molla M., et al. (2024). Predicting microbial growth conditions from amino acid composition. *bioRxiv*:2024.2003.2022.586313. DOI:10.1101/2024.03.22.586313
45. Gado J. E., Knotts M., Shaw A. Y., et al. (2025). Machine learning prediction of enzyme optimum pH. *Nat. Mach. Intell.* **7**:716–729. DOI:10.1038/s42256-025-01026-6
46. Hu S., Wang X., Wang Z., et al. (2024). HPClas: A data-driven approach for identifying halophilic proteins based on catBoost. *mLife* **3**:515–526. DOI:10.1002/mlf2.12125
47. Davin A. A., Woodcroft B. J., Soo R. M., et al. (2025). A geological timescale for bacterial evolution and oxygen adaptation. *Science* **388**:eadp1853. DOI:10.1126/science.adp1853
48. Flamholz A. I., Goldford J. E., Richter P. A., et al. (2024). Annotation-free prediction of microbial dioxygen utilization. *mSystems* **9**:e0076324. DOI:10.1128/msystems.00763-24
49. Groussin M., Boussau B., Charles S., et al. (2013). The molecular signal for the adaptation to cold temperature during early life on Earth. *Biol. Lett.* **9**:20130608. DOI:10.1098/rsbl.2013.0608
50. Groussin M., Boussau B. and Gouy M. (2013). A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. *Syst. Biol.* **62**:523–538. DOI:10.1093/sysbio/syt016
51. Hyatt D., Chen G. L., Locascio P. F., et al. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**:119. DOI:10.1186/1471-2105-11-119
52. Camacho C., Coulouris G., Avagyan V., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinform.* **10**:421. DOI:10.1186/1471-2105-10-421
53. Katoh K. and Standley D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**:772–780. DOI:10.1093/molbev/mst010
54. Capella-Gutierrez S., Silla-Martinez J. M. and Gabaldon T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**:1972–1973. DOI:10.1093/bioinformatics/btp348
55. Price M. N., Dehal P. S. and Arkin A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**:e9490. DOI:10.1371/journal.pone.0009490
56. Nguyen L. T., Schmidt H. A., von Haeseler A., et al. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**:268–274. DOI:10.1093/molbev/msu300
57. Kalyaanamoorthy S., Minh B. Q., Wong T. K. F., et al. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods.* **14**:587–589. DOI:10.1038/nmeth.4285
58. Mirdita M., Schutze K., Moriwaki Y., et al. (2022). ColabFold: Making protein folding accessible to all. *Nat. Methods* **19**:679–682. DOI:10.1038/s41592-022-01488-1
59. Zhang C., Shine M., Pyle A. M., et al. (2022). US-align: Universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods* **19**:1109–1115. DOI:10.1038/s41592-022-01585-1
60. Wang Y., Wegener G., Hou J., et al. (2019). Expanding anaerobic alkane metabolism in the domain of Archaea. *Nat. Microbiol.* **4**:595–602. DOI:10.1038/s41564-019-0364-2
61. Leng H., Wang Y., Zhao W., et al. (2023). Identification of a deep-branching thermophilic clade sheds light on early bacterial evolution. *Nat. Commun.* **14**:4354. DOI:10.1038/s41467-023-39960-x
62. Moody E. R. R., Mahendrarajah T. A., Dombrowski N., et al. (2022). An estimate of the deepest branches of the tree of life from ancient vertically evolving genes. *eLife* **11**:e66695. DOI:10.7554/eLife.66695
63. Huang W.-C., Dombrowski N., Mahendrarajah T. A., et al. (2025). Phylogenetic reconciliation supports a methanogenic ancestor of the archaea and a derived origin for host-associated lineages. *bioRxiv*:2025.11.11.687807. DOI:10.1101/2025.11.11.687807
64. Williams T. A. and Heaps S. E. (2014). An introduction to phylogenetics and the tree of life. In *New Approaches to Prokaryotic Systematics*, Goodfellow M., Sutcliffe I., and Chun J. (eds) (Academic Press), pp:13-44. DOI:10.1016/bs.mim.2014.05.001
65. de la Paz J. A., Nartey C. M., Yuvaraj M., et al. (2020). Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proc. Natl. Acad. Sci. USA* **117**:5873–5882. DOI:10.1073/pnas.1913071117
66. Rodriguez-Rivas J., Croce G., Muscat M., et al. (2022). Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. *Proc. Natl. Acad. Sci. USA* **119**:e2113118119. DOI:10.1073/pnas.2113118119
67. Bisardi M., Rodriguez-Rivas J., Zamponi F., et al. (2022). Modeling sequence-space exploration and emergence of epistatic signals in protein evolution. *Mol. Biol. Evol.* **39**:msab321. DOI:10.1093/molbev/msab321
68. Duchene D. A., Mather N., Van Der Wal C., et al. (2022). Excluding loci with substitution saturation improves inferences from phylogenomic data. *Syst. Biol.* **71**:676–689. DOI:10.1093/sysbio/syab075
69. Jermiin L. S. and Misof B. (2020). Measuring historical and compositional signals in phylogenetic data. *bioRxiv*:2020.2001.2003.894097. DOI:10.1101/2020.01.03.894097
70. Philippe H., Brinkmann H., Lavrov D. V., et al. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLOS Biol.* **9**:e1000602. DOI:10.1371/journal.pbio.1000602
71. Manuel C., Sakalli E., Schmidt H. A., et al. (2025). When the past fades: Detecting phylogenetic signal with SatuTe. *Mol. Biol. Evol.* **42**:msaf090. DOI:10.1093/molbev/msaf090
72. Felsenstein J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**:401–410. DOI:10.2307/2412923
73. Susko E. and Roger A. J. (2021). Long branch attraction biases in phylogenetics. *Syst. Biol.* **70**:838–843. DOI:10.1093/sysbio/syab001
74. Roch S., Nute M. and Warnow T. (2019). Long-branch attraction in species tree estimation: Inconsistency of partitioned likelihood and topology-based summary methods. *Syst. Biol.* **68**:281–297. DOI:10.1093/sysbio/syy061
75. Ingram-Smith C., Woods B. I. and Smith K. S. (2006). Characterization of the acyl substrate binding pocket of acetyl-CoA synthetase. *Biochemistry* **45**:11482–11490. DOI:10.1021/bi061023e
76. Musfeldt M. and Schonheit P. (2002). Novel type of ADP-forming acetyl coenzyme A synthetase in hyperthermophilic archaea: Heterologous expression and characterization of isoenzymes from the sulfate reducer Archaeoglobus fulgidus and the methanogen Methanococcus jannaschii. *J. Bacteriol.* **184**:636–644. DOI:10.1128/JB.184.3.636-644.2002
77. Brasen C. and Schonheit P. (2005). AMP-forming acetyl-CoA synthetase from the extremely halophilic archaeon Haloarcula marismortui: Purification, identification and expression of the encoding gene, and phylogenetic affiliation. *Extremophiles* **9**:355–365. DOI:10.1007/s00792-005-0449-0

78. Liang M. H., Qv X. Y., Jin H. H., et al. (2016). Characterization and expression of AMP-forming acetyl-CoA synthetase from Dunaliella tertiolecta and its response to nitrogen starvation stress. *Sci. Rep.* **6**:23445. DOI:10.1038/srep23445
79. Kuprat T., Ortjohann M., Johnsen U., et al. (2021). Glucose metabolism and acetate switch in archaea: The enzymes in Haloferax volcanii. *J. Bacteriol.* **203**:e00690–20. DOI:10.1128/JB.00690-20
80. Fournier G. P. and Gogarten J. P. (2008). Evolution of acetoclastic methanogenesis in Methanosarcina via horizontal gene transfer from cellulolytic Clostridia. *J. Bacteriol.* **190**:1124–1127. DOI:10.1128/JB.01382-07
81. Rothman D. H., Fournier G. P., French K. L., et al. (2014). Methanogenic burst in the end-Permian carbon cycle. *Proc. Natl. Acad. Sci. USA* **111**:5462–5467. DOI:10.1073/pnas.1318106111
82. He C., Keren R., Whittaker M. L., et al. (2021). Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nat. Microbiol.* **6**:354–365. DOI:10.1038/s41564-020-00840-5
83. Gong W., Hao B., Wei Z., et al. (2008). Structure of the alpha2epsilon2 Ni-dependent CO dehydrogenase component of the Methanosarcina barkeri acetyl-CoA decarbonylase/synthase complex. *Proc. Natl. Acad. Sci. USA* **105**:9558–9563. DOI:10.1073/pnas.0800415105
84. Schwank K., Bornemann T. L. V., Dombrowski N., et al. (2019). An archaeal symbiont-host association from the deep terrestrial subsurface. *ISME J.* **13**:2135–2139. DOI:10.1038/s41396-019-0421-0
85. Probst A. J. and Moissl-Eichinger C. (2015). "Altiarchaeales": Uncultivated archaea from the subsurface. *Life (Basel)* **5**:1381–1395. DOI:10.3390/life5021381
86. Weiss M. C., Preiner M., Xavier J. C., et al. (2018). The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLOS Genet.* **14**:e1007518. DOI:10.1371/journal.pgen.1007518
87. Kurth J. M., Nobu M. K., Tamaki H., et al. (2021). Methanogenic archaea use a bacteria-like methyltransferase system to demethoxylate aromatic compounds. *ISME J.* **15**:3549–3565. DOI:10.1038/s41396-021-01025-6
88. Bueno de Mesquita C. P., Wu D. and Tringe S. G. (2023). Methyl-based methanogenesis: An ecological and genomic review. *Microbiol. Mol. Biol. Rev.* **87**:e0002422. DOI:10.1128/mmbr.00024-22
89. Dombrowski N., Teske A. P. and Baker B. J. (2018). Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. *Nat. Commun.* **9**:4999. DOI:10.1038/s41467-018-07418-0
90. Hoedt E. C., Parks D. H., Volmer J. G., et al. (2018). Culture- and metagenomics-enabled analyses of the Methanosphaera genus reveals their monophyletic origin and differentiation according to genome size. *ISME J.* **12**:2942–2953. DOI:10.1038/s41396-018-0225-7
91. Sprenger W. W., Hackstein J. H. and Keltjens J. T. (2007). The competitive success of Methanomicrococcus blatticola, a dominant methylotrophic methanogen in the cockroach hindgut, is supported by high substrate affinities and favorable thermodynamics. *FEMS Microbiol. Ecol.* **60**:266–275. DOI:10.1111/j.1574-6941.2007.00287.x
92. Thomas C. M., Taib N., Gribaldo S., et al. (2021). Comparative genomic analysis of Methanimicrococcus blatticola provides insights into host adaptation in archaea and the evolution of methanogenesis. *ISME Commun.* **1**:47. DOI:10.1038/s43705-021-00050-y
93. Preiner M., Xavier J. C., Sousa F. L., et al. (2018). Serpentinization: Connecting geochemistry, ancient metabolism and industrial hydrogenation. *Life (Basel)* **8**:41. DOI:10.3390/life8040041
94. Belthle K. S., Martin W. F. and Tuysuz H. (2024). Synergistic effects of silica-supported iron-cobalt catalysts for CO(2) reduction to prebiotic organics. *ChemCatChem* **16**:e202301218. DOI:10.1002/cctc.202301218
95. Preiner M., Igarashi K., Muchowska K. B., et al. (2020). A hydrogen-dependent geochemical analogue of primordial carbon and energy metabolism. *Nat. Ecol. Evol.* **4**:534–542. DOI:10.1038/s41559-020-1125-6
96. Zhou Z., Zhang C. J., Liu P. F., et al. (2022). Non-syntrophic methanogenic hydrocarbon degradation by an archaeal species. *Nature* **601**:257–262. DOI:10.1038/s41586-021-04235-2
97. Hahn C. J., Laso-Perez R., Vulcano F., et al. (2020). "Candidatus Ethanoperedens," a thermophilic genus of archaea mediating the anaerobic oxidation of ethane. *mBio* **11**:e00600–20. DOI:10.1128/mBio.00600-20
98. Kohtz A. J., Nupp S. and Hatzenpichler R. (2025). Cultivation of Methanonezhaarchaeia, the third class of methanogens within the phylum Thermoproteota. *Sci. Adv.* **11**:eaea0936. DOI:10.1126/sciadv.aea0936
99. Spang A., Stairs C. W., Dombrowski N., 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
100. Rao Y. Z., Li Y. X., Li Z. W., et al. (2026). Horizontal gene transfer and gene loss drove the divergent evolution of host dependency in Micrarchaeota. *Natl. Sci. Rev.* **13**:nwaf542. DOI:10.1093/nsr/nwaf542
101. Sakai H. D., Nur N., Kato S., et al. (2022). Insight into the symbiotic lifestyle of DPANN archaea revealed by cultivation and genome analyses. *Proc. Natl. Acad. Sci. USA* **119**:e2115449119. DOI:10.1073/pnas.2115449119
102. Reysenbach A. L. (2015). Thermoprotei class. nov. In Bergey's manual of systematics of archaea and bacteria, Whitman W. B. (ed) (Wiley), pp:1-1. DOI:10.1002/9781118960608.cbm00018
103. Hou J., Wang Y., Zhu P., et al. (2023). Taxonomic and carbon metabolic diversification of Bathyarchaeia during its coevolution history with early Earth surface environment. *Sci. Adv.* **9**:eadf5069. DOI:10.1126/sciadv.adf5069
104. Qi Y. L., Evans P. N., Li Y. X., et al. (2021). Comparative genomics reveals thermal adaptation and a high metabolic diversity in "Candidatus Bathyarchaeia". *mSystems* **6**:e0025221. DOI:10.1128/mSystems.00252-21
105. Kerou M., Eloy Alves R. J. and Schleper C. (2016). Nitrososphaeria. In Bergey's manual of systematics of archaea and bacteria, Whitman W. B. (ed) (Wiley), pp:1-8. DOI:10.1002/9781118960608.cbm00055.
106. Ren M. and Wang J. (2022). Phylogenetic divergence and adaptation of Nitrososphaeria across lake depths and freshwater ecosystems. *ISME J.* **16**:1491–1501. DOI:10.1038/s41396-022-01199-7
107. Ren M., Feng X., Huang Y., et al. (2019). Phylogenomics suggests oxygen availability as a driving force in Thaumarchaeota evolution. *ISME J.* **13**:2150–2161. DOI:10.1038/s41396-019-0418-8
108. Oren A., Ventosa A. and Kamekura M. (2017). Halobacteria. In Bergey's manual of systematics of archaea and bacteria, Whitman W. B. (ed) (Wiley), pp:1-5. DOI:10.1002/9781118960608.cbm00026.pub2.
109. Baker B. A., Gutierrez-Preciado A., Rodriguez Del Rio A., et al. (2024). Expanded phylogeny of extremely halophilic archaea shows multiple independent adaptations to hypersaline environments. *Nat. Microbiol.* **9**:964–975. DOI:10.1038/s41564-024-01647-4
110. Oren A. (2002). Diversity of halophilic microorganisms: Environments, phylogeny, physiology, and applications. *J. Ind. Microbiol. Biotechnol.* **28**:56–63. DOI:10.1038/sj/jim/7000176
111. Boussau B., Blanquart S., Necsulea A., et al. (2008). Parallel adaptations to high temperatures in the Archaean eon. *Nature* **456**:942–945. DOI:10.1038/nature07393
112. Jablonska J. and Tawfik D. S. (2021). The evolution of oxygen-utilizing enzymes suggests early biosphere oxygenation. *Nat. Ecol. Evol.* **5**:442–448. DOI:10.1038/s41559-020-01386-9
113. Lalonde S. V. and Konhauser K. O. (2015). Benthic perspective on Earth's oldest evidence for oxygenic photosynthesis. *Proc. Natl. Acad. Sci. USA* **112**:995–1000. DOI:10.1073/pnas.1415718112
114. Olson S. L., Kump L. R. and Kasting J. F. (2013). Quantifying the areal extent and dissolved oxygen concentrations of Archean oxygen oases. *Chem. Geol.* **362**:35–43. DOI:10.1016/j.chemgeo.2013.08.012
115. Forterre P. (2002). A hot story from comparative genomics: Reverse gyrase is the only hyperthermophile-specific protein. *Trends Genet.* **18**:236–237. DOI:10.1016/s0168-9525(02)02650-1
116. Forterre P., Mirambeau G., Jaxel C., et al. (1985). High positive supercoiling in vitro catalyzed by an ATP and polyethylene glycol-stimulated topoisomerase from Sulfolobus acidocaldarius. *EMBO J.* **4**:2123–2128. DOI:10.1002/j.1460-2075.1985.tb03902.x
117. Kikuchi A. and Asai K. (1984). Reverse gyrase--a topoisomerase which introduces positive superhelical turns into DNA. *Nature* **309**:677–681. DOI:10.1038/309677a0.
118. Brochier-Armanet C. and Forterre P. (2007). Widespread distribution of archaeal reverse gyrase in thermophilic bacteria suggests a complex history of vertical inheritance and lateral gene transfers. *Archaea* **2**:83–93. DOI:10.1155/2006/582916
119. Catchpole R. J. and Forterre P. (2019). The evolution of reverse gyrase suggests a nonhyperthermophilic last universal common ancestor. *Mol. Biol. Evol.* **36**:2737–2747. DOI:10.1093/molbev/msz180
120. Jablonska J. and Tawfik D. S. (2019). The number and type of oxygen-utilizing enzymes indicates aerobic vs. *anaerobic phenotype. Free Radic. Biol. Med.* **140**:84–92. DOI:10.1016/j.freeradbiomed.2019.03.031
121. Sibbald S. J., Eme L., Archibald J. M., et al. (2020). Lateral gene transfer mechanisms and pan-genomes in eukaryotes. *Trends Parasitol.* **36**:927–941. DOI:10.1016/j.pt.2020.07.014
122. Soucy S. M., Huang J. and Gogarten J. P. (2015). Horizontal gene transfer: Building the web of life. *Nat. Rev. Genet.* **16**:472–482. DOI:10.1038/nrg3962
123. Stairs C. W., Eme L., Munoz-Gomez S. A., et al. (2018). Microbial eukaryotes have adapted to hypoxia by horizontal acquisitions of a gene involved in rhodoquinone biosynthesis. *eLife* **7**:e34292. DOI:10.7554/eLife.34292
124. Zhang J. and Yang J. R. (2015). Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* **16**:409–420. DOI:10.1038/nrg3950
125. Yang J. R., Liao B. Y., Zhuang S. M., et al. (2012). Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc. Natl. Acad. Sci. USA* **109**:E831–840. DOI:10.1073/pnas.1117408109
126. Echave J., Spielman S. J. and Wilke C. O. (2016). Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.* **17**:109–121. DOI:10.1038/nrg.2015.18
127. Scanlan D. J., Ostrowski M., Mazard S., et al. (2009). Ecological genomics of marine picocyanobacteria. *Microbiol. Mol. Biol. Rev.* **73**:249–299. DOI:10.1128/MMBR.00035-08
128. Becker E. A., Seitzer P. M., Tritt A., et al. (2014). Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response. *PLOS Genet.* **10**:e1004784. DOI:10.1371/journal.pgen.1004784

129. Ionescu D., Zoccarato L., Cabello-Yeves P. J., et al. (2023). Extreme fluctuations in ambient salinity select for bacteria with a hybrid "salt-in"/"salt-out" osmoregulation strategy. *Front. Microbiomes* **2**:1329925. DOI:10.3389/frmbi.2023.1329925
130. Ghai R., Pasic L., Fernandez A. B., et al. (2011). New abundant microbial groups in aquatic hypersaline environments. *Sci. Rep.* **1**:135. DOI:10.1038/srep00135
131. Narasingarao P., Podell S., Ugalde J. A., et al. (2012). De novo metagenomic assembly reveals abundant novel major lineage of archaea in hypersaline microbial communities. *ISME J.* **6**:81–93. DOI:10.1038/ismej.2011.78
132. Zhao D., Zhang S., Kumar S., et al. (2022). Comparative genomic insights into the evolution of Halobacteria-associated "Candidatus Nanohaloarchaeota". *mSystems* **7**:e0066922. DOI:10.1128/msystems.00669-22
133. Sorokin D. Y., Makarova K. S., Abbas B., et al. (2017). Discovery of extremely halophilic, methyl-reducing euryarchaea provides insights into the evolutionary origin of methanogenesis. *Nat. Microbiol.* **2**:17081. DOI:10.1038/nmicrobiol.2017.81
134. Zhou H., Zhao D., Zhang S., et al. (2022). Metagenomic insights into the environmental adaptation and metabolism of Candidatus Haloplasmatales, one archaeal order thriving in saline lakes. *Environ. Microbiol.* **24**:2239–2258. DOI:10.1111/1462-2920.15899
135. Michoud G., Ngugi D. K., Barozzi A., et al. (2021). Fine-scale metabolic discontinuity in a stratified prokaryote microbiome of a Red Sea deep halocline. *ISME J.* **15**:2351–2365. DOI:10.1038/s41396-021-00931-z
136. Roesser M. and Muller V. (2001). Osmoadaptation in bacteria and archaea: Common principles and differences. *Environ. Microbiol.* **3**:743–754. DOI:10.1046/j.1462-2920.2001.00252.x
137. Protasov E., Nonoh J. O., Kastle Silva J. M., et al. (2023). Diversity and taxonomic revision of methanogens and other archaea in the intestinal tract of terrestrial arthropods. *Front. Microbiol.* **14**:1281628. DOI:10.3389/fmicb.2023.1281628
138. Mazurie A., Bonchev D., Schwikowski B., et al. (2010). Evolution of metabolic network organization. *BMC Syst. Biol.* **4**:59. DOI:10.1186/1752-0509-4-59
139. Raymond J. and Segre D. (2006). The effect of oxygen on biochemical networks and the evolution of complex life. *Science* **311**:1764–1767. DOI:10.1126/science.1118439
140. Takemoto K. and Yoshitake I. (2013). Limited influence of oxygen on the evolution of chemical diversity in metabolic networks. *Metabolites* **3**:979–992. DOI:10.3390/metabo3040979
141. Martemucci G., Costagliola C., Mariano M., et al. (2022). Free radical properties, source and targets, antioxidant consumption and health. *Oxygen* **2**:48–78. DOI:10.3390/oxygen2020006
142. Davies M. J. (2005). The oxidative environment and protein damage. *Biochim. Biophys. Acta* **1703**:93–109. DOI:10.1016/j.bbapap.2004.08.007
143. Carlton J. D., Langwig M. V., Gong X., et al. (2023). Expansion of Armatimonadota through marine sediment sequencing describes two classes with unique ecological roles. *ISME Commun.* **3**:64. DOI:10.1038/s43705-023-00269-x
144. Liang C., Yang B., Cao Y. C., et al. (2024). Salinization mechanism of lakes and controls on organic matter enrichment: From present to deep-time records. *Earth-Sci. Rev.* **251**:104720. DOI:10.1016/j.earscirev.2024.104720
145. Swan B. K., Tupper B., Sczyrba A., et al. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl. Acad. Sci. USA* **110**:11463–11468. DOI:10.1073/pnas.1304246110
146. Giovannoni S. J., Tripp H. J., Givan S., et al. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**:1242–1245. DOI:10.1126/science.1114057
147. Brixi G., Durrant M. G., Ku J., et al. (2025). Genome modeling and design across all domains of life with Evo 2. *bioRxiv*:2025.2002.2018.638918. DOI:10.1101/2025.02.18.638918
148. Hamamsy T., Morton J. T., Blackwell R., et al. (2024). Protein remote homology detection and structural alignment using deep learning. *Nat. Biotechnol.* **42**:975–985. DOI:10.1038/s41587-023-01917-2
149. Varadi M., Bertoni D., Magana P., et al. (2024). AlphaFold Protein Structure Database in 2024: Providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**:D368–D375. DOI:10.1093/nar/gkad1011
150. Jumper J., Evans R., Pritzel A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596**:583–589. DOI:10.1038/s41586-021-03819-2
151. van Kempen M., Kim S. S., Tumescheit C., et al. (2024). Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**:243–246. DOI:10.1038/s41587-023-01773-0
152. Ramoneda J., Hoffert M., Stallard-Olivera E., et al. (2024). Leveraging genomic information to predict environmental preferences of bacteria. *ISME J.* **18**:wrae195. DOI:10.1093/ismejo/wrae195

FUNDING AND ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (NSFC) grants (42422209, 42272354, 42230401 and 42141003) and the National Key Research and Development Program of China for funding (2023YFC3108600).We acknowledge the Genome Taxonomy Database (GTDB, https://gtdb.ecogenomic.org) and National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/) that contribute genomic data for this work. We are grateful for Bingxin Zhou, Yang Tan, Bozitao Zhong, Shuai Li, Dian Ding, Mei Wang, Jiahao Guan, Zichao Zeng, Liuyang Li for useful discussion. The computations in this paper were partially run on our lab, also the Siyuan-1 cluster supported by Chen Li, Xingze Wang, Wu Guo from the Center for High Performance Computing at Shanghai Jiao Tong University.The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Z.L., Y.W. conceived and designed the study, and wrote the manuscript. Z.L. collected genomic data and performed evolutionary analysis. W.F. offered useful suggestions and wrote this manuscript. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ETHICAL STATEMENT AND PATIENT CONSENT

Not applicable.

DATA AND CODE AVAILABILITY

Genomes analyzed in this manuscript are available in GTDB and NCBI. Custom Python scripts for the VAE-PCA framework and Supplemental Data are available at https://doi.org/10.6084/m9.figshare.31849165. All the other data is included in the manuscript and supporting information.

SUPPLEMENTAL INFORMATION

It can be found online at https://doi.org/10.59717/j.xinn-life.2026.100235.