Master Internship Report

Exploring the phylogenetic plausibility of expanding metabolic networks

Dimitra Astra Demertzi

Under the supervision of
prof. dr. Shawn E. McGlynn
and *prof. dr. Wilhelm T.S. Huck*

Earth-Life Science Institute | Tokyo Institute of Technology
Faculty of Science | Department of Physical Organic Chemistry
2024

Contents

1 Introduction

Outline

- OF, OFs, orthologs, paralogs, etc
- evolutionary models: DTL vs DLC
- HMMs & advantage of position-independent evolutionary models
- Metabolic network expansion -> more? what is it, what are the models, etc.
- Metabolic network expansion traditionally used to infer evolution of metabolism. Study objective. Essentially connecting phylogenetic analysis with met. net. exp.

A surge in available data

The advancement of sequencing technologies and the development of state-of-the-art bioinformatics tools have enabled a rapid increase not only in the number of sequenced genomes, but also in the sampled prokaryotic diversity, expanding our view of the tree of life, and especially that of the archaeal branch. This surge in data availability can readily improve the resolution of modern phylogenetic relationships, resolve further their deep evolutionary history, and provide more accurate inference of ancestral gene contents [1]. The first release of GTDB back in 2019, R89, contained 145,904 genomes organized into 24,706 species clusters, while its latest, R220 release this April, has grown to contain 596,859 genomes organized into 113,104 species clusters [2–5]; an increase of more than 300% in just five years.

To date, several studies have investigated the nature and physiology of LACA, LBCA, and LUCA (Last Universal Common Ancestor) through gene content estimation [6–9], with datasets of various sizes and microbial physiologies, originating from a number of different databases. Usually these datasets are optimized for tree rooting and take into account taxon sampling evenness, yet it is still unclear how the incorporation of more species or other species belonging to the same higher taxonomic level will affect the analysis.

Moreover, GTDB, in its effort to standardize prokaryotic taxonomy, employs rank-normalization for the classification of microorganisms at all taxonomic levels [2], which leads to periodic reclassification and renaming of taxa in the database. The change in nomenclature and the constant taxonomic revisions create confusion even when referencing taxa from the same database, and hinder the refinement of deep phylogenetic relationships [1]. Especially in an era of exponential growth in genomic data, there is a pressing need to determine the robustness of standard methodologies and result stability.

2 Methods

Data Retrieval & Datasets Establishment

We obtained the amino acid FASTA files containing all protein-encoding gene sequences predicted with Prodigal [10] for all representative archaeal and bacterial species, from the Genome Taxonomy Database (GTDB, release R08-RS214, accessed in March 2024) [2–5]

As of March 2024, the database contained 402,709 genomes organized into 85,205 species clusters, with one representative genome of each named species, according to Parks et al. (2020) [3]. We kept only the genomes which met the high quality completeness (≥ 90%) and low contamination (≤ 5%) criteria, as defined by MIMAG [11]. Out of the 85,205 (4,416 archaeal and 80,789 bacterial) representative (REP) genomes, 52,614 (1,785 archaeal and 50,829 bacterial) genomes met these criteria – henceforth called high quality (HQ) genomes.

Datasets were generated for the phylum, class, order, family, and genus taxonomic ranks for the domains of archaea and bacteria separately, both for the REP and HQ genomes. Per rank, a single representative genome was randomly selected with the pandas function .sample() of the random module. Besides the domain-specific datasets, merged ones were also created for the phylum, class, and order ranks by simple combination of the previously generated folder directories, to perform a phylogenetic analysis for the entire prokaryotic tree of life. Table 1 shows the distribution of genomes across taxonomic levels for both archaea and bacteria domains, which reflects dataset sizes.

Table 1: Distribution of representative (REP) and high quality (HQ) genomes across taxonomic levels for both archaea and bacteria domains. Dataset size corresponds to the number of genomes in that taxonomic level.

	Taxonomic Level					
	Phylum	Class	Order	Family	Genus	Species
Archaea						
REP genomes	21	63	149	509	1586	4416
HQ genomes	**16**	**51**	**94**	**206**	**582**	**1785**
HQ genome proportion*	76.2%	80.9%	63.1%	40.5%	36.7%	40.4%
Bacteria						
Rep genomes	181	490	1653	4305	19153	80789
HQ genomes	**150**	**374**	**1139**	**2787**	**12086**	**50089**
HQ genome proportion	82.9%	76.3%	68.9%	64.8%	63.1%	62.0%

* Percentage of high quality genomes within representative genomes.

Ancestral Gene Content Reconstruction

A phylogenetic analysis of datasets presented in Table 2 was performed with OrthoFinder v2.5.5 [12, 13], which employs a number of specialized phylogenetic tools, such as DIAMOND v2.1.9 [14],

Table 2: OrthoFinder phylogenetic analysis performed for the following datasets.

Taxonomy Level	**Domain**		
	Archaea	Bacteria	All
Phylum	REP | HQ	HQ	HQ
Class	HQ	HQ	HQ
Order	HQ		
Family	HQ		
Genus	HQ		
Species	HQ		

MCL v22.282 [15], ETE v3.1.3 [16], running on default parameters. After the initial construction of orthogroups (gene families) for each respective dataset, gene tree reconciliation and ancestral gene content were determined with the GTDB species trees per domain, instead of the rooted species tree computed by OrthoFinder. The superfluous GTDB branches were pruned with Biopython's Phylo module to match the dataset topology. The smaller-sized datasets were run locally, while the rest were run by the Kyoto Supercomputer. The genome accession numbers constituting each dataset can be found in the supplementary material (Table x).

Gene Annotation using EggNOG-mapper

A number of ancestral genomes were selected for further analysis based on the species tree topology of the smallest dataset for the archaea domain, as shown in Table ref nodes per dataset. For bacteria, metabolic networks were reconstructed only for the Last Bacterial Common Ancestor (LBCA).

The initial functional annotation and KEGG orthology (KO) number assignment were performed by eggNOG-mapper v2.1.12 [17] with default settings, using the eggNOG 5.0 database [18] on the medoid sequence; the sequence having the shortest genetic distance to all other sequences in the group. This was calculated under the BLOSUM62 substitution matrix with the .align function from Biopython's Align module, for maximum 100 sequences per orthogroup. For large orthogroups, a hundred sequences were selected at random with the .sample function of the random module.

The annotation was performed twice per ancestral genome, against the domain-specific and prokaryote eggNOG databases. For each shared hit, priority was given to the hit with the higher bit-score. For shared hits with identical bit-scores, the hit with the lower e-value was selected, and in case both bit-scores and e-values were identical, the hit from the prokaryote eggNOG database was selected. All hits unique to one of the two database runs were kept for downstream analysis.

Gene Annotation using HMM profiles

The LACA inferred genome (N0) for the phylum-level archaea dataset was selected to be annotated with Hidden Markov Model (HMM) profiles. Each orthogroup initially underwent multiple

sequence alignment (MSA) with MAFFT v7.525 [19] using the L-INS-i iterative refinement method with local alignment. The resulting MSA was used to build an HMM profile with HMMER v3.4 hmmbuild program. It was then searched against the UniprotKB database [20] for archaeal proteins (accessed in June 2024) with the hmmsearch program.

Metabolic Network Reconstruction

To reconstruct the metabolic network for each chosen ancestral genome, we utilized a database compiled by Goldford et al. (2024) [21], containing elementally consistent biochemical reactions from the Kyoto Encyclopedia of Genes and Genomes (KEGG), with added detailed organic and inorganic cofactor dependencies gathered from various other databases. The KEGG reaction IDs obtained from eggNOG-mapper were then mapped to the reactions in the database. Reaction reversibility was determined with the eQuilibrator python API [22] using the following parameters: pH = 7.0, pMg (magnesium ion concentration) = 3.0, ionic strength of 0.25 M, temperature T = 298.15 K, and metabolite concentrations between 0.01 and 10 mM. All forward or reverse reactions with minimum reaction free energy above zero were removed from the network. A reaction was kept reversible if no free energy estimate was available.

Seed Sets Generation

All seed sets utilized in this study are available in the data/seedsets folder of a dedicated github repository (Metabolic Network Expansion Report). We utilized the seed set created by Goldford et al. (2024) [21], consisting of various metal species and elemental sources of phosphorus, nitrogen, sulfur, oxygen, hydrogen and carbon found in KEGG, as a basis for all our generated seed sets. With the prebiotic soup concept in mind, we expanded the seed set by adding compounds found in the Murchison meteorite and the Miller-Urey spark-discharge-like experiments, for which KEGG IDs were available, according to Vincent et al. (2021) [23]. For the rest of our seedsets, we utilized the Bertz complexity metric computed for 3,588 compounds by Goldford et al. (2024) for all KEGG module compounds, and filtered the list to create seed sets of increasing molecular complexity. The KEGG module compound data were retrieved using the TogoWS REST service (hrefhttp://togows.dbcls.jphttp://togows.dbcls.jp).

Network expansion algorithm

The network expansion algorithm was implemented using the BioXP python package with some modifications, written initially by Harrison B. Smith and colleagues (https://github.com/hbsmith/BioXP). *move this sentence to introduction: Briefly, seed compounds are allowed to react given the reactions in the network, which then produce product compounds. The product compounds are added to the seed set, and the process is repeated until convergence [24, 25].

3 Results & Discussion

Comparative Genomics Analysis

To identify genes tracing back to the last common ancestors of archaea (LACA) and bacteria (LBCA), we started from 85,205 genomes, each one the representative of species clusters, as defined by the Genome Taxonomy Database (GTDB) [2–5].

Genome completeness can lead to false-positive ortholog and erroneous species phylogeny inference [26]. To ensure that phylogenetic relationships and ancestral genome reconstruction would be as accurate as possible, we filtered the genomes based on completeness and contamination cutoffs of ≥ 90% and ≤ 5%, respectively. This resulted in 52,614 high-quality (HQ) genomes, 1,785 archaeal and 50,829 bacterial, spanning 16 archaeal and 150 bacterial phyla.

To evaluate the effect of taxon sampling and dataset size on ancestral gene content inference, we constructed eight genomic datasets of various sizes (16-1139 genomes). These span the phylum, class, order, family, and genus taxonomic levels for archaea, and at the phylum, class, and order taxonomic levels for bacteria. Dataset sizes can be found in table 1. Even though increasing taxon sampling has been shown to improve phylogenetic accuracy, taxon evenness markedly affects tree topology [27, 28]. To capture, therefore, the full extant phylogenetic diversity and keep the taxon sampling even and unbiased while rendering the largest of our analyses computationally tractable, we selected a single genome for each taxon, per genomic dataset/taxonomic rank in our study.

To review the robustness of random taxon sampling we performed our analysis for three separately produced datasets both for the phylum and class taxonomic levels of archaea. Figure 1A shows the average percentage of genes assigned to orthogroups (OGs) for the two aforementioned triplicate runs, while 1C depicts the mean percentage of gene assignment in OGs for each of the archaea datasets. This statistic can act as a first-level quality control of an OrthoFinder (OF) analysis. A below 80% mean assignment of genes in orthogroups indicates that important orthology relationships for some remaining genes are missing, likely due to poor species sampling [13]. Even though the number of included-in-the-analysis genomes approximately doubles per taxonomic level, the gene assignment percentage only slightly increases. The increase is justified because, as taxonomic levels become narrower, genetic distances between species decrease, allowing state-of-the-art algorithms to cluster a larger number of genes.

On a first assessment, performing a phylogenetic analysis on 51 archaea genomes (class taxonomic level) seems enough to capture the orthology relationships of protein-encoding genes. Plots 1A and 1C, however, present the average percentage of three individual OF runs, and the mean percentage of all individual species/taxa in a single OF run, respectively. They therefore conceal variance that is crucial for the analysis interpretation. To address this, we converted the percentage of genes assigned in OGs per species to a binary classification, where species with less than 80% gene assignment were designated a "poorly sampled" status and given a value equal to zero, while the rest were given a value of one. The normalized distribution of poorly sampled taxa for each taxonomic level analysis can be seen in Figure 1D.

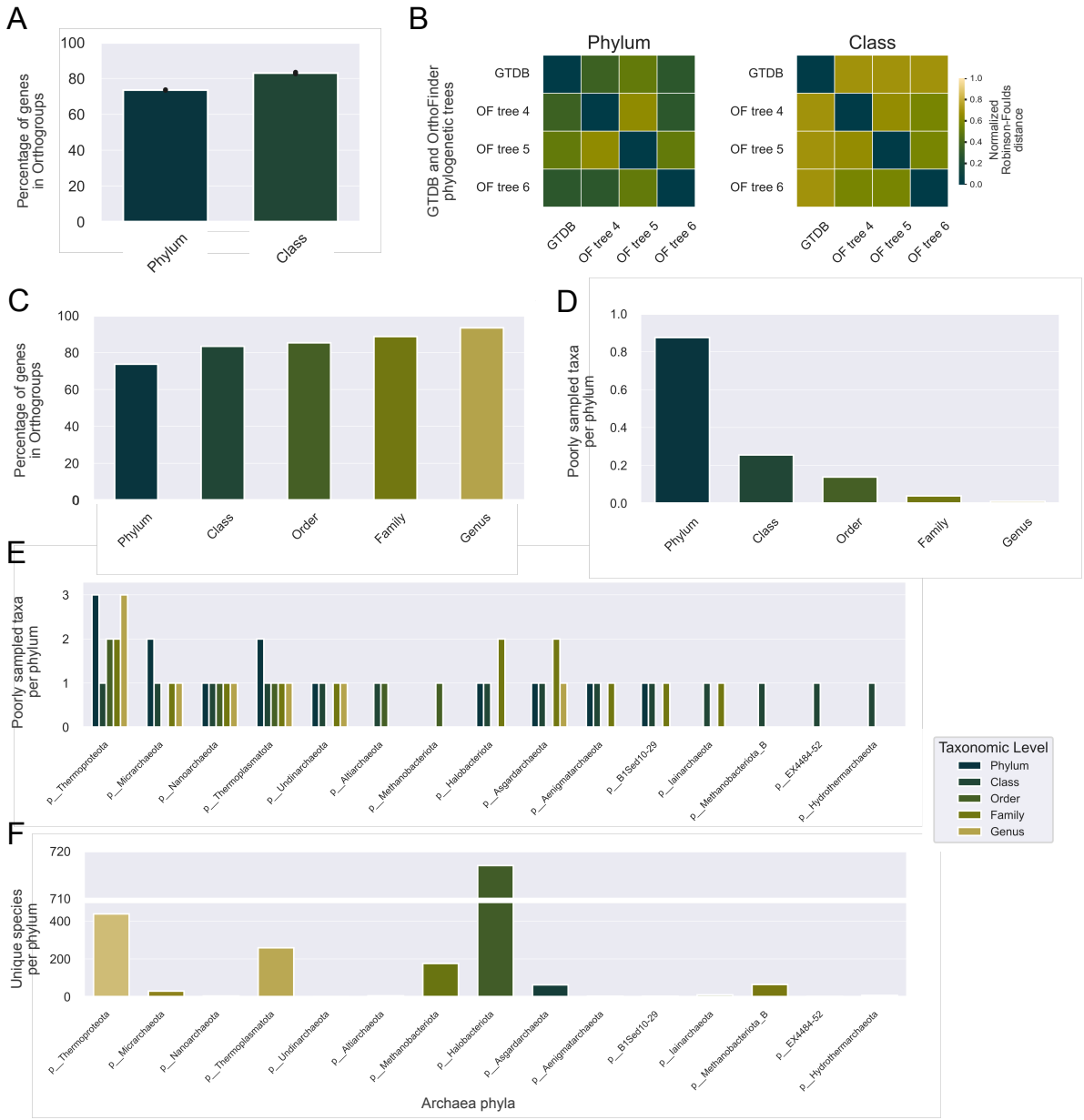


Figure 1: Comparative genomics statistics of phylogenetic OrthoFinder (OF) analysis. (A) Shows the average percentage of genes assigned in orthogroups for the taxonomic levels of phylum and class for archaea. (B) Presents the Robinson-Foulds (RF) distance between the trees generated by OF as part of the aforementioned analyses, as well as their RF from the GTDB archaeal tree. (C) Compares the percentage of genes assigned in orthogroups for every archaeal taxonomic level. (D) Presents the normalized distribution of poorly sampled taxa for each taxonomic level analysis. (E) Detailed distribution of poorly sampled taxa for each taxonomic level analysis, per phylum, and (F) the number of unique species per phylum taxon.

A closer inspection of the statistics for individual taxa (Figure 1D and 1E) revealed that the percentage of poorly sampled genomes dropped below 5% only at the family level (1D). Interestingly, even though genomes belonging to particular phyla appear to be consistently categorized as "poorly sampled", this behavior cannot be linked to the number of species belonging to the same taxon. The Thermoproteota phylum, for example, has the highest count of "poorly sampled" species

across all datasets, yet is the second richest archaeal phylum. Nanoarchaeota, on the other hand, whose genes are also consistently left out of orthogroups, comprise only of 4 species (Figure 1F). One has to take into account that species distribution remains uneven throughout the taxonomic levels, with some classes, orders, families, and genera that belong to the same phylum enriched more than others. Yet, with the decrease in genetic distance between species as datasets grow, and taxonomic levels become narrower, we would expect richer phyla to have a higher assignment of genes to OGs in narrower taxonomic levels.

An initial comparison of the species trees produced by OF's STAG [29] and STRIDE [30] en-suite algorithms in Dendroscope [31] (found in Appendix Figures A1 - A4) exposed unique tree topologies. Because of the significant missing gene orthology relationships at the phylum and class level, this did not come as a surprise. The more surprising finding came with calculating the Robinson-Foulds (RF) distance, a direct measure of phylogenetic tree similarity. RF calculates the number of nodes that are dissimilar between the phylogenetic trees under comparison, with lower values indicating a higher similarity degree. Even though the comparative genomics statistics results present the class level analyses as more-encompassing of orthology relationships, the inferred by OF species trees are more dissimilar than those for the phyla (Figure 1B). This could be rationalized by the fact that a smaller phylum dataset with longer genetic distances would be more likely to produce the same species tree topology. Nonetheless, it raises concerns about algorithms such as STRIDE using all genes, rather than only highly conserved ones, to infer a species tree. As Martinez-Gutierrez and Aylward [28] have pointed out, *"more genes and genomes do not necessarily improve phylogenetic accuracy"*. In light of this and considering our varying dataset sizes, we selected the GTDB domain-specific species trees for our downstream analysis.

Inferring ancestral genomes

We wanted to utilize the full potential of the genomic data at hand for reconstructing ancestral metabolic networks, and avoid simplified approaches such as phylogenetic presence/absence profiles [32] or those using near-universal gene family distribution as filtering criteria [7]. We therefore performed our phylogenetic analysis with OF [12, 13], a comprehensive platform for comparative genomics that provides a standardized, accurate, fast, and scalable orthology inference approach.

The relationship between number of each ancestor's descendants and ancestral genome size can be seen in Figure 2. The number of descendants for node zero is a direct representation of the respective dataset size, as all descendants are utilized to infer its gene content. Leaf parents like nodes 3, 5, 10, 11, 12, 13, on the other hand, tend to have the smallest number of descendants. These relationships can be seen in the phylum-level tree topology of Figure 2 panel B. Of interest here is the uneven addition of new taxa to the various tree clades, with some clades experiencing rapid enrichment, while others remaining with few taxa. This unevenness could be a direct reflection of field sampling bias, but can be more likely be attributed to the culturability, and thus higher quality of genomic data, of specific taxa. For example, nodes 3 and 5, which represent a superphylum-level clade called DPANN, archaea with extremely small genome sizes and few cultured members [33,

34], have very few descendants in all taxonomic-level datasets. What stands out the most, however, in Figure 2 panel, is the strong positive correlation between the inferred ancestral genome size and the number of descendants/initial sampling size, with a Pearson correlation coefficient close to 1 for all nodes.

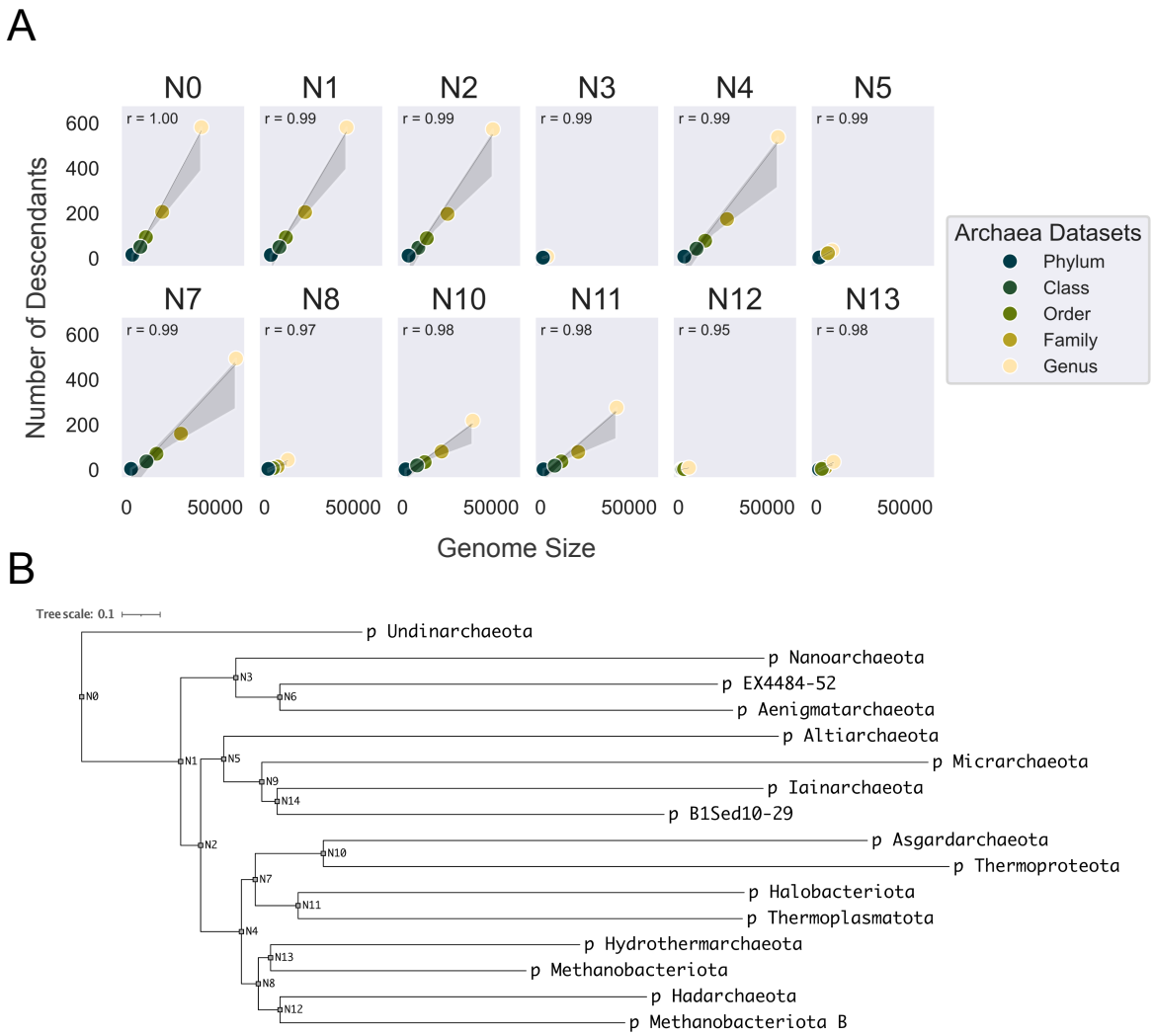


Figure 2: Relationship between genome size and number of node descendants. (A) Shows the number of descendants of each node as a function of the inferred ancestral genome sizes, for all taxonomic levels of the archaea datasets. Colors are assigned to the five taxonomic levels, and turn lighter with narrower taxonomies. (B) Depicts the position of the nodes in the GTDB species tree, as well as their descendants. The phylum-level GTDB pruned tree has been used for simplification purposes.

This relationship is not surprising, and can be attributed to the Duplication-Loss-Coalescence (DLC) parsimonious model for gene trees-species tree reconciliation OF utilizes. This model does not take into account horizontal gene transfer (HGT) events, and is known to overestimate gene content in ancestral genomes [35]. HGT, however, is not only omnipresent in prokaryotic evolution, but necessary for it [36]. Especially when considering that any ancestral gene content reconstruction is necessarily incomplete, as the extinct gene families cannot be accounted for, it seems unreasonable that an ancient, last common ancestor would have the metabolic versatility

of the entire modern biosphere. The inferred ancestral genomes of this study should therefore be considered with caution at this stage. Another model, the Duplication-Transfer-Loss (DTL) model, which accounts for HGT, is able to mitigate the tendency to infer unrealistically large ancestral genomes, getting bigger with each added genome, in the absence of HGT [35]. We therefore plan to redo our analysis by employing the DTL model, which has been shown to be more realistic and robust to gene tree uncertainty [37, 38].

Ancestral Genome Functional Annotation

For the metabolic network reconstruction of the ancestral genomes, we functionally annotated a single sequence for each of the OGs inferred to be present in the respective ancestor node. We based this decision on the rationale that an OG comprises genes that have descended from a single gene in the last common ancestor (LCA) of a group of species, and that sequences within the same OG are evolutionarily closer to each other than to sequences outside of that OG. Yet, the OGs OrthoFinder infers include both orthologs and paralogs, and functional annotation is known to be most reliable when based on orthologs, as they are expected to retain function more often than paralogs [39]. To assess the impact of choosing an OG representative sequence on functional annotation, we compared the functional annotations of the first sequence within each OG against the medoid sequence of the same OG (Figure 3). The medoid sequence is the sequence with the shortest genetic distance to all other sequences in the OG, and is therefore considered a better representative of the OG than the first—essentially a random—sequence. Since OF solves the gene length bias in orthogroup inference—which tends to cluster sequences of similar length together—the produced OGs include sequences of varying lengths. It may therefore be beneficial to perform a multiple sequence alignment for each OG, in the future, before choosing a representative sequence for functional annotation. Even though we do not check here for the length of the representative sequence relative to the median sequence length of an OG, there may be a selection bias towards lengths that are more common in the OG, which is also not necessarily erroneous.

Figure 3 presents the comparison statistics from the functional annotation of the two sequences belonging to the same OG in the form of a waffle plot; the dataset used for this part of the study is the phylum level archaeal one run against the Archaea (2157) eggNOG v5 database. EggNOG-mapper takes the protein sequences we provide and performs functional annotation after sequence alignment, based on the best hit. Out of 2849 OGs, for 68% the first and medoid sequences differed, for 26% they were the same, and 6% were individual hits (Fig. 3 orange), meaning that eggNOG-mapper found a match in the database for only one of them. Henceforth, we compare KEGG reaction and EC assignment only for different hits. Most hits do not correspond to neither a reaction ID nor EC number (Fig. 3 gray blocks). Considering that the GTDB data used in our study only contain sequences predicted by Prodigal to be protein-coding, the absence of a match for the majority of the sequences is perplexing; it raises questions regarding the incompatibility between tools and databases used throughout the globe for genomic and metagenomic workflows. For hits that do find a match, 73% share the same KEGG reaction ID, and another 71% share the same EC number (Fig. 3 purple and blue). With regard to the KEGG EC assignment, if we relax our constraints

and compare EC assignments up to the third digit, which specifies the nature of the reaction [40], the percentage of shared EC assignments increases to 79%. This simple statistical analysis stresses how important the choice of representative sequence is for functional annotation, even for related sequences that may have descended from the same gene.

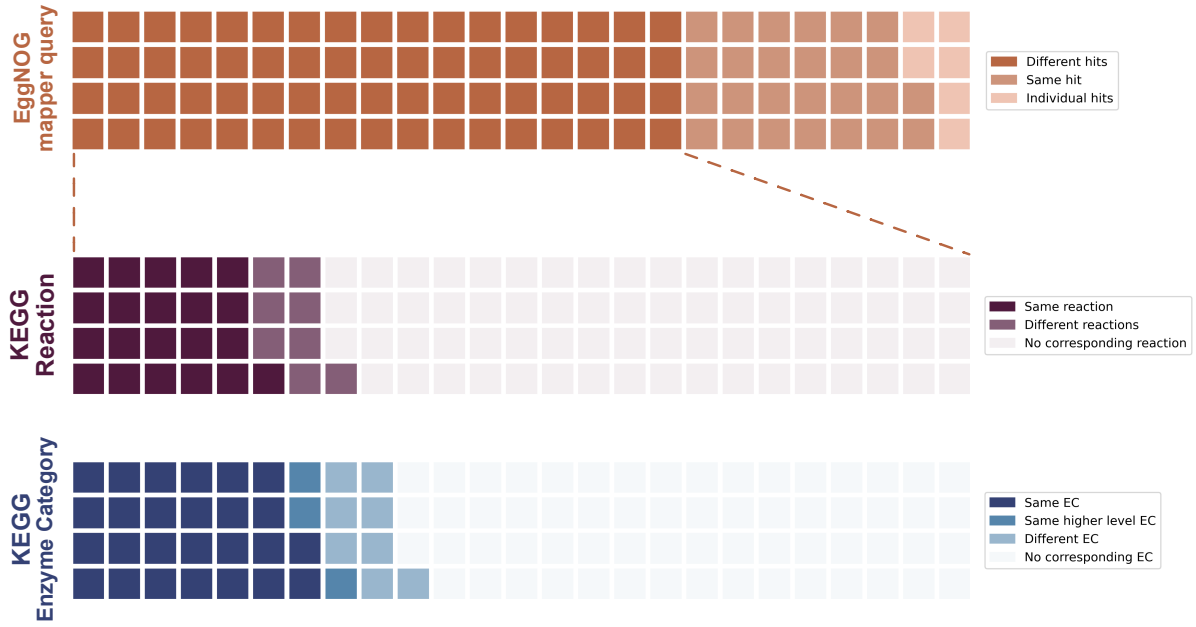


Figure 3: Comparison of functional annotations between the first and medoid sequence of each ancestral Node zero (N0) for the archaea phylum level dataset. The waffle plot shows the percentage of a quantity—orange for number of OGs (used as a proxy for number of hits), purple for KEGG reaction number of different queries, and blue for KEGG enzyme commission (EC) number of different queries. Every axis—orange, purple, and blue—shows a different statistic

For our inferred ancestral genomes belonging to the archaea phylum and class level datasets, the medoid sequence was calculated and utilized for OG functional annotation. For the rest of the datasets (found in Table 2), we sampled a hundred sequences from each OG at random, and calculated the medoid only for those. Since the medoid is calculated by pairwise alignment of all sequences in an OG, the computational burden increases according to the Gauss's summation formula (Equation 1), where n is the number of sequences. As our datasets grow in size, the number of sequences attributed to each OG increases to a point where calculating the medoid becomes extremely time-consuming. For example, an OG with 1000 sequences—which becomes common in our larger datasets—would require 500,500 pairwise alignments. We therefore opted for a computationally tractable approach, calculating the medoid for a hundred randomly sampled sequences for OGs larger than that. Instead of calculating the medoid for all OGs, we could also have annotated all sequences of an OG, and then choose the most common eggNOG OG (eggNOG orthogroup) or COG (cluster of orthologous genes) category as the representative one for the OG, as done by Xavier et al. [7].

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \tag{1}$$

As mentioned above, the functional annotation does not only depend on the sequence itself, but on the database against which it is aligned. A feature of eggNOG-mapper version 2 [17] allows for the generation of taxon-specific eggNOG databases. We therefore generated three databases spanning the prokaryotic domain, one for the Archaea (2157), one for Bacteria (2), and a general one including both domains (2157, 2). We then performed functional annotation of genes predicted to be present for certain internal tree nodes (presented in Figure 2) against the domain-specific and the general eggNOG databases, chose for the best hit between the two runs, as described in the Methods section, and reconstructed each metabolic network based on the KEGG reaction IDs assigned by eggNOG-mapper for each best hit. Extant genomes were annotated only against the general eggNOG database.

We also tried performing functional annotation of entire OGs, instead of single protein sequences, with profile hidden markov models (HMMs), as a more sensitive, probabilistic approach. In contrast to traditional substitution matrices, such as the BLOSUM matrix [41] we utilized to determine the medoid, profiles are position-specific scoring models and take into account specific—to each sequence—conservation patterns [42, 43], and can therefore offer enhanced alignment and functional annotation quality. However, this methodology is even more time-consuming than simply calculating the medoid, and was not feasible with our current computational resources. In future analyses, we plan to use HMMs for functional annotation of the OGs, and compare the results with those obtained by the medoid method.

Metabolic Network Reconstruction

One of the first things we noticed when mapping our inferred reaction IDs to the KEGG database was the absence of multiple IDs per dataset and often incomplete or missing fields, leading to inconsistent information between reactions; for example, some reactions include the KEGG module or pathway, while others do not. Even though the KEGG databases provide a valuable resource for the field of bioinformatics and metabolism modeling [44], KEGG was primarily designed for visualization purposes, with its reactions often unbalanced [45], and even elementally inconsistent [21]. For these reasons, we utilized the database compiled by Goldford et al. (2024) [21], which extends the KEGG reactions database by adding detailed organic and inorganic cofactor dependencies from various other databases, while excluding elementally inconsistent reactions. Reaction directionality was performed with eQuilibrator [22], to ensure a more realistic network reconstruction.

To inspect the connectivity of the reconstructed metabolisms of putative, ancient microorganisms, we visualized the networks using iPath [46], an interactive metabolic pathway explorer that is based on four KEGG global maps. The hypothetical, genome-scale metabolic model of LACA for the family-level archaea dataset can be seen in Figure 4, while the rest of LACA-inferred metabolic networks can be found in Appendix Figs. A.5 - A.8. Even at the phylum-level, the network seems relatively well-connected, with isolated reactions being dispersed across the entire map; the small-

est of the reconstructed metabolic networks, after all, contains 1192 reactions for the Goldford database, which includes multiple versions of the same reaction. Only one version of each reaction, the KEGG version, can be visualized using iPath.

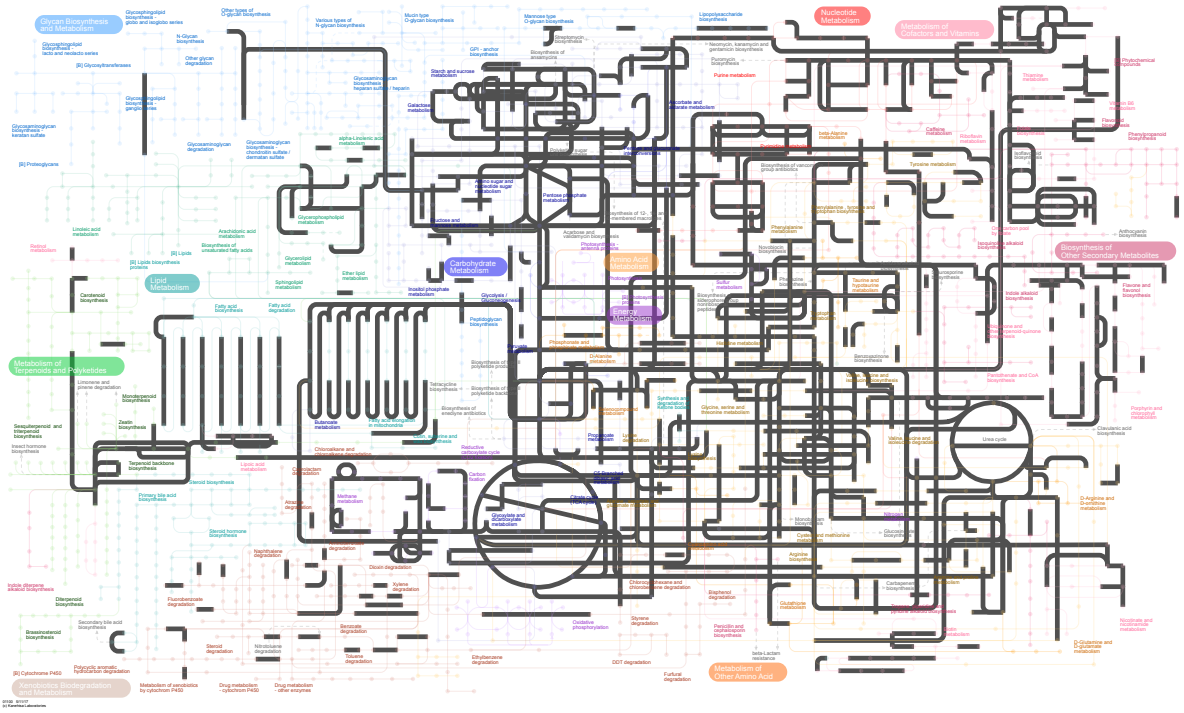


Figure 4: Reconstruction of LACA's metabolic network from extant life, for the family-level archaea dataset. In black: enzymes and metabolic pathways that were inferred to be present in LACA.

Phylogenetic-based metabolic network reconstruction enables the mapping of genetic information to genome distance from the tree root. We aimed to explore the emergence and evolution of individual ECs across the tree of life (ToL), a topic that has not been previously investigated. For this, we acquired the distance of all nodes from the tree root and used it as a proxy for evolution. All enzyme categories are present in the reconstructed LACA metabolism (Figure 5B). The relative abundance of ligases and oxidoreductases is higher in the inferred ancient metabolisms, while that of transferases declines. Lyases and isomerases are universally more limited and do not become enriched over time. This divergence may indicate a different rate of innovation for various ECs or suggest that specific types of enzymes are more prone to either loss or gain.

No other significant differences or patterns are observable, so it may be beneficial to divide the tree into multiple clades and track EC evolution within specific groups of species rather than across the entire tree. This approach will be especially important if we increase the deep branch resolution by reconstructing the metabolic networks of all possible internal tree nodes. Even though the relative abundance of the six EC categories varies between inferred and extant microorganisms, their distribution across the tree follows a similar pattern, reflecting the species distribution across the tree (Fig. 5A). Small variations, such as the increase in transferases at a distance of around 1.5 from the tree root, may indicate an enrichment of this enzyme category in certain species, possibly due to specialization or an evolutionary advantage. The evolution of ECs across the ToL for the rest of the archaeal datasets can be found in Appendix Figs. A.10 - A.13.

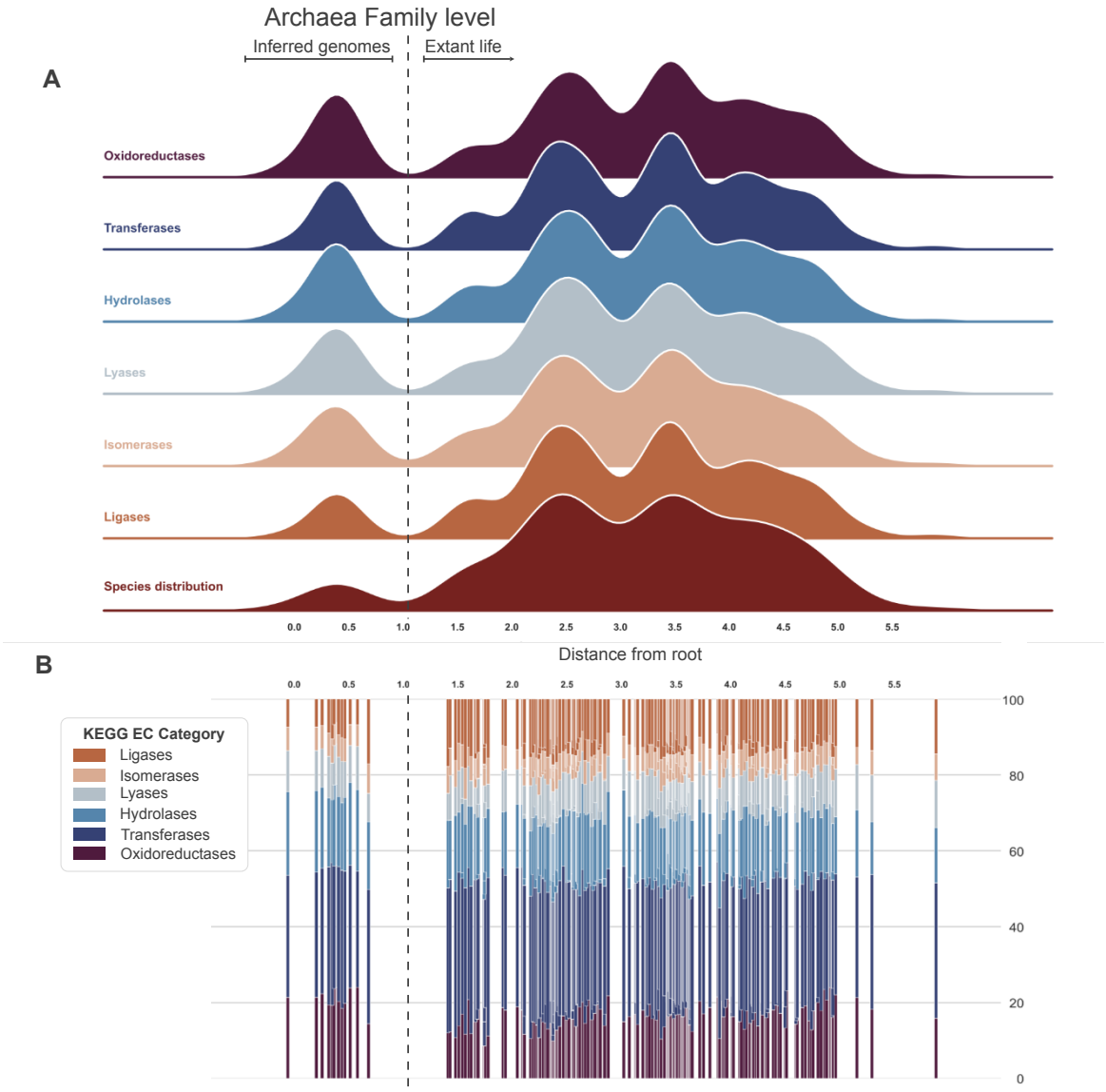


Figure 5: Evolution of individual enzyme categories for the family-level archaea dataset. The dashed line separates the inferred ancient metabolisms, on the left, from the extant ones, on the right. The ridgeplot of panel (A) displays the distribution of each category as a function of distance from the tree root, with the last axis presenting the species distribution. (B) shows the relative abundance of each category at that particular distance as a stacked barplot.

Metabolic Network Expansion

One of the first objectives of this internship was to investigate the evolution of metabolism across the ToL, and the metabolic potential of the inferred ancient microorganisms, with Metabolic Network Expansion (MNE). MNE is a graph-based structural analysis of large-scale metabolic systems [24], that has previously been used to explore the metabolic scope of extant microorganisms. While the method itself does not take into account the phylogenetic relationships between species, Handorf et al. [25], among others, have noted that it exhibits evolution-like features. The temporal

order of reaction addition to the expanding network, however, solely depends on the nature of the seed compounds utilized to initiate the expansion, and can therefore be described as predetermined.

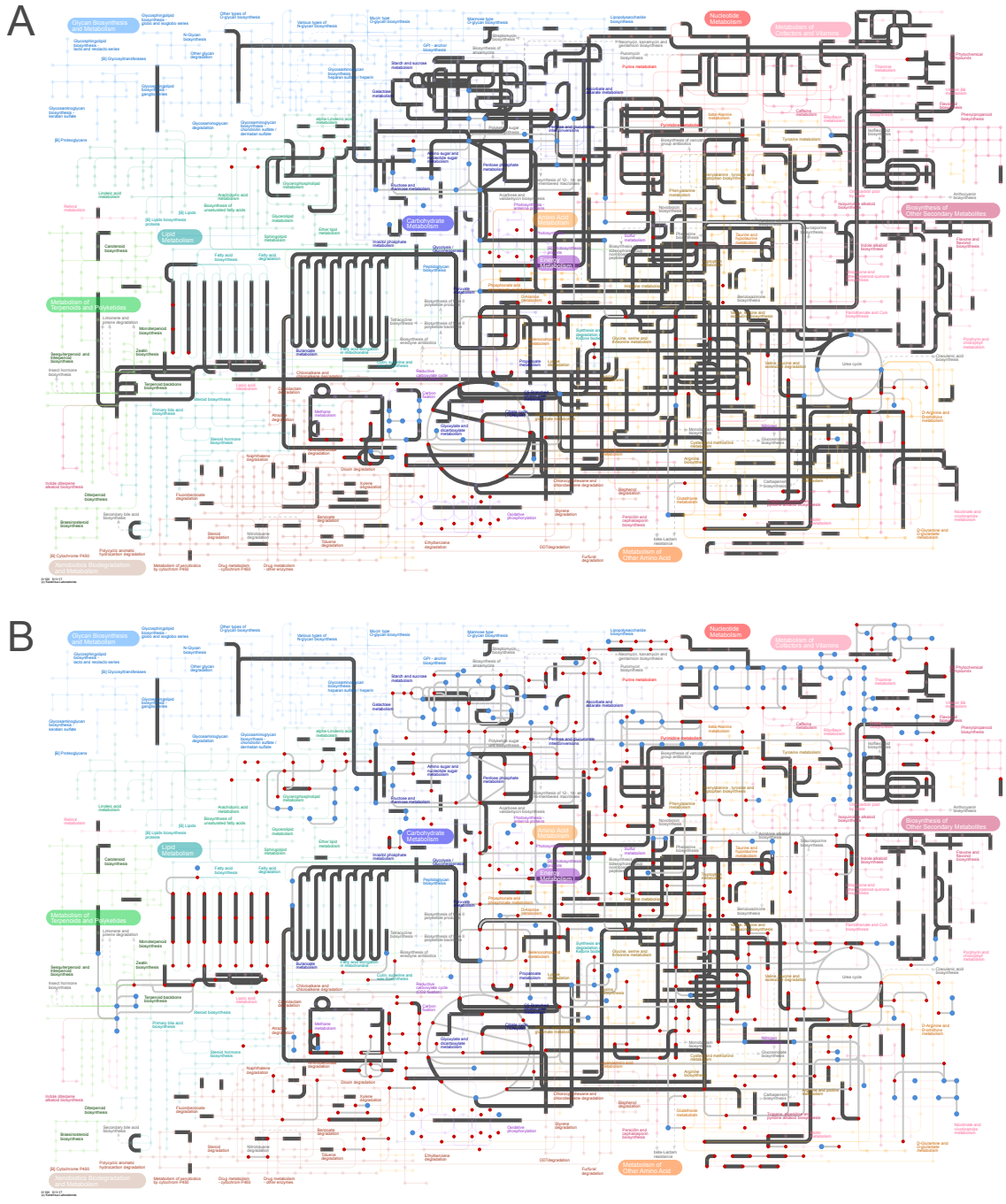


Figure 6: Reconstruction of LACA's metabolic network from extant life, for the family-level archaea dataset. In black: enzymes and metabolic pathways that were inferred to be present in LACA, which are not part of the expansion. In gray: enzymes and metabolic pathways of the expanded network. Red dots: respective seed set compounds. Blue dots: scope compounds after network expansion. (A) shows the expansion of the network with the 10% completeness seed set, while (B) shows the expansion with the 60% completeness seed set.

For example, Figure 6 shows the expansion of the family-level LACA for two different seed sets:

the one with the lowest 10% complexity compounds based on the Bertz complexity index (A) and the one with the lowest 60% complexity compounds. The 10% seed set expansion is more limited, with fewer seed compounds (red dots) unevenly distributed throughout the metabolic map. Depending on the positions of the seeds, different parts of the metabolism will be accessed (gray lines), with a specified order of addition. A visual representation like the one above quickly reveals the pre-determined nature of MNE. If anything, MNE suggests the flow of information in the system more than it reveals evolution-like traits.

While the inclusion of more seed compounds leads to a more extensive network, the richness of the seed set and clustering of the seeds in specific map regions limit the expansion to a few iterations. The first iteration introduces the largest change in the network, and the total difference in numbers between the scope (blue dots) and seed size is small. Additionally, since the seed sets are randomly generated, some seeds fall outside the reconstructed network. This is illustrated in Figure 6, and more clearly in Appendix Figure A.9, where several red dots representing the seed compounds are not connected to any of the gray lines that denote the network reactions. The seed set size used for expansion will always be equal to or smaller than the initially generated seed set size, so the ratio of the two will range between 0 and 1. We use this ratio instead of the initial or expansion-used seed set sizes in our MNE statistical analysis. A straightforward way to compensate for this lack of overlap between the seed set and the reconstructed metabolic network, while increasing the spread of the seeds in the network is to select a set number of compounds from each KEGG module. However, it may be more beneficial to manually select prebiotic or at least lower complexity compounds from a variety of molecular classes. In this way, the seeds will be more evenly distributed throughout the network, and the expansion will be more extensive, potentially with more iterations.

For a given seed set, the seed set size ratio increases rapidly with smaller genome sizes but soon reaches a plateau (Fig. 7A). As demonstrated in Figures 4 and A.5 - A.8, the reconstructed metabolic networks of LACA across various taxonomic levels show minimal variation, despite substantial increases in genome size for the respective LACA genomes. The saturation curve suggests that the metabolic network size stabilizes once a certain genome size is reached, indicating that the maximum number of enzymes present in the inferred genomes and available for inclusion in the network has been attained.

For the various seed sets, the seed set size ratio decreases as the total seed set size increases (Fig. 7A), implying that compounds of higher complexity are progressively less integrated into the expansion. In other words, the larger the seed set, the more seeds fall outside the network. Both the seed set size and the scope size (Fig. 7B) of an expansion show a similar correlation with genome size, as demonstrated by their linear relationship (Fig. 7C): larger seed sets produce larger scope sizes. The scope size distribution (Fig. 7D)

need to comment on bimodality, symmetry, or skewed towards smaller values. need to comment on how that is correlated to seed set size rather than initial genome size (themaybe a better metric would be met. network size) need to comment on how the scope size increases drastically when going from the 40 to the 60% complexity seed set, but then kind of stagnates.

and then I need to comment on the same for the extant genomes.

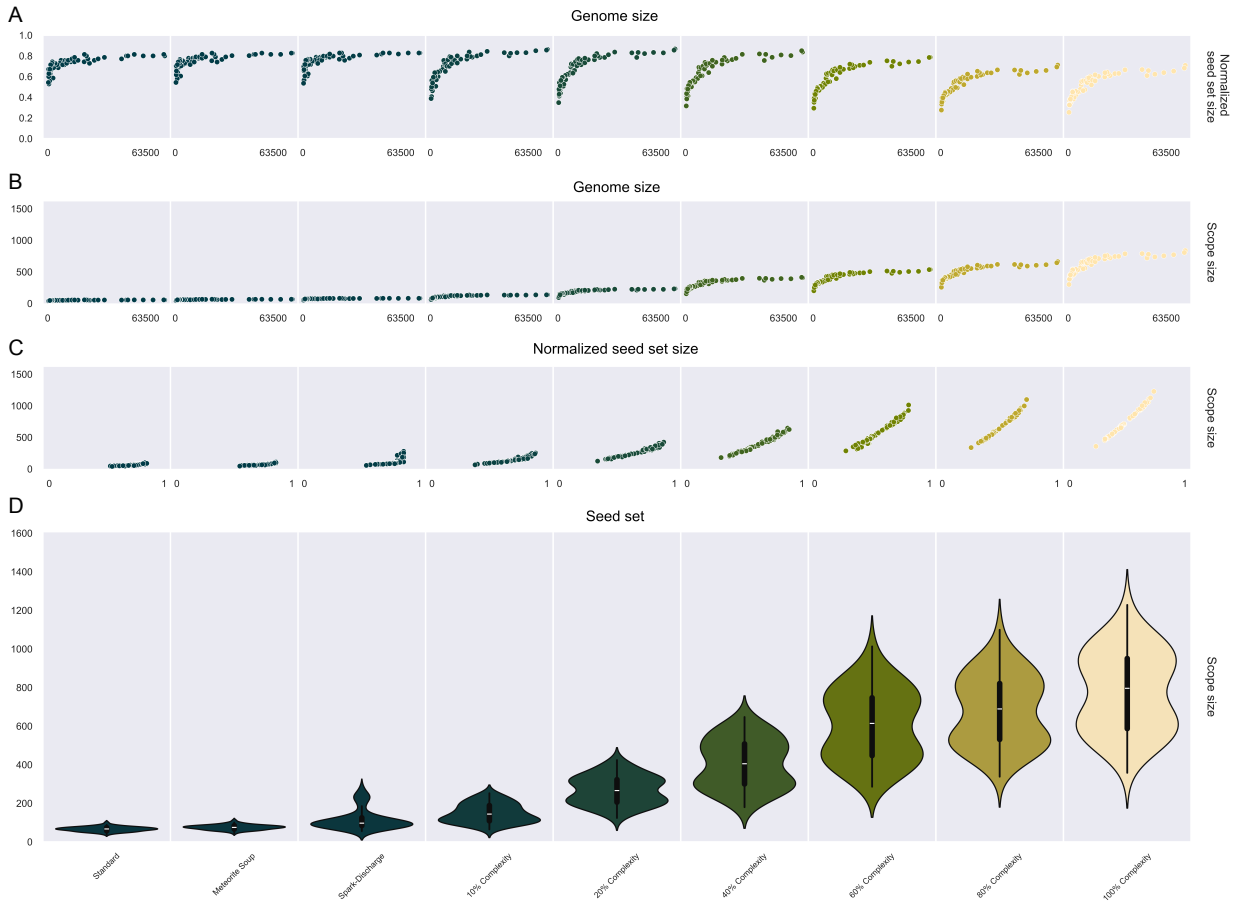


Figure 7: Overview of the MNE for all inferred ancient metabolisms. (A) shows the microorganisms' inferred genome size as a function of the normalized seed set size for each seed set. The number of protein-encoding genes is used as a proxy for genome size, while the normalized seed set size reflects the number of seed compounds found in the network divided by the initial number of seeds. (B) displays genome size as a function of scope size for each seed set, with scope size defined as the number of compounds in the network after expansion. (C) illustrates the normalized seed set size as a function of scope size for each seed set. (D) presents the scope size distribution for each seed set in the form of a violin plot.

It may be beneficial to explore the minimal seed sets of the reconstructed networks. According to Handorf et al. (2008) [47], the minimal seed set of an entire metabolic network is the smallest set of compounds capable of producing all its metabolites. This approach has shown that minimal seed sets reflect the nutritional requirements of the microorganisms under analysis. In a reverse ecology concept, as Borenstein et al. (2008) [48] have demonstrated, this approach can predict the nutrients a microorganism can acquire from the environment. Therefore, when applied to the hypothetical reconstructed metabolisms of ancient microorganisms, this method could provide insights into the environmental conditions in which they could have thrived.

MNE of Extant Metabolisms

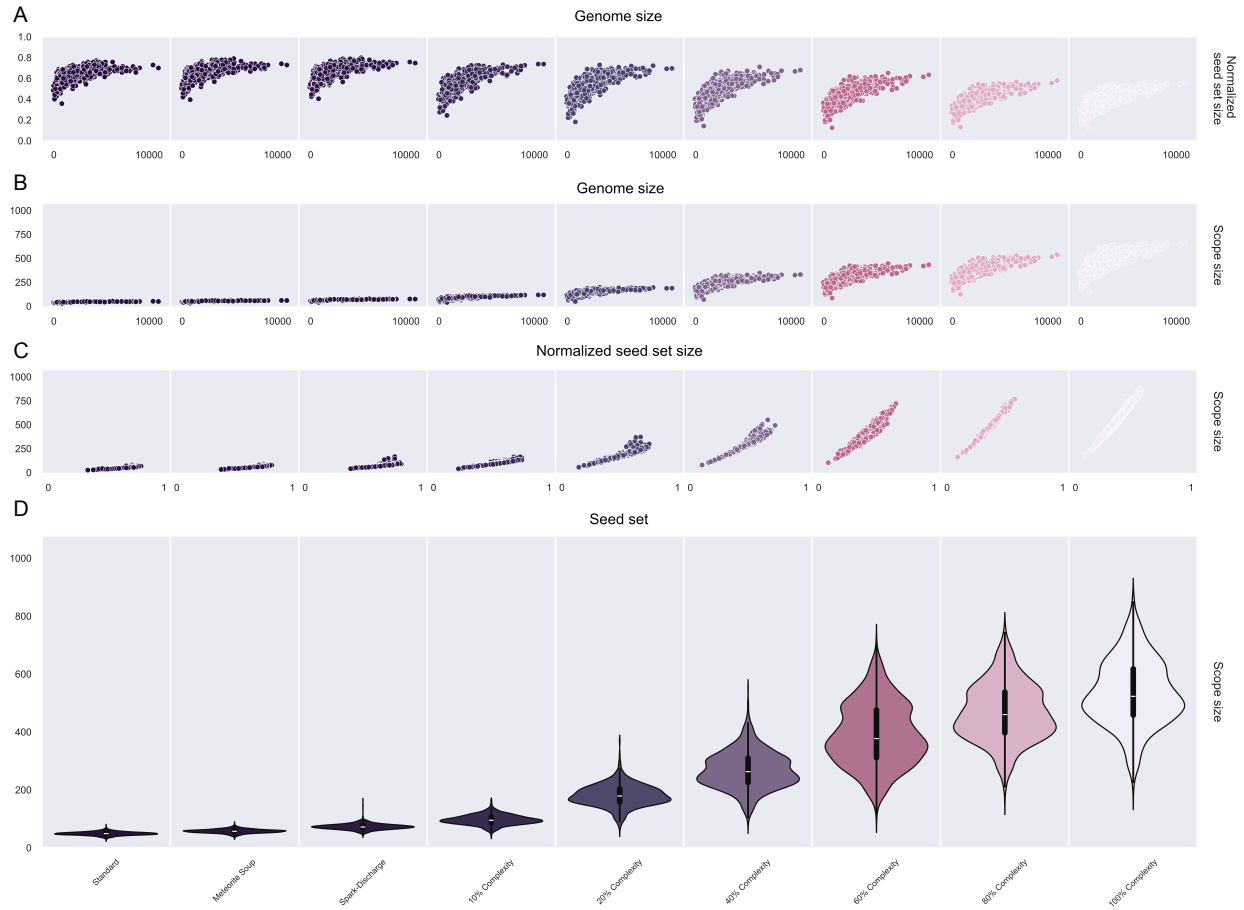


Figure 8: Overview of the MNE for all extant metabolisms. (A) shows the microorganisms' inferred genome size as a function of the normalized seed set size for each seed set. The number of protein-encoding genes is used as a proxy for genome size, while the normalized seed set size reflects the number of seed compounds found in the network divided by the initial number of seeds. (B) displays genome size as a function of scope size for each seed set, with scope size defined as the number of compounds in the network after expansion. (C) illustrates the normalized seed set size as a function of scope size for each seed set. (D) presents the scope size distribution for each seed set in the form of a violin plot.

These two show scope size as a function of distance from tree root. Maybe it'd be better to remove these two actually.

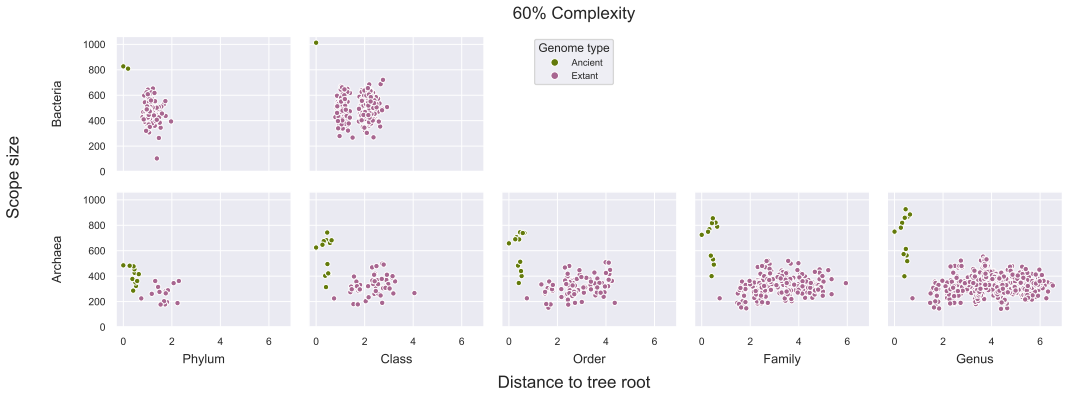


Figure 9

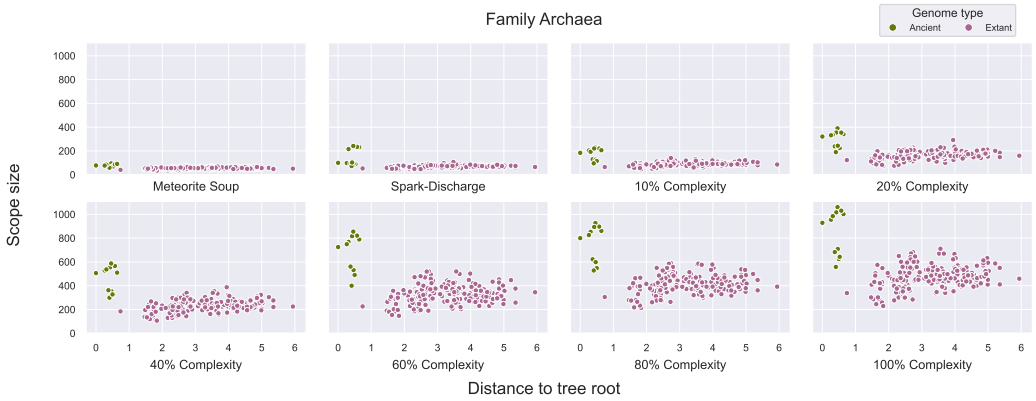


Figure 10

4 References

1. Tahon, G., Geesink, P. & Ettema, T. J. Expanding Archaeal Diversity and Phylogeny: Past, Present, and Future. *Annu. Rev. Microbiol.* **75,** 359–381 (2021).
2. Parks, D. H. *et al.* A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life. *Nat Biotechnol* **36,** 996–1004 (2018).
3. Parks, D. H. *et al.* A Complete Domain-to-Species Taxonomy for Bacteria and Archaea. *Nat Biotechnol* **38,** 1079–1086 (2020).
4. Parks, D. H. *et al.* GTDB: An Ongoing Census of Bacterial and Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy. *Nucleic Acids Research* **50,** D785–D794 (2022).
5. Rinke, C. *et al.* A Standardized Archaeal Taxonomy for the Genome Taxonomy Database. *Nat Microbiol* **6,** 946–959 (2021).
6. Williams, T. A. *et al.* Integrative Modeling of Gene and Genome Evolution Roots the Archaeal Tree of Life. *Proc. Natl. Acad. Sci. U.S.A.* **114** (2017).
7. Xavier, J. C. *et al.* The Metabolic Network of the Last Bacterial Common Ancestor. *Commun Biol* **4,** 413 (2021).
8. Moody, E. R. R. *et al.* The Nature of the Last Universal Common Ancestor and Its Impact on the Early Earth System. *Nat Ecol Evol* (2024).
9. Coleman, G. A. *et al.* A Rooted Phylogeny Resolves Early Bacterial Evolution. *Science* **372,** eabe0511 (2021).
10. Hyatt, D. *et al.* Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC Bioinformatics* **11,** 119 (2010).
11. The Genome Standards Consortium *et al.* Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea. *Nat Biotechnol* **35,** 725–731 (2017).
12. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics. *Genome Biol* **20,** 238 (2019).
13. Emms, D. M. & Kelly, S. OrthoFinder: Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy. *Genome Biol* **16,** 157 (2015).
14. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND. *Nat Methods* **18,** 366–368 (2021).
15. Van Dongen, S. Graph Clustering Via a Discrete Uncoupling Process. *SIAM J. Matrix Anal. & Appl.* **30,** 121–141 (2008).
16. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* **33,** 1635–1638 (2016).

17. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* **38** (ed Tamura, K.) 5825–5829 (2021).

18. Huerta-Cepas, J. *et al.* eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses. *Nucleic Acids Research* **47,** D309–D314 (2019).

19. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30,** 772–780 (2013).

20. The UniProt Consortium *et al.* UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51,** D523–D531 (2023).

21. Goldford, J. E., Smith, H. B., Longo, L. M., Wing, B. A. & McGlynn, S. E. Primitive Purine Biosynthesis Connects Ancient Geochemistry to Modern Metabolism. *Nat Ecol Evol* (2024).

22. Beber, M. E. *et al.* eQuilibrator 3.0: A Database Solution for Thermodynamic Constant Estimation. *Nucleic Acids Research* **50,** D603–D609 (2022).

23. Vincent, S. G. T., Jennerjahn, T. & Ramasamy, K. in *Microbial Communities in Coastal Sediments* 79–117 (Elsevier, 2021).

24. Ebenhoh, O., Handorf, T. & Heinrich, R. Structural Analysis of Expanding Metabolic Networks. *Genome Informatics* **15,** 35–45 (2004).

25. Handorf, T., Ebenhöh, O. & Heinrich, R. Expanding Metabolic Networks: Scopes of Compounds, Robustness, and Evolution. *J Mol Evol* **61,** 498–512 (2005).

26. Kuzniar, A., Van Ham, R. C., Pongor, S. & Leunissen, J. A. The Quest for Orthologs: Finding the Corresponding Gene across Genomes. *Trends in Genetics* **24,** 539–551 (2008).

27. Graybeal, A. Is It Better to Add Taxa or Characters to a Difficult Phylogenetic Problem? *Systematic Biology* **47** (ed Cannatella, D.) 9–17 (1998).

28. Martinez-Gutierrez, C. A. & Aylward, F. O. Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. *Molecular Biology and Evolution* **38** (ed Battistuzzi, F. U.) 5514–5527 (2021).

29. Emms, D. & Kelly, S. *STAG: Species Tree Inference from All Genes* http://biorxiv.org/lookup/doi/10.1101/267914 (2024). Pre-published.

30. Emms, D. M. & Kelly, S. STRIDE: Species Tree Root Inference from Gene Duplication Events. *Molecular Biology and Evolution* **34,** 3267–3278 (2017).

31. Huson, D. H. & Scornavacca, C. Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Systematic Biology* **61,** 1061–1067 (2012).

32. Kreimer, A., Borenstein, E., Gophna, U. & Ruppin, E. The Evolution of Modularity in Bacterial Metabolic Networks. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 6976–6981 (2008).

33. Dombrowski, N., Lee, J.-H., Williams, T. A., Offre, P. & Spang, A. Genomic Diversity, Lifestyles and Evolutionary Origins of DPANN Archaea. *FEMS Microbiology Letters* **366** (2019).

34. Dombrowski, N. *et al.* Undinarchaeota Illuminate DPANN Phylogeny and the Impact of Gene Transfer on Archaeal Evolution. *Nat Commun* **11,** 3939 (2020).

35. Doolittle, W. F. *et al.* How Big Is the Iceberg of Which Organellar Genes in Nuclear Genomes Are but the Tip? *Phil. Trans. R. Soc. Lond. B* **358** (eds Allen, J. F. & Raven, J. A.) 39–58 (2003).

36. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral Gene Transfer and the Nature of Bacterial Innovation. **405** (2000).

37. Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. Efficient Exploration of the Space of Reconciled Gene Trees. *Systematic Biology* **62,** 901–912 (2013).

38. Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V. & Boussau, B. Genome-Scale Phylogenetic Analysis Finds Extensive Gene Transfer among Fungi. *Phil. Trans. R. Soc. B* **370,** 20140335 (2015).

39. Gabaldón, T. & Koonin, E. V. Functional and Evolutionary Implications of Gene Orthology. *Nat Rev Genet* **14,** 360–366 (2013).

40. McDonald, A. G., Boyce, S. & Tipton, K. F. ExplorEnz: The Primary Source of the IUBMB Enzyme List. *Nucleic Acids Research* **37,** D593–D597 (Database 2009).

41. Henikoff, S. & Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci. U.S.A.* **89,** 10915–10919 (1992).

42. Mount, D. W. Using Hidden Markov Models to Align Multiple Sequences. *Cold Spring Harb Protoc* **2009,** pdb.top41 (2009).

43. Gribskov, M., McLachlan, A. D. & Eisenberg, D. Profile Analysis: Detection of Distantly Related Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **84,** 4355–4358 (1987).

44. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28,** 27–30 (2000).

45. Wrzodek, C., Büchel, F., Ruff, M., Dräger, A. & Zell, A. Precise Generation of Systems Biology Models from KEGG Pathways. *BMC Syst Biol* **7,** 15 (2013).

46. Darzi, Y., Letunic, I., Bork, P. & Yamada, T. iPath3.0: Interactive Pathways Explorer V3. *Nucleic Acids Research* **46,** W510–W513 (2018).

47. Handorf, T., Christian, N., Ebenhöh, O. & Kahn, D. An Environmental Perspective on Metabolism. *Journal of Theoretical Biology* **252,** 530–537 (2008).

48. Borenstein, E., Kupiec, M., Feldman, M. W. & Ruppin, E. Large-Scale Reconstruction and Phylogenetic Analysis of Metabolic Environments. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 14482–14487 (2008).

A Appendix

A1. Phylogenetic Analysis

All shown phylogenetic trees have been visualized with Dendroscope [31].

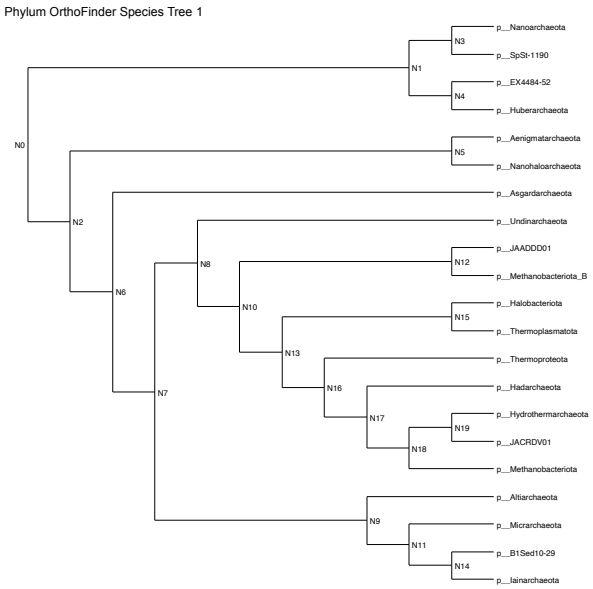


Figure A1: OrthoFinder-inferred species tree; first analysis.

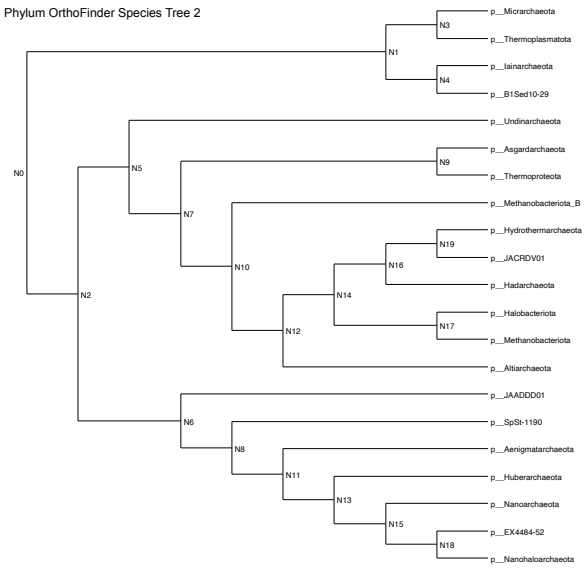


Figure A2: OrthoFinder-inferred species tree; second analysis.

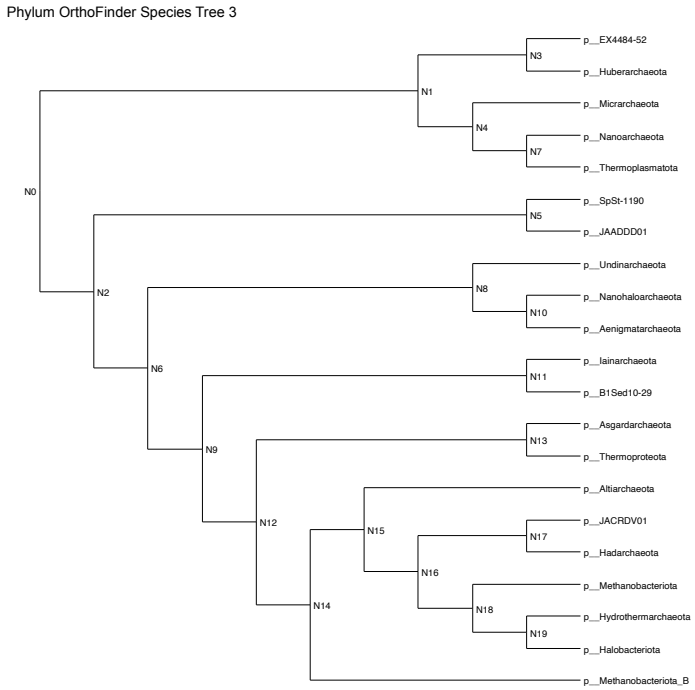


Figure A3: OrthoFinder-inferred species tree; third analysis.

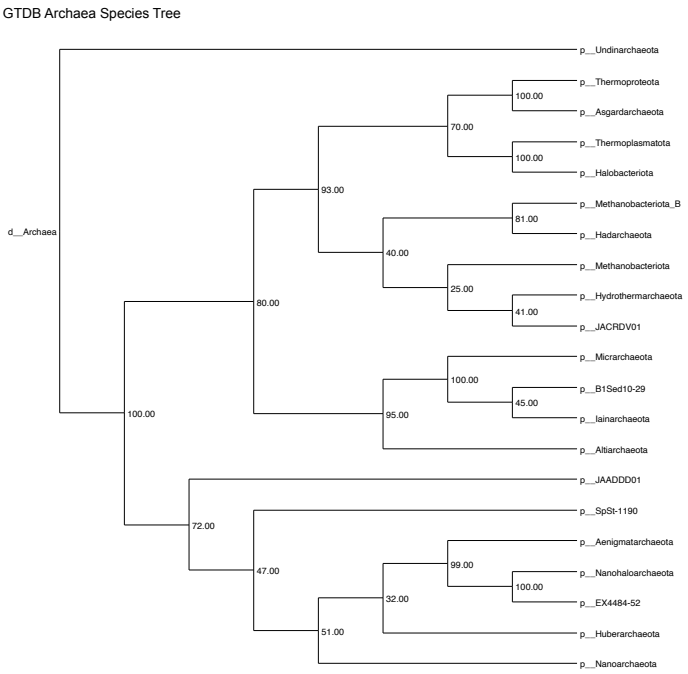


Figure A4: GTDB archaea species tree.

A2. Metabolic Network Reconstruction of LACA for all Archaea Datasets

Metabolic Potential of LACA.

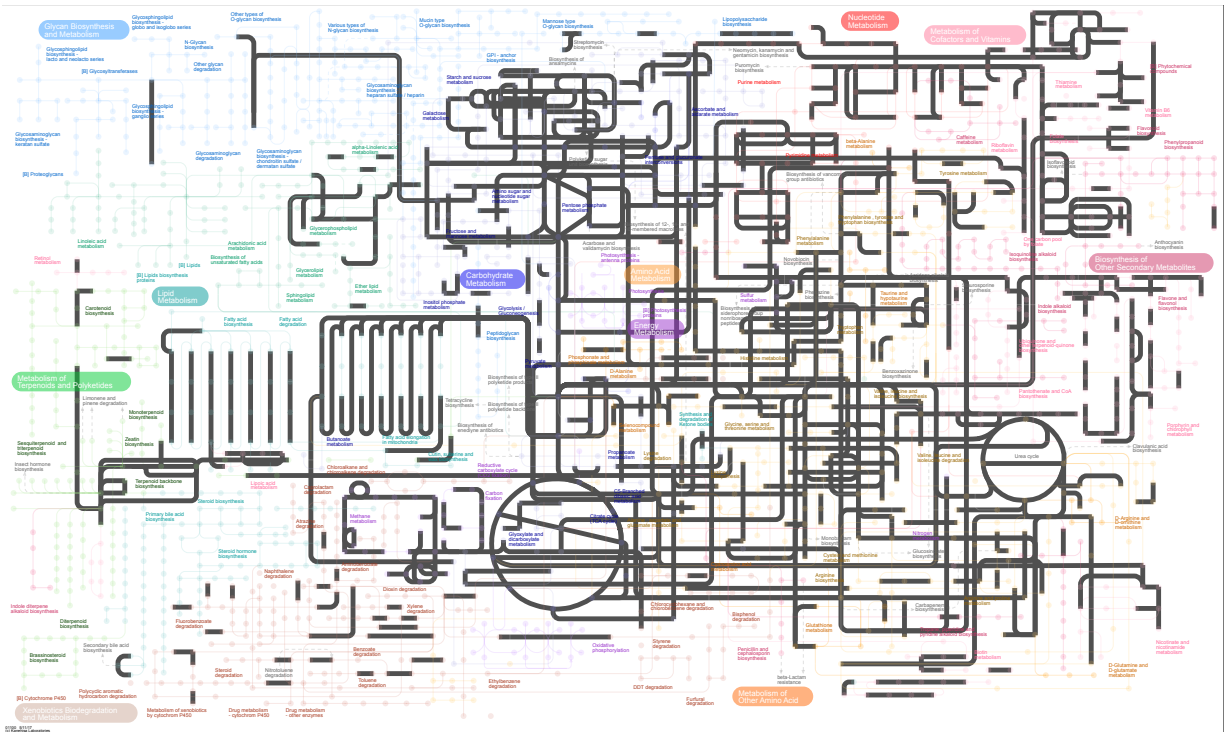


Figure A.5: Reconstruction of LACA's metabolic network from extant life, for the phylum-level archaea dataset. In black: enzymes and metabolic pathways that were inferred to be present in LACA.

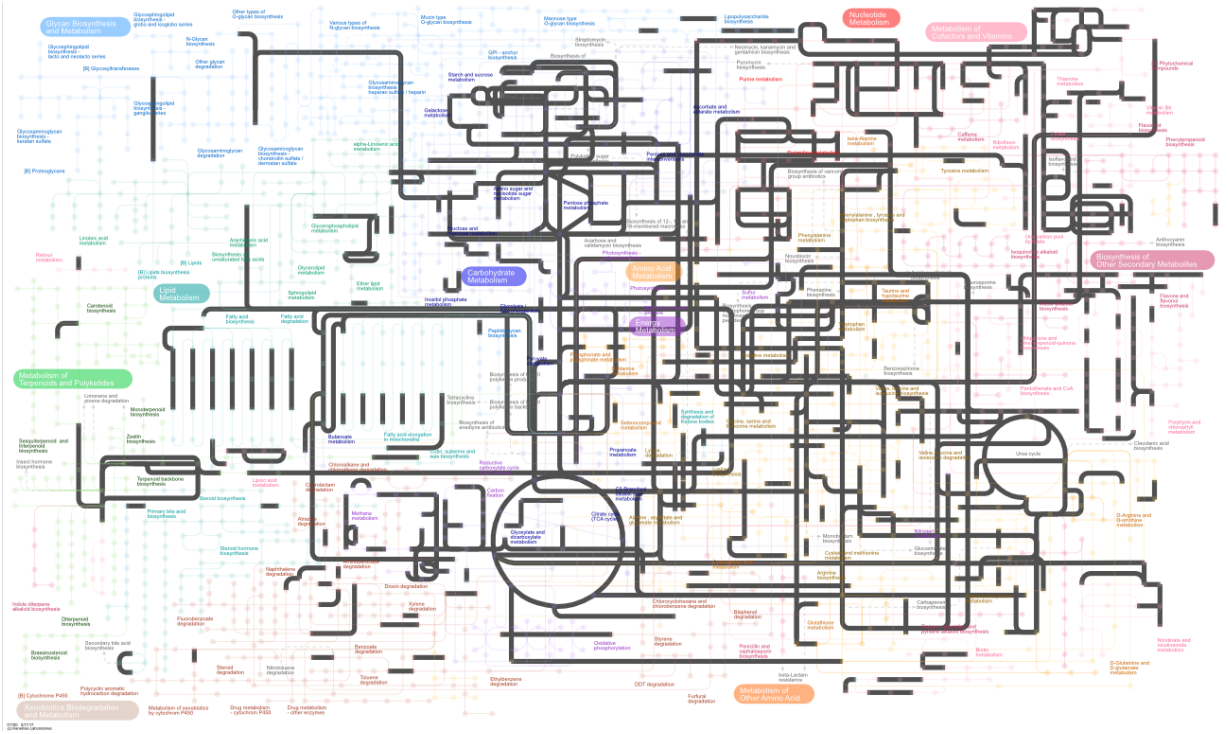


Figure A.6: Reconstruction of LACA's metabolic network from extant life, for the class-level archaea dataset. In black: enzymes and metabolic pathways that were inferred to be present in LACA.

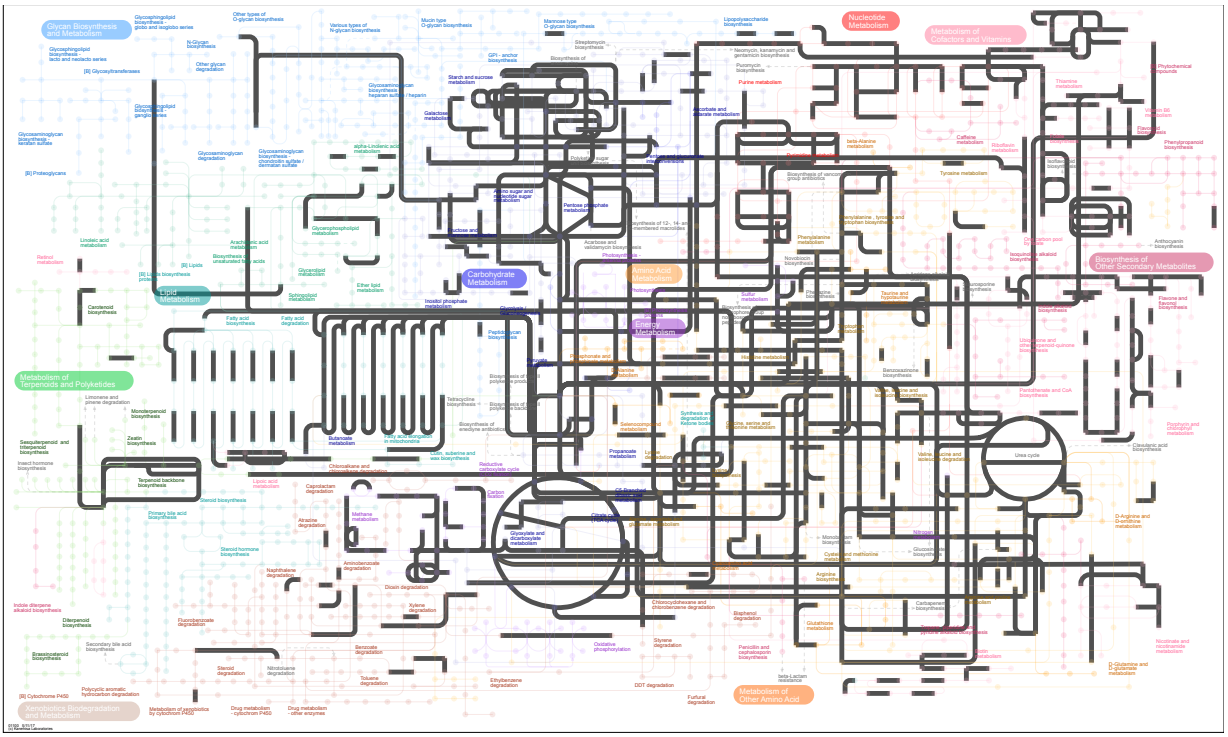


Figure A.7: Reconstruction of LACA's metabolic network from extant life, for the order-level archaea dataset. In black: enzymes and metabolic pathways that were inferred to be present in LACA.

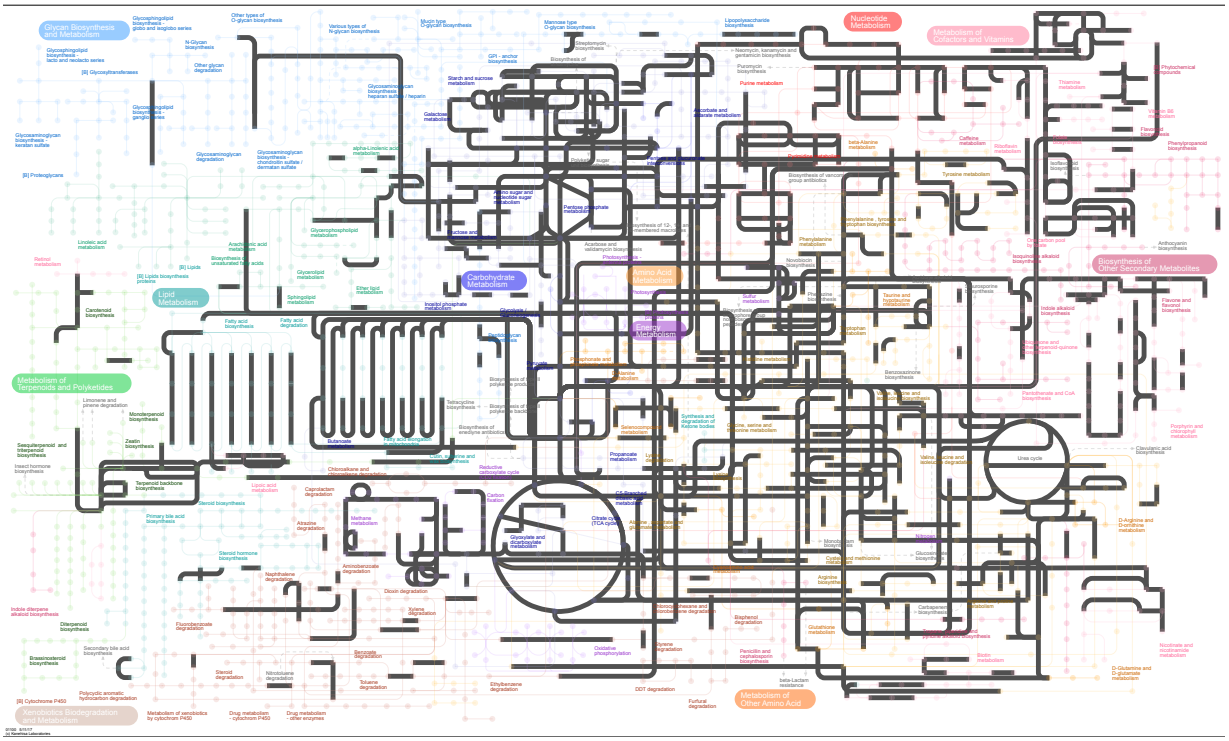


Figure A.8: Reconstruction of LACA's metabolic network from extant life, for the genus-level archaea dataset. In black: enzymes and metabolic pathways that were inferred to be present in LACA.

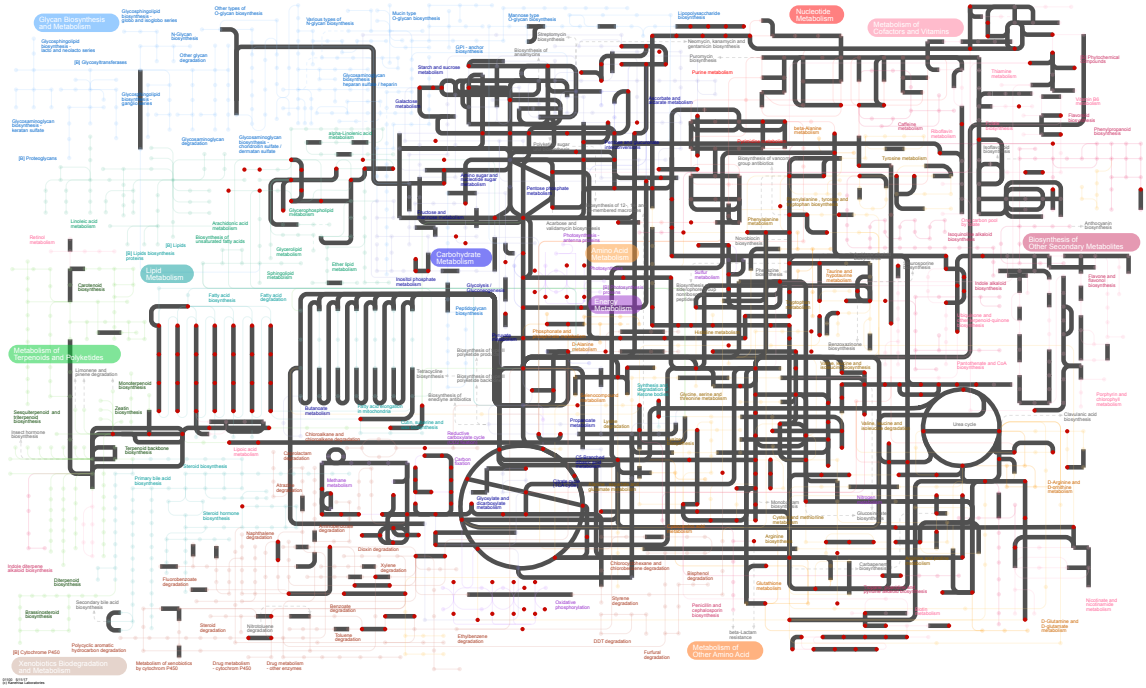


Figure A.9: Reconstruction of LACA's metabolic network from extant life, for the family-level archaea dataset. In black: enzymes and metabolic pathways that were inferred to be present in LACA. Red dots: seed compounds of 60% completeness seed set.

EC number evolution in LACA.

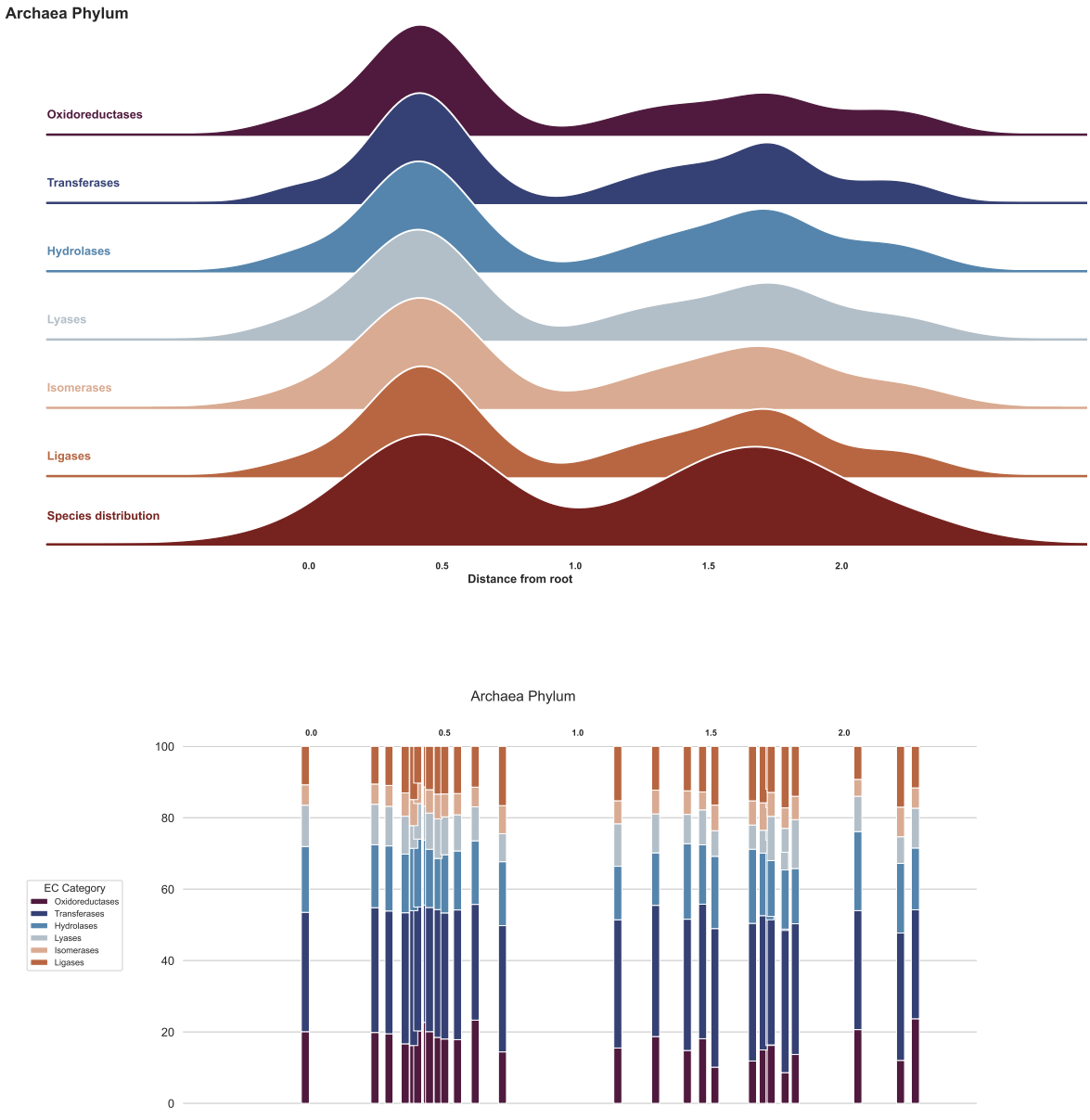
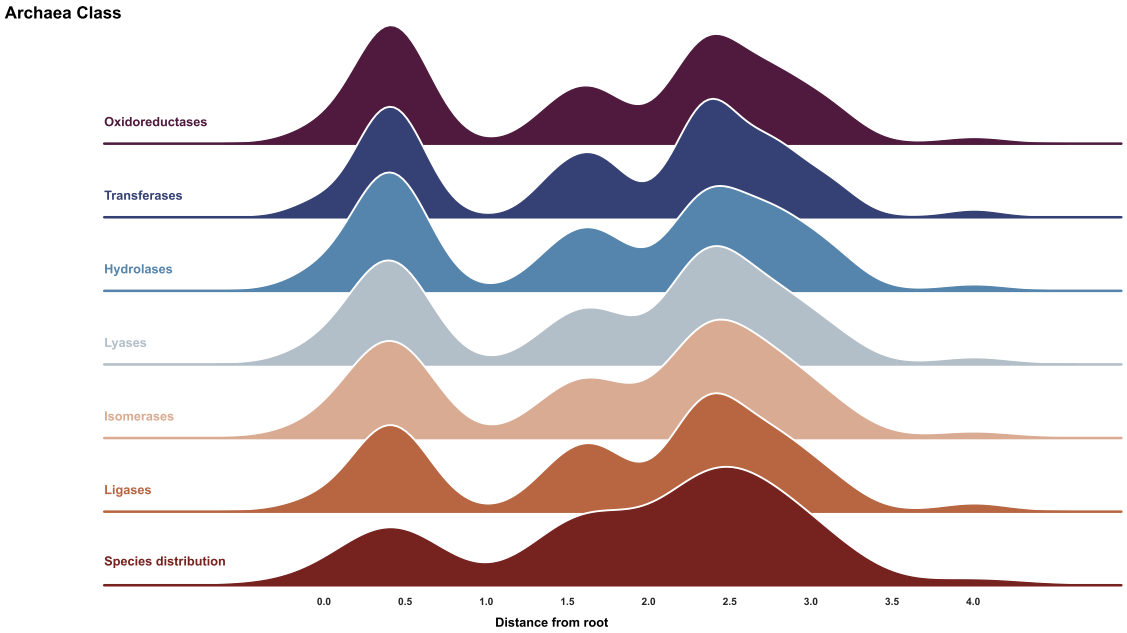


Figure A.10: Evolution of individual enzyme categories for the phylum-level archaea dataset. The ridgeplot displays the distribution of each category as a function of distance from the tree root, with the last axis presenting the species distribution, while the barplot the relative abundance of each category at that particular distance as a stacked barplot.



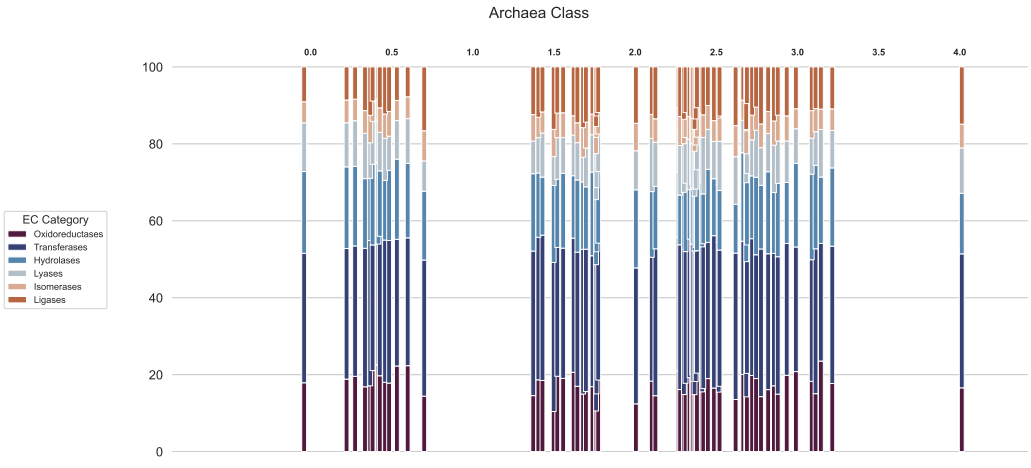
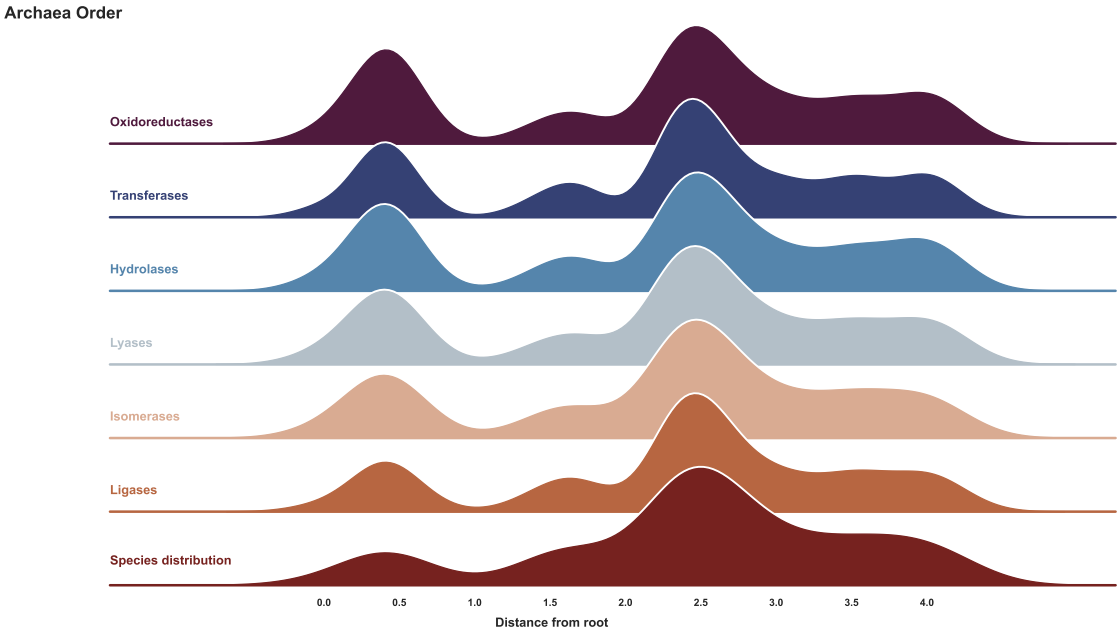


Figure A.11: Evolution of individual enzyme categories for the class-level archaea dataset. The ridgeplot displays the distribution of each category as a function of distance from the tree root, with the last axis presenting the species distribution, while the barplot the relative abundance of each category at that particular distance as a stacked barplot.



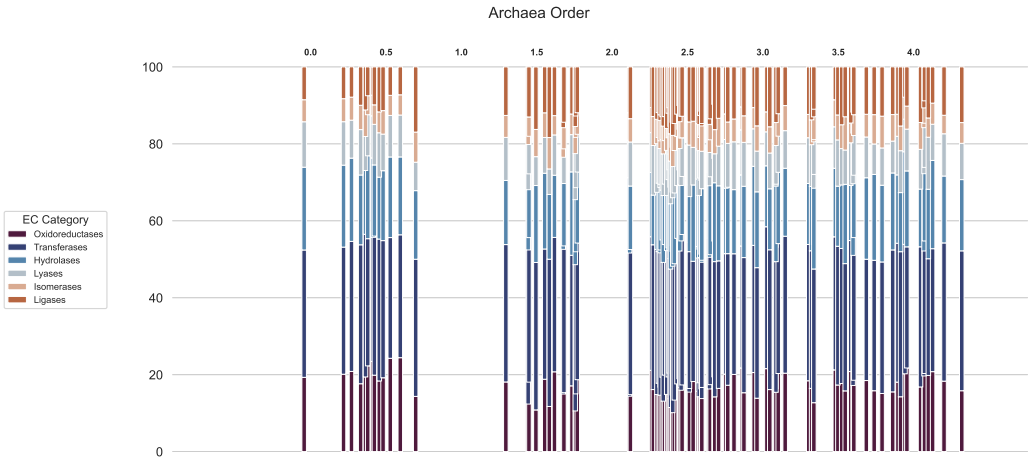


Figure A.12: Evolution of individual enzyme categories for the order-level archaea dataset. The ridgeplot displays the distribution of each category as a function of distance from the tree root, with the last axis presenting the species distribution, while the barplot the relative abundance of each category at that particular distance as a stacked barplot.

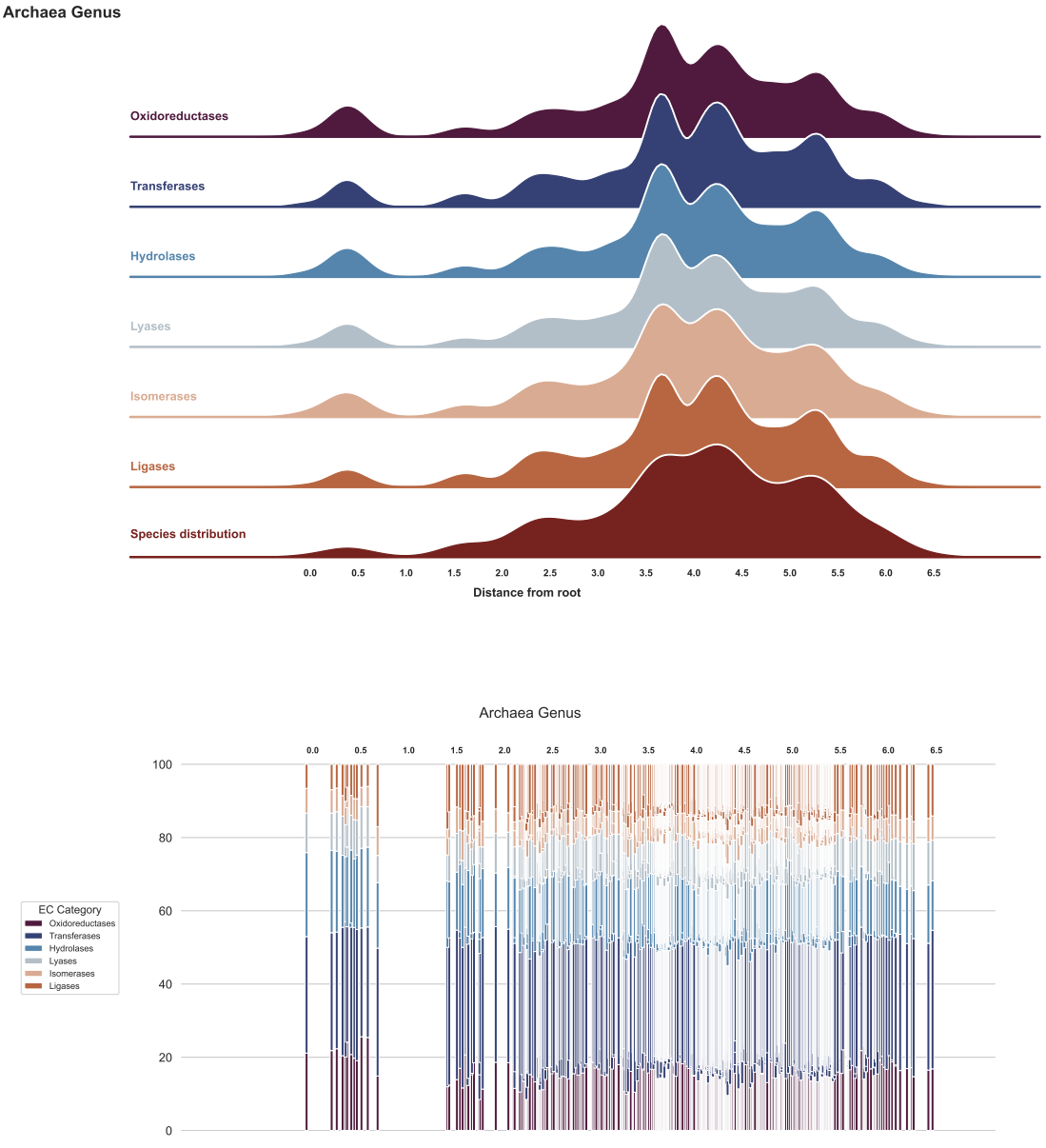


Figure A.13: Evolution of individual enzyme categories for the genus-level archaea dataset. The ridgeplot displays the distribution of each category as a function of distance from the tree root, with the last axis presenting the species distribution, while the barplot the relative abundance of each category at that particular distance as a stacked barplot.

A3. Expansion Scope Size as a Function of Distance to Root

For every taxonomic level dataset, per seed set.

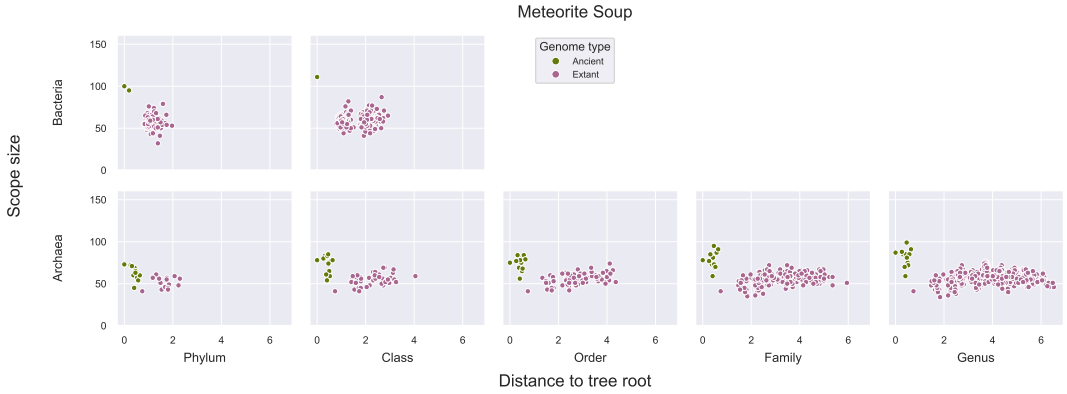


Figure A.14: Goldford + meteoritic soup

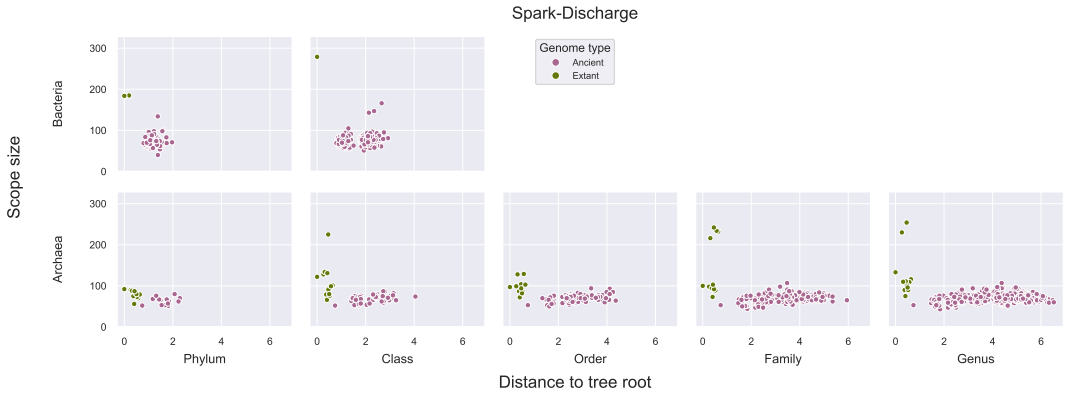


Figure A.15: Goldford + spark-discharge

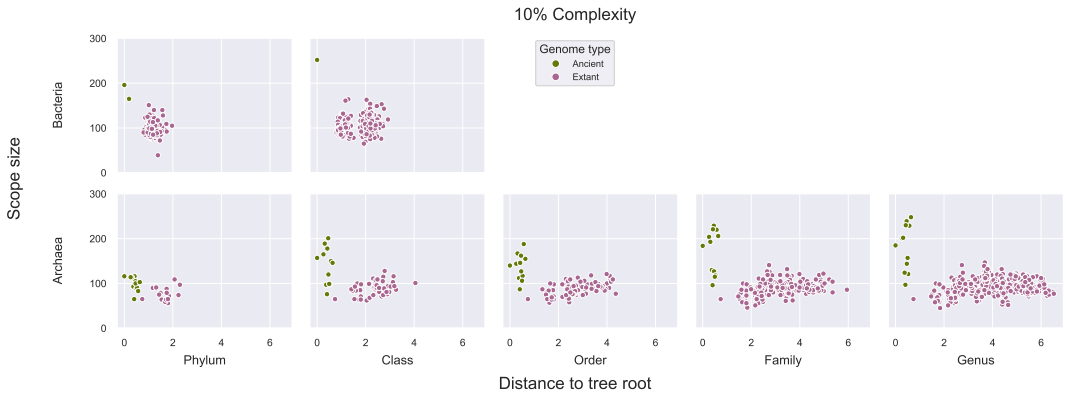


Figure A.16: 10% complexity

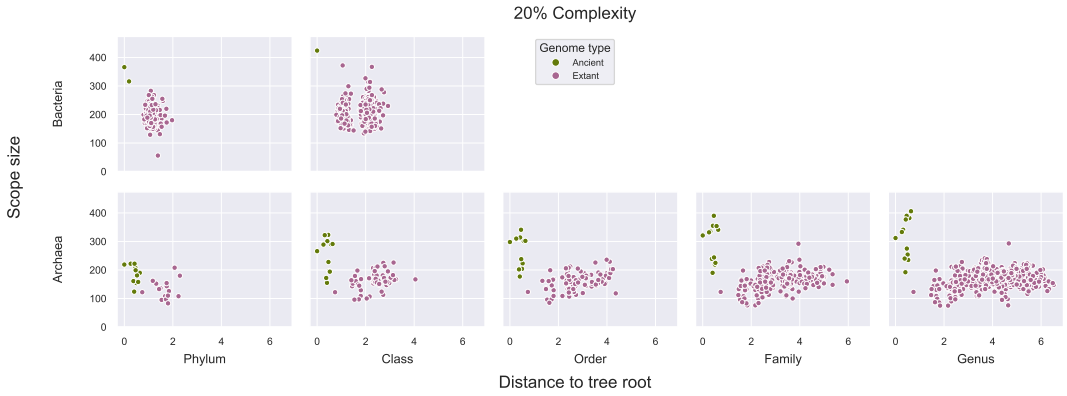


Figure A.17: 20% complexity

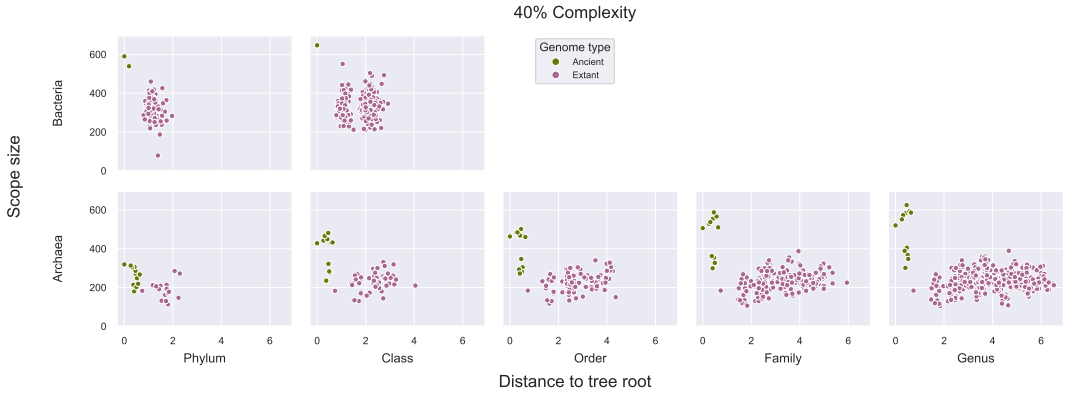


Figure A.18: 40% complexity

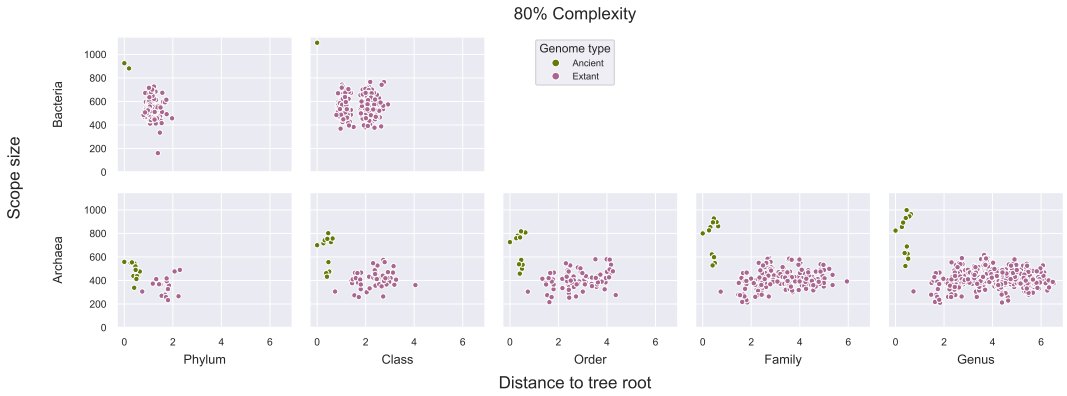


Figure A.19: 80% complexity

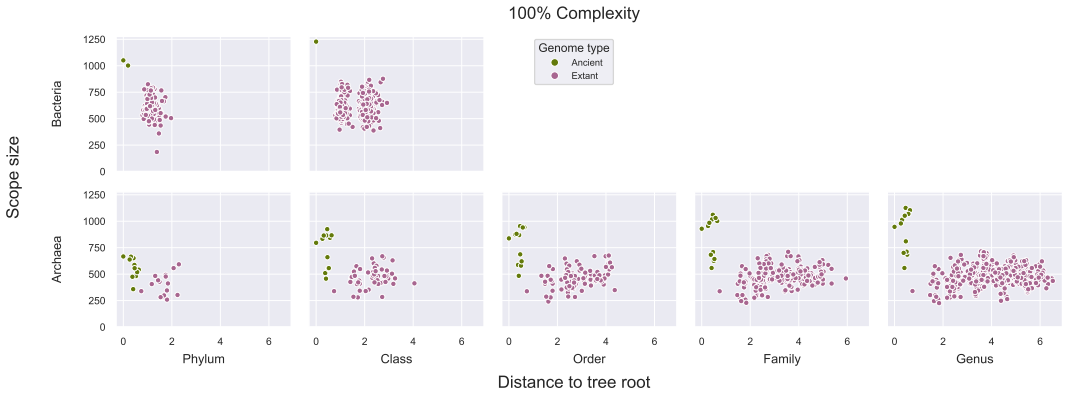


Figure A.20: 100% complexity

For every seed set, per taxonomic level dataset.

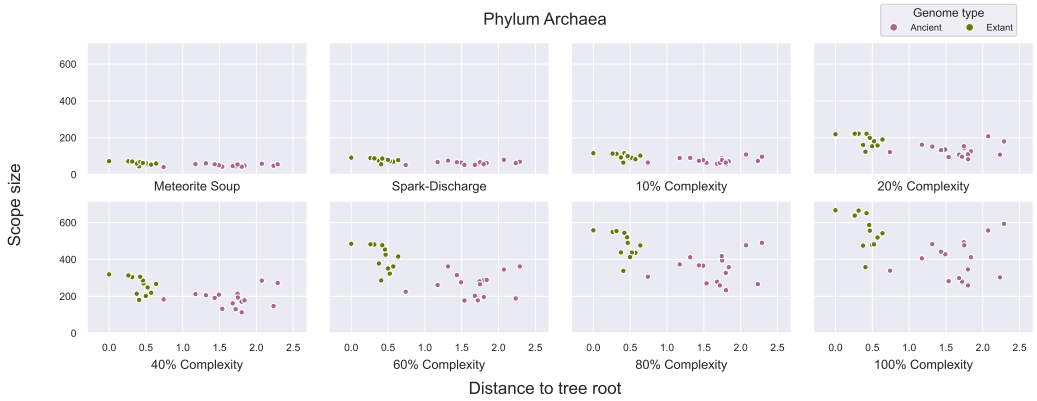


Figure A.21: Phylum level archaea

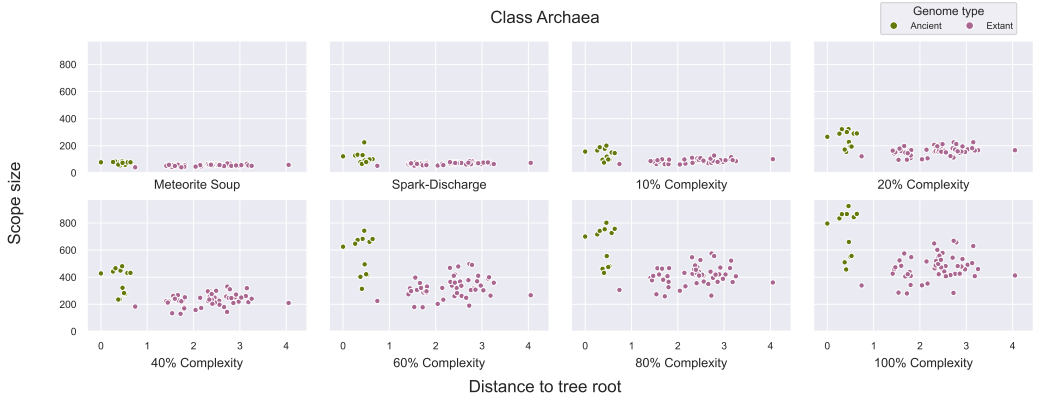


Figure A.22: Class level archaea

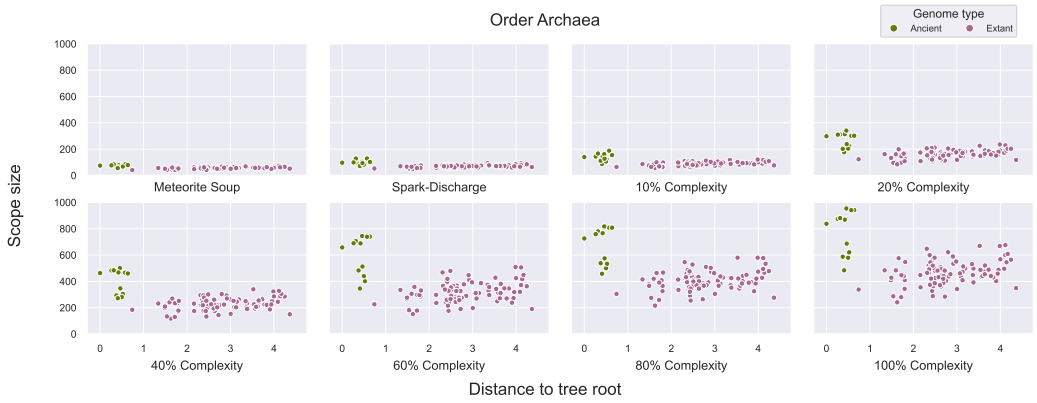


Figure A.23: Order level archaea

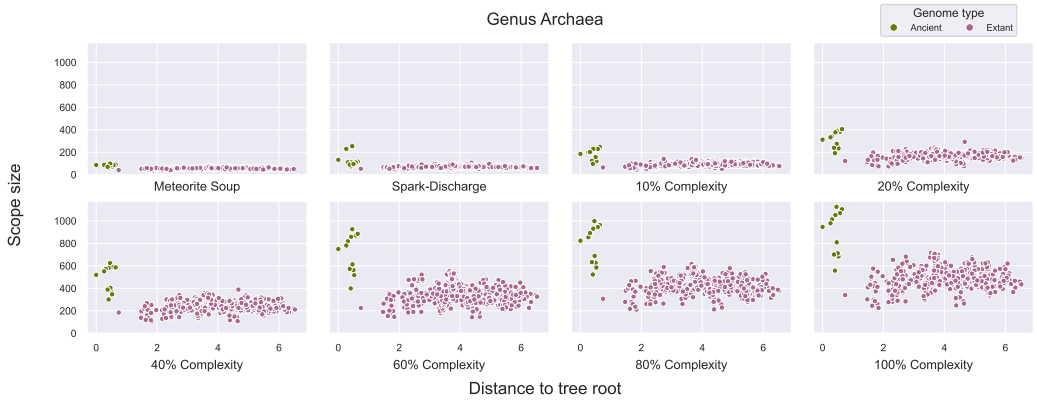


Figure A.24: Genus level archaea

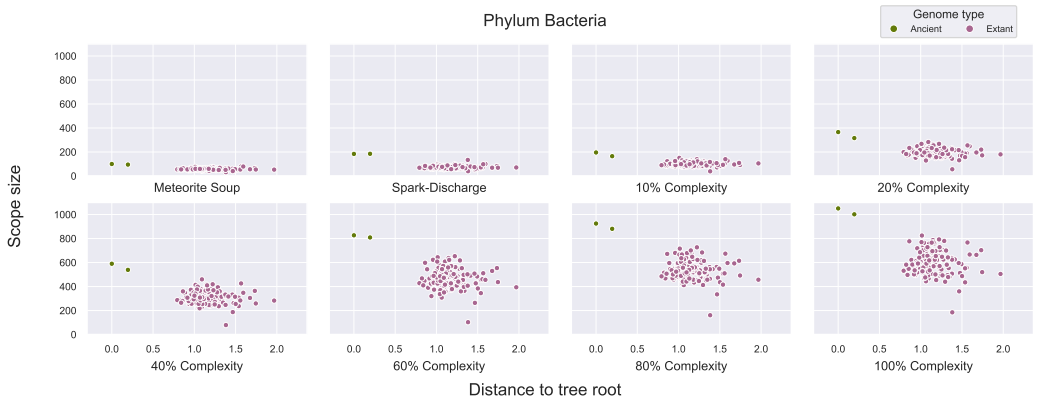


Figure A.25: Phylum level bacteria

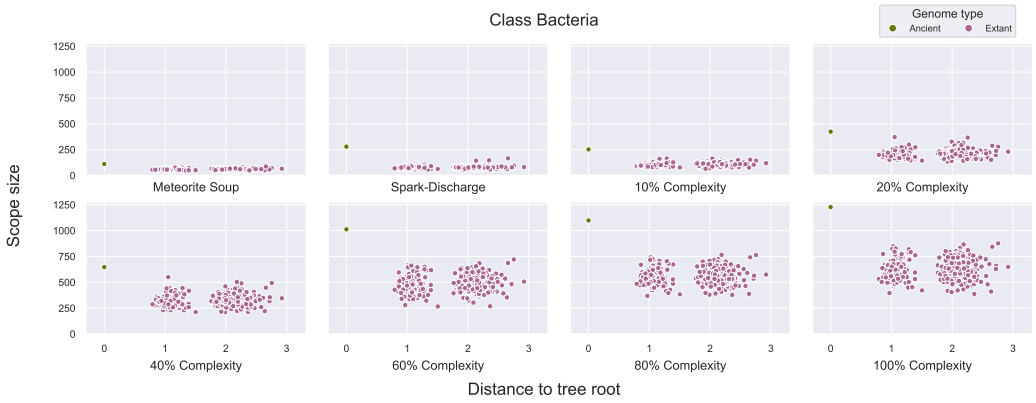


Figure A.26: Class level bacteria