Applying Bioinformatics and Machine Learning Techniques to Investigate the Cultivability and Slow Growth of Uncultured Microbes

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Iyanu Mumeen Oduwole

December 2025

DEDICATION

I dedicate this work to Almighty Allah for his guidance and grace and to my family most especially my mother and my wife for their unwavering support throughout this journey

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisors, Dr. Scott Emrich and Dr. Andrew Steen, for their unwavering support, mentorship, and guidance throughout my Ph.D. journey. I am especially thankful to Dr. Andrew Steen for continuing to support me financially and academically, even after his transition to USC, and for his steadfast belief in my ability to complete the program. I am also deeply grateful to Dr. Tatiana Vishnivetskaya and Dr. Tongye Shen for serving on my dissertation committee and offering valuable feedback. Special thanks to Dr. Karen Lloyd, Dr. Taylor Royalty, and Dr. Ashley Babjac for their contributions to my first dissertation chapter, and to Dr. Adrienne Hoarfrost for her significant input on the final chapter. I extend heartfelt appreciation to my family, especially my mother, Taibat Oduwole, and my beloved wife, Monsurat Aina for being my greatest sources of strength and encouragement. I would also like to acknowledge my GST cohorts (Eugene, Tasneem, and Keertana) and lab mates for their mental support throughout the program. Finally, I thank my Nigerian friends for their continued encouragement, and the Muslim community of Knoxville for providing a spiritual home that nurtured my faith during this journey.

ABSTRACT

Uncultured microbes constitute the vast majority of microbial diversity on Earth, presenting significant obstacles to understanding their ecological roles and physiological adaptations through traditional cultivation techniques. The advent of metagenome-assembled genomes (MAGs) has greatly expanded our insights into uncultured microbes, but major challenges remain in linking genomic traits to culturability and growth potential. This dissertation addresses a subset of these challenges by investigating genomic traits associated with culturability, exploring unexpected slow growth adaptation in cold environments and applying deep learning models to predict microbial growth rate directly from the genomic sequences. In the first part of this work, I analyzed 52,515 MAGs from the Genomes from Earth's Microbiomes (GEM) catalog to examine functional novelty and culturability across phylogenetic lineages. Uncultured MAGs, particularly among Archaea, were found to encode significantly more divergent proteins. Using pathway enrichment analysis, LASSO regression, and permutation-based feature importance, I identified a core set of genomic traits predictive of culturability, suggesting that culturability is linked to conserved functional signatures across microbial lineages. Cultured MAGs were enriched in vitamin and cofactor biosynthesis, which were also strong predictors of microbial culturability. The second part of this study focuses on permafrost soils, where energy availability is chronically low. Here, I tested the hypothesis that thermophilic traits may confer an advantage to extremely slow-growing microbes in cold environments. Genomic indicators of thermophilicity were positively correlated with predicted doubling times in permafrost, revealing

a counterintuitive pattern in which thermophilic traits appear to be adaptive for persistence under slow-growth conditions. This trend was reversed in seasonally thawed active layer soils, where periodic energy influx supports faster microbial growth. To improve prediction of microbial growth rate, I developed *LGMrib*, a transformer-based deep learning model fine-tuned on ribosomal protein-coding sequences. This model significantly outperforms traditional codon usage-based methods, particularly for slow-growing microbes. Applied across 25,000+ genomes, it reveals pervasive cultivation biases toward fast growers across diverse environments. Collectively, this work integrates computational and machine learning techniques to advance our understanding of microbial growth potential, culturability, and environmental adaptation, also offering frameworks for guiding cultivation strategies and predicting microbial traits from metagenomic data.

TABLE OF CONTENTS

CHAPTER I 1

INTRODUCTION 1

Background 1

Uncultured microbes are the majority and phylogenetically distinct 1

Culture independent Techniques: Recovering Metagenomes Assembled Genomes (MAGs) 3

Abundance of uncultured microbes in extreme habitats 3

Predicting microbial phenotypic traits directly from sequence data using computational and machine learning approach 4

Overview of research in this dissertation 5

References 8

Appendix I 13

Figures 13

CHAPTER II Functional Genomic Signatures Predict Microbial Culturability Across The Tree Of Life 15

Abstract 16

Introduction 17

Results 19

Cultured MAGs resemble cultured isolates 19

Uncultured MAGs harbor substantially greater functional novelty 19

Alignment metrics reflect genome source and quality 20
Phylogenetic resolution predicts functional similarity 21
Cultured microbes are enriched in certain metabolic pathways 21
LASSO regression identified metabolic predictors of culturability across phyla.....23
Culturabilty correlates with growth strategy and phylogeny.....24
Discussion.....25
Methods.....29
Description of data sets.....29
Cultured Status.....30
Comparing The Genes in Cultured Genomes To Uncultured Genomes.....31
Fisher's enrichment test analysis 32
Description of the LASSO regression model pipeline for analyzing cultured status 32
Details of necessary machine learning preprocessing.....33
Details of LASSO-based regression model for predicting cultured and uncultured status ...34
Details regarding calculating and interpreting coefficients, confidence intervals and permutation importance 35
Shuffling data to create a random baseline 36
Evaluating model performance 37
References.....38
Appendix II 47
Figures.....47

Supplemental Results 58
Supplemental Tables 61
Supplemental figures 64
CHAPTER III Thermophilicity correlates with slow growth potential in cold subsurface 73
Abstract 74
Introduction 75
Results 77
Slow growers in permafrost and active layer 77
Thermophilicity is positively correlated with doubling time in permafrost but not in active layer soils 78
Thermophilicity of functional genes is correlated to slow growth in permafrost 79
Thermophilic traits and minimum doubling time are independent of phylogeny in permafrost 80
Thermophilic slow growers share few horizontally transferred genes with mesophilic permafrost microbes 81
Discussion 83
Permafrost shows thermophilic intrinsic adaptation to slow growth 83
Weak phylogenetic signal in traits correlation associated with permafrost. 86
Methods 88
Environmental description 88
Metagenome data sets 89

Growth rate Prediction 90
Optimal growth temperature and optimal temperature of enzymes and thermophilicity prediction 91
Phylogenetic signals and regression 91
Horizontal gene transfer prediction 92
Data availability 92
References 93
Appendix III 109
Tables 109
Figures 110
Supplemental figures 116
Supplemental Table 118
CHAPTER IV 119
Transformer-based Prediction of Microbial Growth Rates from Genomic Data 119
Abstract 120
Introduction 121
Results 123
LookingGlass models indicate that growth signals are stronger in ribosomal related genes 123
LookingGlass model improves minimum doubling time prediction 124
LookingGlass model embeddings differentiate fast growers from slow growers 125

LookingGlass model reveals culturable fast and slow growers across earth environments 126
Discussion 127
Methods 130
Isolates trait dataset preparation 130
COGs and Ribosomal sequences annotation 130
GEM dataset preparation 131
Description of Lookingglass model 132
Lookingglass Model-2 Fine-tuning Procedure 132
Assessing model performance 133
Classification into fast and slow growers 134
Assessing the classification performance 134
Data availability 134
Code availability 134
References 135
Appendix IV 144
Table 144
Figures 145
Supplemental figures 148
CHAPTER V 150
CONCLUSION 150
Future Direction 151

VITA 152

LIST OF TABLES

Table 3.1: The phylogenetic signals of traits in permafrost (SIB_PF) and active layer (SVA_AL). PMDT indicated predicted minimum doubling time; POGT indicates predicted optimal growth temperature. 1094

Table 3.2: PGLS regression of PMDT and POGT in permafrost (SIB_PF) and active layer (SVA_AL)..... 109

Table 4.1: Performance of LookingGlass-derived models finetuned on different annotated gene subsets 1446

Supplemental table 2.1: Counts of cultured and uncultured MAGs with average gene identity < 70% to SwissProt..... 61

LIST OF FIGURES

Figure 1.1: Estimated fractions of uncultured microbes from earth habitats as shown by Lloyd et al. The Graphical representation was adapted from Hug et al. The cultivated lineages (19%) are colored in pink while the remaining (81%) constitute uncultured taxa beginning from the genus level. Also, the top 25% belonging to uncultured phyla are colored in cream. Exception in human microbial populations where only 1% having uncultured phyla. 13

Figure 1.2: Steps to recover Metagenomics Assembled Genomes (MAGs) from metagenomes. DNA from a mixed microbial community is extracted, sequenced, assembled into contigs and binned into draft genomes using bioinformatics tools (astrobiomike.github.io). 14

Figure 2.1: Similarity of predicted protein sequences in MAGs and isolates to SwissProt entries. A and C show normalized bitscores and percent identities, respectively, for archaeal genomes; B and D show the same metrics for bacterial genomes. The distributions represent the average similarity of all predicted protein-coding genes to their closest match in the SwissProt. Random sequences (gray) represent simulated protein sequences with no known homology and serve as a negative control. 476

Figure 2.2: Similarity of predicted proteins in Archaea (left) and Bacteria (right) to their closest SwissProt homolog, stratified by taxonomic proximity to cultured isolates. For each metagenome-assembled genome (MAG) in the GEM catalog, the lowest shared taxonomic rank with a cultured isolate was determined. Protein similarity—measured as percent identity to the nearest SwissProt match—is shown as a function of this taxonomic

proximity. Taxonomic rank assignments for GEM MAGs are based on the Genome Taxonomy Database (GTDB), as implemented by Nayfach et al. Cultured lineages were identified based on RefSeq annotations in the same GTDB release, excluding genomes derived from single-cell amplification or other metagenome-assembled genomes (MAGs or SAGs)..... 487

Figure 2.3: COG pathways that are enriched in cultured and uncultured MAGs (P < 0.005, fisher's exact test). Pathways related to co-factors, vitamins, energy production, and CRISPR Cas are over-represented in cultured MAGs. 508

Figure 2.4: Clusters of Orthologous Groups (COG) pathways that are significantly enriched in cultured MAGs in at least one phylum and not enriched in uncultured MAGs in any phylum. Each column in the plot represents a COG pathway, and each row represents a phylum. Solid circles indicate phyla where the corresponding pathway is significantly enriched in cultured MAGs (Fisher's exact test, FDR-adjusted P < 0.05). 51

Figure 2.5: Machine learning (LASSO) predicts important COG pathways that are predictive of microbial cultured or uncultured status. Pathways are filtered based on the permutation importance. Only pathways with permutation importance (predictive power > 0.1) are included in the plot..... 52

Figure 2.6: Machine learning (LASSO) predicts COG pathways that are strongly predictive of cultured status by phylum. Only COG pathways that predict cultured status with high permutation importance (>0.1) are included in the plot. Solid circles indicate the phyla for which this COG pathway is significantly predicts cultured status. COG pathways that are

both enriched in cultured MAGs and predict microbial cultured status are starred and are predominantly vitamin biosynthetic pathways. 54

Figure 2.7: Heat map showing COG pathways predicting cultured (blue) and uncultured (red) species within each phylum. The intensity of the color represents the strength of the prediction (see methods). Various clusters are numbered and phyla in which fast growers are known to be abundant are highlighted in green. Microbes that are known to be fast growers and have higher cultured representatives cluster together while slow growers show greater variability in clustering. 56

Figure 3.1: Correlations between predicted minimum doubling time and either (a) predicted growth temperature by TOME; or (b) thermophilic index in permafrost and active layer from the Tibetan Plateau alpine region. Thermophilic index is the fraction amount of amino acid set: Ile, Val, Tyr, Trp, Arg, Glu, Leu (IVYWREL) which serves as indicator for thermophilicity and thermostability Abbreviations: AP_AL - alpine active layer; AP_PF - alpine permafrost. 110

Figure 3.2: Thermophilicity is correlated with minimum doubling time. (a) Predicted minimum doubling time (hrs) of microbes from permafrost and active layer using gRodon (Weissman et al., 2021). The vertical dashed lines represent the median of the predicted minimum doubling time. Abbreviation as follows: SVA_AL - Svalbard active layer; STO_AL - Stordalen Mire palsa active layer; SIB_PF - Siberian permafrost. (b-d) Relationship between thermophilicity and growth rate in permafrost and active layer soils. Thermophilicity was measured by predicted optimal growth temperature as predicted by

TOME (G. Li et al., 2019) (b), by the optimal activity temperature for NDPK as predicted by TOMER (Gado et al., 2020), and as the thermophilic index. All correlation analyses were two tailed and the shaded areas around the regression line represent the 95% confidence interval. 111

Figure 3.3: Correlations between predicted minimum doubling time and thermophilic index either specific to (a) annotated carbohydrate transport and metabolism genes (NCBI COG category 'G') or specific to (b) energy production and conversion genes (NCBI COG category 'C'). The correlation analyses in a and b are two tailed. The shaded areas around the regression line show the 95% confidence interval. 113

Figure 3.4: Predicted horizontal gene transfer events within the permafrost and active layer microbial communities. Bands connect the donors and recipients. The band's color corresponds to the donors and the band width represents the number of HGTs. The customized groups include Thermophilic slow growers, TS (predicted optimal growth temperature > 40°C, predicted minimum doubling time > 5 hrs), Thermophilic fast growers, TF (predicted optimal growth temperature > 40°C, predicted minimum doubling time < 5 hrs), Mesophilic slow growers, MS (predicted optimal growth temperature between 20°C and 40°C, predicted minimum doubling time > 5 hrs), Mesophilic fast growers, MF (predicted optimal growth temperature between 20°C and 40°C, predicted minimum doubling time < 5 hrs). The 5 hrs cut off is based on gRodon classification of slow and fast growers (Weissman et al., 2021). The number outside the grid represents the matrix percentage of each donor. 114

Figure 4.1: Comparison of predicted versus empirical minimum doubling times using three models: the fine-tuned LookingGlass model on ribosomal protein sequences (LGM_rib), the original gRodon, and gRodon with temperature correction (gRodon_tmp). LGM_rib achieved the lowest root mean square error (RMSE) and the highest correlation with observed doubling times. The black dashed line represents the identity line, indicating perfect agreement between predicted and actual values. All models tend to underestimate doubling times for slow-growing microbes, as shown by the clustering of points below the identity line at higher doubling times. The red dashed line marks the 5-hour threshold used by gRodon to distinguish fast- and slow-growing organisms..... 145

Figure 4.2: A 5-hour doubling time threshold improves microbial growth classification and cluster separation in LGM_rib embeddings. Classification performance using three doubling time thresholds (2.6, 5, and 8 hours): a) AUC_ROC b) Precision_Recall for identifying fast growers c) Precision_Recall for identifying slow growers. The LGM_rib model achieves the highest AUC-ROC at the 5-hour threshold. (d–f) t-SNE projections of LGM_rib embeddings show clustering of fast (blue) and slow (orange) growers under different classification thresholds: (b) 5 hours, (c) 2.6 hours (empirical median), and (d) 8 hours (75th percentile). The 5-hour threshold results in the clearest separation between fast and slow growers, with more distinct clustering of fast-growing microbes. Points from training and test datasets are marked separately. 146

Figure 4.3: LGM_rib predicts minimum doubling time of microbes inferred to be cultured (green) and uncultured (gray) in GEM dataset across a) ecosystem and b) GEM phyla. The

vertical dashed blue line represents a doubling time of 5 hours, distinguishing fast growers from slow growers. Only phyla with over 50 representatives and ecosystem with 75 representatives are shown in the plot. 147
Supplemental Figure 2.1: Distribution of mean of alignment length of MAGs, isolates, and random sequences to SwissProt in A) Archaea and in B) Bacteria. C and D represent the relationship between alignment length and identity to SwissProt of genes (1000 samples) from cultured isolates, uncultured MAGs, cultured MAGs Archaea and Bacteria respectively 64
Supplemental Figure 2.2: Cumulative fractions of taxonomic cultured representatives in GEM data sets. The GEM MAG lowest taxonomic rank was assigned by Nayfach et al. (2021) using all Genome Taxonomy Database (GTDB) reference genomes (Chaumeil et al., 2020) while the taxonomic cultured levels were assigned based on the RefSeq category of the same GTDB release excluding single cell amplified genomes and the assembled metagenomes..... 65
Supplemental Figure 2.3: Proportion of genes assigned by the Agnostos DB clusters profile search database at different GEM taxonomic cultured levels. 66
Supplemental Figure 2.4: Permutation importance by GEM phyla. Pathways from less cultured phyla tend to have COG pathways with higher permutation importance. 67
Supplemental Figure 2.5: Machine learning workflow for LASSO regression analysis. We created a custom dataset containing metadata, pathway features, annotation features, and cultured labels (See Methods). Each of these data were then split into subsets representing

their individual phyla and the metadata were removed from the subsets. We removed any subsets that did not meet the minimum data requirements (see Methods). We lastly ran our LASSO pipeline 5 times on each dataset and agglomerate the counts of significant phyla into a final output file. 68
Supplemental Figure 2.6: Performance metrics (Accuracy) for GEM annotations and pathways. A value of 0.0 is the worst and 1.0 is the best. The orange represents real data (i.e. shuffled=" No") and the blue represents our random baseline (i.e. shuffled="Yes"). Phyla are omitted from interpretation if the margin of error of shuffled performance is within our actual predictions across all metrics. 69
Supplemental Figure 2.7: Performance metrics (AUC) for GEM annotations and pathways. A value of 0.5 is worst for AUC. The orange represents real data (i.e. shuffled=" No") and the blue represents our random baseline (i.e. shuffled="Yes"). Phyla are omitted from interpretation if the margin of error of shuffled performance is within our actual predictions across all metrics. 70
Supplemental Figure 2.8: Precision for GEM annotations and pathways. A value of 0.0 is worst while a value of 1.0 is the best. The orange represents real data (i.e. shuffled=" No") and the blue represents our random baseline (i.e. shuffled="Yes"). Phyla are omitted from interpretation if the margin of error of shuffled performance is within our actual predictions across all metrics. 71
Supplemental Figure 2.9: Recall for GEM annotations and pathways. A value of 0.0 is worst while a value of 1.0 is the best. The orange represents real data (i.e. shuffled=" No") and the

blue represents our random baseline (i.e. shuffled="Yes"). Phyla are omitted from interpretation if the margin of error of shuffled performance is within our actual predictions across all metrics. 72

Supplemental Figure 3.1: Comparing the predicted minimum doubling time of all the active layers and permafrost to human gut. MAGs in the active layer and permafrost have higher minimum doubling time (slower growth rate) than that of the human gut. 116

Supplemental Figure 3.2: Correlation of minimum doubling time with the counts of annotated genes relating to energy production and conversion (a) and carbohydrate transports and metabolism (b). MAGs with predicted slower growth rates have lesser number of genes in energy production and carbohydrate transports. 117

Supplemental Figure 4.1: Classification metrics for differentiating fast from slow growers using different empirical minimum doubling time thresholds. The vertical dashed line at 5hrs indicates the gRodon threshold. The metrics were computed using logistic regression trained on sequence embeddings from the Lookingglass models. At 5hrs, the AUC and PR-AUC already reached the plateau indicating a pragmatic cutoff for fast and slow growers. 148

Supplemental Figure 4.2: Cultured microbes are biased towards fast growers except in human associated environments. Predicted minimum doubling time of a) all GEM microbes b) Excluding human related environments. The dashed line indicates 5hrs threshold which differentiates fast growers from the slow growers. 117

CHAPTER I

INTRODUCTION

Background

Uncultured microbes are the majority and phylogenetically distinct

Various estimations, coupled with the exponential increase in genomic data, have revealed that prokaryotes primarily divided into the domains Archaea and Bacteria are taxonomically diverse, with estimates suggesting the existence of hundreds to thousands of distinct phyla (Hug et al., 2016; Lewis et al., 2021; Rinke et al., 2013; Yarza et al., 2014). Based on 16S rRNA gene sequences, current estimates place the number of microbial species at approximately 400,000 and the number of genera at around 60,000. However, these figures account for less than 1% of the actual microbial diversity on Earth, which is believed to exceed these numbers by several orders of magnitude (Curtis et al., 2002; Lennon & Locey, 2020).

Utilizing comprehensive genomic datasets, including 16S rRNA sequences, metagenomes, and metatranscriptomes, Lloyd et al. (2018) estimated that only 19% of microbial cells from environmental samples have any phylogenetic affiliation with cultured representatives (Lloyd et al., 2018)(**Figure 1.1**). Importantly, this estimate excludes microbial population associated with human environments, further emphasizing the vast underrepresentation of environmental microbes in culture collections. Lloyd et al. further proposed a conceptual framework distinguishing between two major groups of uncultured microbes: (1) viable but non-culturable (VBNC) cells, which may regain

culturability under specific growth conditions, and (2) phylogenetically divergent non-cultured (PDNC) lineages, which likely represent organisms that are fundamentally difficult or impossible to grow under conventional laboratory conditions. PDNC organisms are often characterized by high phylogenetic novelty and may possess unique biological traits such as extremely slow growth rates, oligotrophic lifestyles, or obligate syntrophic dependencies that prevent them from thriving in standard axenic cultures (Browne et al., 2016; Carini, 2019; Connon & Giovannoni, 2002).

These features imply that the physiology and ecology of PDNC cells may be drastically different from those of culturable taxa. As such, researchers increasingly rely on culture-independent approaches, such as metagenomics, single-cell genomics, and stable isotope probing, to infer the functional potential and ecological roles of these elusive microbes (Rinke et al., 2013; Solden et al., 2016). Recent innovations like the iChip and co-culturing techniques have made strides in increasing the number of cultivable species by mimicking natural environments and enabling small molecule exchanges critical for microbial growth (Berdy et al., 2017; Nichols et al., 2010), but the majority of environmental microbes remain recalcitrant to cultivation. Therefore, advancing our understanding of Earth's microbial diversity, particularly that of uncultured and poorly characterized taxa requires cultured independent techniques and continued development and application of computational, multi-omics, and ecological modeling frameworks that go beyond traditional cultivation-based methods.

Culture independent Techniques: Recovering Metagenomes Assembled Genomes (MAGs)

Metagenomics is a powerful culture-independent approach which allows recovery of partial or nearly complete microbial genomes from environmental samples (Handelsman, 2004; Quince et al., 2017). Metagenomics approach could offer insight into microbial diversity, metabolic potential and novel genes in any mixed microbial communities. This technique usually starts with extracting genomic DNA from mixed microbial community which is then sequenced in most cases using random shortgun sequencing approach to generate millions of short DNA fragments referred to as "reads" (Thomas et al., 2012; Tyson et al., 2004). Using computational assembling approach, these short reads are then assembled into longer contiguous sequences known as contigs which are overlapping DNA that represents portion of microbial genomes in the sample (Alneberg et al., 2014; Sczyrba et al., 2017). Based on sequence characteristics such as nucleotide composition and sequence coverage, these contigs can be grouped into bins refers to Metagenomes Assembled Genomes (MAGs) (Kang et al., 2019; Uritskiy et al., 2018). A MAG gives genomic access to the individualized microbial species or strains in the microbial community and can serve as a proxy to cultured isolates in the laboratory (**Figure 1.2**).

Abundance of uncultured microbes in extreme habitats

Extreme habitats such as permafrost which has subzero temperature, limited nutrient availability presents significant challenge to microbial survival and growth (Altshuler et al., 2017). Despites these harsh constraints, unique microbial life thrives in

this environment, although it remains largely understudied due limited cultured representatives. Conventional cultivation techniques are biased towards fast growing microbes and often overlook the slow growing ones which are mostly adapted to extreme conditions (Weissman et al., 2021). To understand the ecological roles, metabolic capabilities and adaptive strategies of these microbes, metagenomics approach involving of binning MAGs offer a comprehensive approach for characterizing microbial diversity, uncovering novel genomic traits and metabolic pathways for understanding the ecological roles of these microbes such as in biogeochemical cycling,

Predicting microbial phenotypic traits directly from sequence data using computational and machine learning approach

Most microbes do not have closely related, experimentally characterized cultured relatives, which limits the ability to infer physicochemical properties such as growth temperature, minimum doubling time, and pH of uncultured microbes. As a result, the vast majority of microbial diversity remains unexplored. Recent advances in sequencing technologies, coupled with a dramatic decrease in sequencing costs, have exponentially increased the availability and diversity of microbial genomic data. For instance, the Genome Taxonomy Database (GTDB) currently contains over 250,000 bacterial genomes and over 4,000 archaeal genomes (Parks et al., 2018). Some of the largest metagenomic datasets known include the Genomic Earth Microbiomes (GEM) dataset, which contains 52,515 MAGs binned from 10,540 metagenomes across diverse environments (Nayfach et al., 2021), and the OceanDNA MAGs, comprising 53,325 MAGs assembled from 2,057 marine metagenomes (Nishimura & Yoshizawa, 2022).

These large collections of genomic datasets have opened new opportunities to apply computational methods that predict functional microbial traits without the need for experimental annotations. For instance, Li et al. (2019) developed the Temperature Optima for Growth and Enzymes (TOGE) model using a support vector regression approach based on dipeptide frequencies to predict microbial growth temperature (Li et al., 2019). Similarly, Barnum et al. (2021) applied machine learning models to infer optimum growth temperature, pH, oxygen tolerance, and salinity directly from amino acid composition features of genomes (Barnum et al., 2024). gRodon is another example that uses codon usage bias to infer microbial growth rate from genomic sequences (Weissman et al., 2021).

Furthermore, deep learning techniques have enabled the development of foundational biological models for various applications. For example, EvoFormer-2 was trained on trillions of nucleotide sequences from over 128,000 genomes spanning humans, plants, microbes, and other species (Brixi et al., 2025). Transformer-based biological models are increasingly common and include ESM-2, trained on millions of protein sequences from the UniProt database (Lin et al., 2023), and the LookingGlass model, trained on billions of nucleotide sequences from the GTDB database. (Hoarfrost et al., 2022).

Overview of research in this dissertation

In these dissertation chapters, cultivability and slow growth adaptation of uncultured microbes were investigated using MAGs. The first chapter utilizes GEM dataset to examine if cultured microbes are biologically distinct from uncultured

microbes through sequence and taxonomic novelties. Key pathways that are predictive of microbial culturedness were identified using LASSO regression machine learning model. This chapter showed that uncultured microbes have higher genomic novelty, characterized by slow growth and certain metabolic pathways such as co factors and vitamin biosynthesis can predict microbial culturedness.

The second chapter focuses on uncultured microbes in permafrost. MAGs from permafrost and active layer environments which include Alpine, Siberia, Svalbard and Stordalen Mire were analyzed to test the hypothesis that thermophile-like traits are adaptive strategies for slow growth in frozen environments. First, computational predictions were performed to estimate minimum doubling time, optimal growth temperature, and the optimal temperature of the nucleoside diphosphate kinase (NDPK) enzyme. In permafrost, thermophilicity was found to correlate positively with minimum doubling time, suggesting that higher thermophilic traits are associated with slower-growing microbes. Conversely, in the active layer, thermophilicity correlated with faster growth. Furthermore, these traits exhibited low phylogenetic signals in permafrost but not in the active layer, suggesting that the observed relationships are independent of phylogeny. Finally, thermophilic slow growers in permafrost shared fewer horizontally transferred genes with mesophilic taxa compared to thermophilic fast growers in the active layer which indicates that slow growing thermophiles may have distinct evolution.

The third chapter applies a transformer based deep learning model, LookingGlass model to predict microbial growth rate. The model was fine-tuned using phenotypic trait data, specifically minimum doubling time and growth temperature, and then applied to

predict the doubling time of unknown microbes. Among the different gene sets evaluated, models trained on ribosomal sequences outperformed the benchmark method gRodon, a widely used codon usage-based growth rate predictor. The study validated the 5-hour doubling time threshold proposed by gRodon to distinguish fast from slow growers and used the model to predict growth rates for over 25,000 microbes with detectable ribosomal proteins from the GEM dataset. The results confirmed that cultured microbes are predominantly fast growers.

This dissertation shows that computational and machine learning technique can uncover genomic and phenotypic characteristics of uncultured microbes by providing insights that complements and extend beyond traditional cultivation-based methods. Integrating machine learning and bioinformatics can further expand our understanding on microbial adaptation, physiochemical properties and ecological function.

References

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, *11*(11), 1144–1146.

Altshuler, I., Goordial, J., & Whyte, L. G. (2017). Microbial Life in Permafrost. In R. Margesin (Ed.), *Psychrophiles: From Biodiversity to Biotechnology* (pp. 153–179). Springer International Publishing. https://doi.org/10.1007/978-3-319-57057-0_8

Barnum, T. P., Crits-Christoph, A., Molla, M., Carini, P., Lee, H. H., & Ostrov, N. (2024). Predicting microbial growth conditions from amino acid composition. *bioRxiv*, 2024–03.

Berdy, B., Spoering, A. L., Ling, L. L., & Epstein, S. S. (2017). In situ cultivation of previously uncultivable microorganisms using the ichip. *Nature Protocols*, *12*(10), 2232–2242.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., & Merchant, A. T. (2025). Genome modeling and design across all domains of life with Evo 2. *BioRxiv*, 2025–02.

Browne, H. P., Forster, S. C., Anonye, B. O., Kumar, N., Neville, B. A., Stares, M. D., Goulding, D., & Lawley, T. D. (2016). Culturing of 'unculturable'human microbiota reveals novel taxa and extensive sporulation. *Nature*, *533*(7604), 543–546.

Carini, P. (2019). A “Cultural” Renaissance: Genomics Breathes New Life into an Old Craft. *mSystems, 4*(3), e00092-19. https://doi.org/10.1128/mSystems.00092-19

Connon, S. A., & Giovannoni, S. J. (2002). High-Throughput Methods for Culturing Microorganisms in Very-Low-Nutrient Media Yield Diverse New Marine Isolates. *Applied and Environmental Microbiology*, *68*(8), 3878–3885. https://doi.org/10.1128/AEM.68.8.3878-3885.2002

Curtis, T. P., Sloan, W. T., & Scannell, J. W. (2002). Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences, 99*(16), 10494–10499. https://doi.org/10.1073/pnas.142680199

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, *68*(4), 669–685. https://doi.org/10.1128/MMBR.68.4.669-685.2004

Hoarfrost, A., Aptekmann, A., Farfañuk, G., & Bromberg, Y. (2022). Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nature Communications*, *13*(1), 1–12.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., & Ise, K. (2016). A new view of the tree of life. *Nature Microbiology, 1*(5), 1–6.

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ, 7*, e7359.

Lennon, J. T., & Locey, K. J. (2020). More support for Earth's massive microbiome. *Biology Direct*, *15*(1), 5. https://doi.org/10.1186/s13062-020-00261-8

Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z., & Ettema, T. J. (2021). Innovations to culturing the uncultured microbial majority. *Nature Reviews Microbiology*, *19*(4), 225–240.

Li, G., Rabe, K. S., Nielsen, J., & Engqvist, M. K. (2019). Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. *ACS Synthetic Biology*, *8*(6), 1411–1420.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574

Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J., & Crosby, L. (2018). Phylogenetically novel uncultured microbial cells dominate earth microbiomes. *MSystems*, *3*(5), e00055-18.

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., & Huntemann, M. (2021). A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, *39*(4), 499–509.

Nichols, D., Cahoon, N., Trakhtenberg, E. M., Pham, L., Mehta, A., Belanger, A., Kanigan, T., Lewis, K., & Epstein, S. (2010). Use of ichip for high-throughput in

situ cultivation of “uncultivable” microbial species. *Applied and Environmental Microbiology, 76*(8), 2445–2450.

Nishimura, Y., & Yoshizawa, S. (2022). The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. *Scientific Data*, *9*(1), 305.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, *36*(10), 996–1004.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), 833–844.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., & Gies, E. A. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–437.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., & Dahms, E. (2017). Critical assessment of metagenome interpretation—A benchmark of metagenomics software. *Nature Methods*, *14*(11), 1063–1071.

Solden, L., Lloyd, K., & Wrighton, K. (2016). The bright side of microbial dark matter: Lessons learned from the uncultivated majority. *Current Opinion in Microbiology, 31*, 217–226.

Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics—A guide from sampling to data analysis. *Microbial Informatics and Experimentation*, *2*(1), 3. https://doi.org/10.1186/2042-5783-2-3

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., & Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43.

Uritskiy, G. V., DiRuggiero, J., & Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, *6*(1), 158. https://doi.org/10.1186/s40168-018-0541-1

Weissman, J. L., Hou, S., & Fuhrman, J. A. (2021). Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *Proceedings of the National Academy of Sciences*, *118*(12), e2016810118. https://doi.org/10.1073/pnas.2016810118

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K.-H., Whitman, W. B., Euzéby, J., Amann, R., & Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, *12*(9), 635–645.

Appendix I

Figures

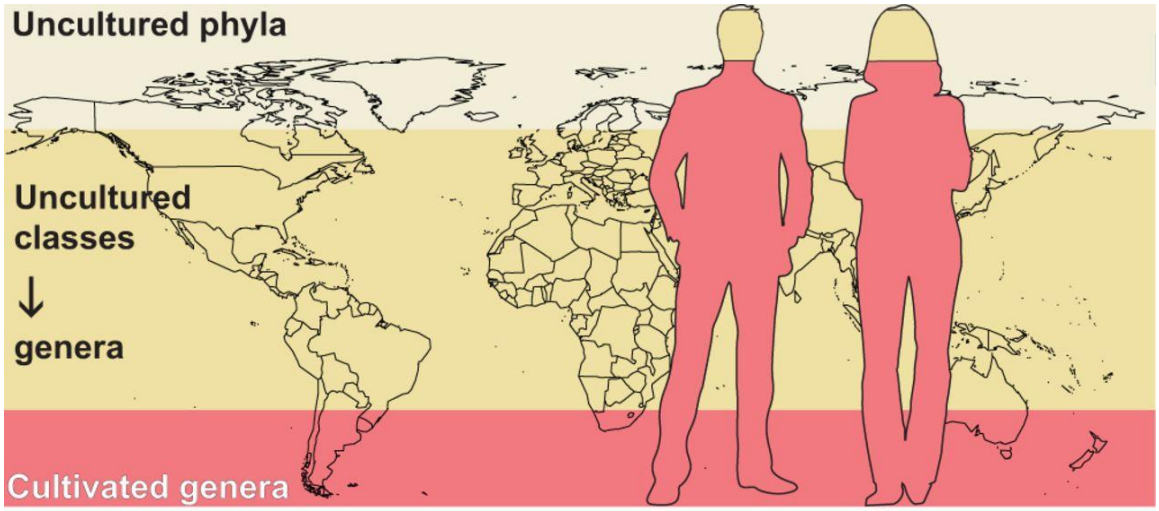


Figure 1.1: Estimated fractions of uncultured microbes from earth habitats as shown by Lloyd et al. The Graphical representation was adapted from Hug et al. The cultivated lineages (19%) are colored in pink while the remaining (81%) constitute uncultured taxa beginning from the genus level. Also, the top 25% belonging to uncultured phyla are colored in cream. Exception in human microbial populations where only 1% having uncultured phyla.

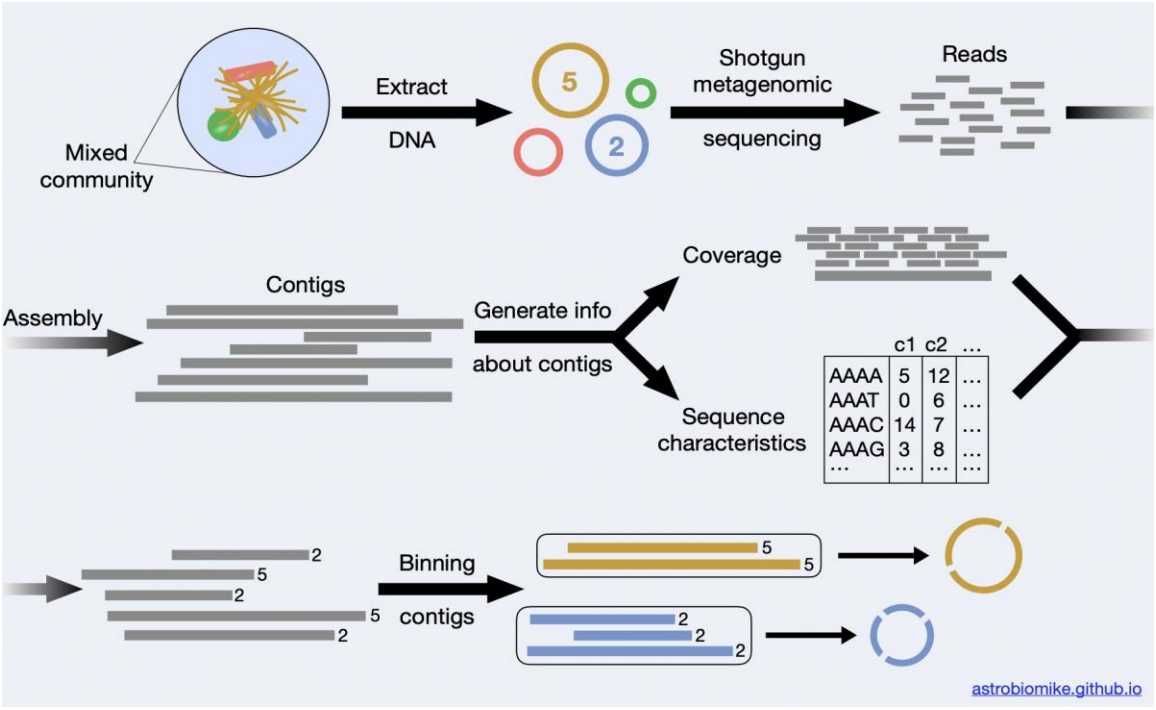


Figure 1.2: Steps to recover Metagenomics Assembled Genomes (MAGs) from metagenomes. DNA from a mixed microbial community is extracted, sequenced, assembled into contigs and binned into draft genomes using bioinformatics tools (astrobiomike.github.io).

CHAPTER II

FUNCTIONAL GENOMIC SIGNATURES PREDICT MICROBIAL CULTURABILITY ACROSS THE TREE OF LIFE

Abstract

Most microbial taxa on Earth remain uncultivated, limiting our ability to study their physiology, ecology, and roles in environmental processes. Although metagenome-assembled genomes (MAGs) have expanded access to uncultured diversity, the genomic basis for culturability remains poorly understood. Here, we analyze the 52,515 MAGs from the Genomes from Earth's Microbiomes (GEM) catalog to test two hypotheses: that genomes from uncultured microbes encode more functionally novel genes than those from cultured taxa, and that specific genomic features are systematically associated with culturability across phyla. To assess functional novelty, we aligned predicted proteins to SwissProt and measured sequence dissimilarity to the nearest curated homolog. We find that uncultured MAGs, particularly among Archaea, harbor substantially more divergent proteins. To identify genomic traits predictive of culturability, we combined pathway-level enrichment with LASSO regression and permutation-based feature importance. Cultured MAGs were consistently enriched in Clusters of Orthologous Groups (COG) pathways related to vitamin and cofactor biosynthesis (e.g., thiamine, folate, B12), energy metabolism (e.g., TCA cycle), and CRISPR-Cas systems—functions often depleted in uncultured counterparts. LASSO models identified a subset of these pathways as strong predictors of cultured status even in poorly sampled phyla, suggesting conserved genomic signatures of culturability. In contrast, pathways such as purine biosynthesis and NADH dehydrogenase were associated with uncultured lineages, highlighting potential barriers to cultivation. Together, these results highlight both the functional novelty of uncultured microbes and candidate metabolic traits associated with

culturability. These insights can also inform future cultivation strategies and improve interpretation of MAGs from diverse environmental lineages.

Introduction

Most microbial cells on Earth belong to taxa that have never been grown in pure culture, despite decades of cultivation efforts (Lloyd et al., 2018; Rinke et al., 2013; Steen et al., 2019). The reasons many microbes remain uncultivated are only partially understood (Lewis et al., 2021; Mu et al., 2021; Vartoukian et al., 2010). This cultivation gap limits our experimental access to microbial diversity and constrains efforts to characterize microbial physiology, ecological roles, and biotechnological potential.

One contributing factor may be uneven culturing effort; some lineages may be underrepresented simply due to limited attempts to grow them from certain environments (Lloyd et al., 2018). A second explanation is that many microbes have complex or context-dependent growth requirements, including dependence on specific environmental cues or close associations with other taxa, which may not easily replicate under standard lab conditions (Berdy et al., 2017; D'Onofrio et al., 2010; Ghoul & Mitri, 2016; Imachi et al., 2020; Kamagata & Tamaki, 2005). Third, inherently slow-growing taxa may be outcompeted in enrichment cultures or may fail to reach detectable levels within typical incubation times (Mu et al., 2021; Prakash et al., 2013).

Although cultivation-independent techniques, particularly metagenome-assembled genomes (MAGs), have transformed our view of microbial diversity and expanded access to uncultured lineages (Garza & Dutilh, 2015; Handelsman, 2004), the genomic basis of microbial culturability remains poorly understood. We reasoned that

comparing cultured and uncultured genomes could identify functional trends associated with culturability and highlight protein-coding novelty in lineages that remain recalcitrant to cultivation. Using over 50,000 MAGs from the Genomes from Earth's Microbiomes catalog (Nayfach et al., 2021), we tested two hypotheses.

First, we hypothesized that uncultured microbes encode more functionally novel genes than cultured taxa. To assess novelty, we aligned predicted proteins to SwissProt, a manually curated database of experimentally characterized proteins and measured amino acid sequence dissimilarity to the closest known homolog. Given that sequence divergence often reflects functional divergence (Dunbrack Jr, 2006; Rost, 1999), we used this metric to infer protein-coding novelty. Second, we asked whether microbial culturability can be predicted from genomic content. Specifically, we examined the distribution of Clusters of Orthologous Groups (COG) pathways across genomes and used LASSO regression (Tibshirani, 1996), a regularized modeling approach that selects a sparse set of predictive features, coupled with permutation-based feature importance (Winkler et al., 2015), to identify genomic pathways most strongly associated with cultured status.

Previous studies have shown that uncultured microbes are phylogenetically distinct from cultured lineages (Lloyd et al., 2018; Salam et al., 2021). Our analysis extends these findings by showing that uncultured lineages —especially among Archaea— harbor extensive functional novelty, while cultured MAGs are consistently enriched in core metabolic pathways, particularly those involved in vitamins biosynthesis and energy metabolism. These patterns suggest that conserved genomic features may

underlie culturability and may provide targets for improved cultivation strategies across phylogenetically diverse microbes.

Results

Cultured MAGs resemble cultured isolates

We analyzed 52,550 metagenome-assembled genomes (MAGs) from the GEM catalog. Of these, 14,620 genomes were classified as "cultured" based on matched taxonomic cultured representatives and sequence similarity to cultured isolates (>95% ANI, see Methods); the remaining genomes were classified as "uncultured." Cultured MAGs and cultured isolates exhibited similar levels of predicted *functional novelty*, defined as amino acid sequence dissimilarity to SwissProt proteins. Prior work suggests that protein sequences with ≥70% identity often retain similar biological functions (Chung & Subbiah, 1996). The comparable sequence similarity levels between cultured MAGs and isolates indicate that the process of sequencing DNA from the environment, assembly, and binning do not introduce substantial artifacts that inflate the apparent novelty (**Figure 2.1 A-D**).

Uncultured MAGs harbor substantially greater functional novelty

In contrast to the cultured MAGs?, uncultured MAGs contained significantly more novel protein-coding genes than either cultured MAGs or isolates, based on both sequence identity and bitscore normalized to alignment length (**Fig. 1A-D**). To assess whether this increased novelty could result from spurious alignments, we compared all results to a control set of randomly generated sequences. All biological sequences—cultured and uncultured—were substantially less divergent than the random set (**Figure**

2.1 A-D), supporting the conclusion that the observed novelty reflects genuine biological differences rather than alignment noise. Only 7% (2,725 out of 37,847) of uncultured MAGs have predicted proteins exhibited more than 70% average amino acid identity with SwissProt homologs, compared to 42% (6,100 out of 14,620) in cultured MAGs and 51% (3,596 out of 7,035) in cultured isolates. These findings suggest that uncultured microbes encode extensive, unexplored protein diversity. Uncultured archaeal MAGs were especially novel, exhibiting greater divergence from SwissProt proteins than uncultured bacterial MAGs relative to their respective cultured counterparts (**Figure 2.1 A-D**). This highlights uncultured archaeal lineages as particularly rich sources of uncharacterized functional diversity.

Alignment metrics reflect genome source and quality

To evaluate whether genome quality contributes to these patterns, we examined alignment length distributions. Cultured MAGs had shorter alignment lengths than isolates (**Supplemental Figure 2.1A**), likely reflecting lower sequencing depth or more fragmented assemblies. As expected, uncultured MAGs had the shortest alignments, consistent with lower estimated completeness.

Across all genomes, alignment length was positively correlated with sequence identity—supporting the idea that more complete proteins yield more confident homology matches. By contrast, alignment length and identity were negatively correlated in random sequences (**Supplemental Figure 2.1B**). Correlation strength was highest in isolates (R = 0.42), intermediate in cultured MAGs (R = 0.18) and lowest in

uncultured MAGs (R = 0.14) (**Supplemental Figure 2.1B**). These findings further support the biological validity of our results concerning functional novelty.

Roughly 9% of predicted proteins across all MAGs (combining Bacteria and Archaea) had no detectable match to SwissProt proteins, even with a permissive *e*-value threshold (*e* <100). These unmatched genes may represent pseudogenes, misannotated open reading frames (ORFs), or real proteins that have no homologues in any database.

Phylogenetic resolution predicts functional similarity

Bacterial MAGs were more frequently assigned to specific taxonomic ranks than archaeal MAGs, consistent with a higher proportion of cultured representatives among bacterial lineages (**Supplemental Figure 2.2**). In both domains, genomes assigned to more specific ranks (e.g., species) encoded proteins that were more similar to SwissProt entries (**Figure 2.2**). This trend suggests that functional annotation is tightly linked to phylogenetic proximity to cultured taxa. A complementary analysis using the Agnostos HMM database yielded consistent results: the proportion of unclassified genes increased with taxonomic distance from cultured lineages (**Supplemental Figure 2.3**). Genomes classified only at high taxonomic ranks (e.g., phylum) had the highest fraction of unannotated genes, whereas those resolved to the species level had the lowest.

Cultured microbes are enriched in certain metabolic pathways

To identify potential genomic features associated with culturability, we compared COG pathway enrichment in cultured versus uncultured MAGs, initially focusing on comparisons within individual phyla. We restricted this analysis to 28 phyla with sufficient numbers of both cultured and uncultured representatives, each containing at

least one COG pathway significantly enriched in cultured or uncultured MAGs as determined by Fisher's exact test.

Every COG pathway was significantly enriched in cultured MAGs in at least one phylum, and 46 out of 60 COG pathways were also enriched in uncultured MAGs in at least one phylum (**Figure 2.3**). In most phyla, cultured MAGs were enriched for pathways involved in cofactor and vitamin biosynthesis, energy metabolism, and the CRISPR-Cas adaptive immunity. By comparison, uncultured MAGs were enriched for pathways such as NADH dehydrogenase, Iron-sulfur (Fe_2S) biosynthesis, and asparagine biosynthesis (**Figure 2.3**). These differences suggest that uncultured microbes may rely on alternative metabolic strategies, reduced biosynthetic capabilities, or niche-specific adaptations that hinder growth under laboratory conditions.

We further identified 15 pathways that were enriched exclusively in cultured MAGs and in no uncultured MAGs across all phyla (**Figure 2.4**). These pathways may serve as genomic markers of culturability. Many were involved in cofactor biosynthesis—for example, thiamine biosynthesis was enriched in cultured members of 7–8 phyla including *Bacteroidota*, *Actinobacteriota*, and *Proteobacteria*. Cobalamin (B12) biosynthesis was enriched in *Firmicutes*, *Verrucomicrobiota*, *Cyanobacteriota*, and *Proteobacteria*. The TCA cycle was enriched in cultured members of the *Proteobacteria*, *Desulfobacterota*, *Fusobacteriota*, and *Firmicutes* groups. These findings suggest that a genome's ability to support core metabolic processes and cofactor biosynthesis may promote microbial growth and survival under axenic laboratory conditions.

LASSO regression identified metabolic predictors of culturability across phyla

To identify genomic features predictive of culturability, we applied LASSO regression to model culturability using the full set of COG pathway annotations. Unlike enrichment analysis, which tests each pathway independently, LASSO jointly considers all features while penalizing model complexity, producing sparse models that retain only the most informative predictors. We focused on pathways with nonzero LASSO coefficients and moderate permutation importance scores (>0.1), which reflect each pathway's contribution to model performance. These pathways were identified in both well-sampled and poorly cultured phyla– including *Acidobacteriota*, *Planctomycetota*, *Spirochaetota*, and *Chloroflexota* (**Supplemental Figure 2.4**), suggesting that LASSO captures generalizable predictors of culturability across microbial diversity.

Consistent with the previous enrichment results, the strongest predictors of cultivability were dominated by cofactor and vitamin biosynthesis pathways. Thiamine, folate and pantothenate/CoA biosynthesis were frequently retained across phyla. For example, pantothenate/CoA biosynthesis predicted culturability in well-represented groups such as *Deinococcota*, *Proteobacteria*, *Spirochaetota*, *Desulfobacterota*, and *Firmicutes_C*. Thiamine and folate biosynthesis also predicted cultured status in multiple groups including *Spirochaetota*, *Desulfobacterota*, and *Synergistota* (**Figure 2.5**). In contrast, purine biosynthesis and NADH dehydrogenase were predictive of uncultured status and showed high permutation importance in multiple phyla, possibly reflecting metabolic dependencies or adaptations incompatible with growth in isolation (**Figure 2.5**).

Culturabilty correlates with growth strategy and phylogeny

A heatmap of LASSO-derived feature importance across phyla (**Figure 2.7**) revealed distinct clustering patterns that partially align with known microbial growth strategies. Fast-growing, frequently cultured phyla—such as *Bacteroidota*, *Firmicutes_C*, *Proteobacteria*, and *Actinobacteriota*—cluster together (**Figure 2.7, clade 2**) and shared similar sets of culturability-associated pathways, suggesting conserved genomic predictors of laboratory growth.

In comparison, slow-growing or poorly cultured phyla such as *Spirocaetota*, *Desulfobacteriota*, *Acidobacteriota*, *Dictyoglomota*, and *Crenaracheaota* form a separate clade (**Figure 2.7, clade 1**) characterized by more variable or phylum-specific predictors. Strains from these phyla may require co-culture, longer incubations, or oligotrophic media that are not routinely used in cultivation experiments.

Extremophilic phyla such as *Halobacteriota*, *Thermotogota*, and *Bipolaricaulota* form their own distinct cluster (**Figure 2.7, clade 3**), suggesting that the genomic features associated with laboratory culturability in these lineages differ from those in mesophiles. This could reflect both biological traits and biases in culturing methods.

Finally, several phyla including *Campylobacteriota*, *Nitrospirota*, *Chloroflexota*, exhibit mixed clustering patterns, containing both fast- and slow-growing lineages. This likely reflects a broader range of ecological niches and growth strategies within these phyla, resulting in more heterogeneous culturability-associated genomic features.

Discussion

All microbes, including the most studied model organisms, contain a substantial fraction of genes with unknown or poorly characterized functions. For example, even *E. coli* K-12 has ~35% of genes lacking experimental annotation (Ghatak et al., 2019). In our analysis of over 52,000 metagenome-assembled genomes (MAGs) from the GEM catalog, we find that this "functional dark matter" (Bernard et al., 2018; Rinke et al., 2013) is substantially more prevalent among uncultured microbes. After comparing uncultured MAGs to both cultured MAGs and isolate genomes, we find that uncultured taxa exhibit markedly greater predicted functional novelty (**Figure 2.1**), as reflected in lower amino acid identity to SwissProt proteins and lower bitscores normalized by alignment length.

This novelty increases with genetic distance from cultured representatives: genes from MAGs that belong to cultured species show the highest similarity to characterized proteins, followed by those within the same genus, and so on (**Figure 2.2**). This trend likely stems from the historical reliance on model organisms, selected in part for their ability to grow in pure culture, for protein function characterization (Garza & Dutilh, 2015; Lewis et al., 2021; Prakash et al., 2021).

Prior work has reported widespread genomic novelty among uncultured microbes from diverse environments (Almeida et al., 2021; Bernard et al., 2018; Godzik, 2011; Nayfach et al., 2021; Rinke et al., 2013; Solden et al., 2016). Our results reinforce and extend this view, showing that uncharacterized genes are not random or artifactual, but are concentrated in lineages phylogenetically distant from cultured taxa and

disproportionately prevalent in archaeal MAGs. The high proportion of genes with unknown function—especially in uncultured archaea—suggests extensive, unexplored functional diversity with likely biological significance (Escudeiro et al., 2022; Pavlopoulos et al., 2023; Rodríguez del Río et al., 2022; Sunagawa et al., 2020).

To explore how genomic content relates to culturability, we examined COG pathway enrichment across 28 microbial phyla. Cultured MAGs were consistently enriched for pathways involved in cofactor and vitamin biosynthesis, energy metabolism (e.g., TCA cycle), and CRISPR-Cas defense (**Figure 2.3**). These cofactor and vitamin biosynthesis likely support autonomous growth in axenic lab conditions. TCA cycle and CRISPR Cas system are fundamental for rapid energy production, growth, and viral defense, potentially explaining their prevalence in highly cultured microbes. The TCA cycle can provide the energy efficiency to support the higher metabolic demands of fast-growing microbes (Berney & Cook, 2010; Szenk et al., 2017) while the CRISPR-Cas system can help microbes fend off viral infections, especially in environments where exposure to phages could limit their survival (Burstein et al., 2016). In contrast, uncultured MAGs were enriched for pathways such as NADH dehydrogenase activity, iron-sulfur cluster biosynthesis, and asparagine metabolism—features potentially associated with ecological dependencies or niche-specific adaptations (Boyd et al., 2014; Mand & Metcalf, 2019; Mee et al., 2014; Norton et al., 2008)

Notably, we identified 15 pathways that were exclusively enriched in cultured MAGs and absent from uncultured ones (**Figure 2.4**). These include biosynthetic pathways for thiamine, folate, pantothenate/CoA, and cobalamin, which are well-

documented microbial growth factors (Kamagata & Tamaki, 2005; Stewart, 2012; Xavier et al., 2017). Our findings support the hypothesis that biosynthetic self-sufficiency is a strong predictor of culturability. For instance, Zheng et al identified uncultured *Burkholderiales* genomes from sludge samples as dependent on vitamin B12 (cobalamin) and their experimental findings revealed that the addition of vitamin B12 stimulated the growth of these taxa (Zheng et al., 2024).

Vitamins are among the most frequently exchanged metabolites in microbial communities (Guillen et al., 2021), and microbes often rely on co-produced vitamins for growth (Davis et al., 2005; Overmann, 2006). For example, *Variovorax* species support thiamine auxotrophs in bioreactor microbial communities while depending on pantothenate supplied by other species (Hessler et al., 2023). Such interdependencies—often mediated by unknown small molecules—are difficult to replicate in vitro (Alain & Querellou, 2009; Hamm et al., 2019). Facilitating small-molecule exchange via co-culture systems or in situ cultivation platforms (e.g., iChip) has significantly improved culturability (Berdy et al., 2017; Imachi et al., 2020; Zengler & Zaramela, 2018). These findings suggest that many uncultured microbes rely on metabolite sharing, limiting their growth in isolation.

To identify genomic features that generalize across phyla, we applied LASSO regression to model culturability. LASSO consistently identified cofactor biosynthesis—particularly thiamine, folate, and pantothenate—as predictive of culturability, even in poorly sampled or rarely cultured phyla such as *Acidobacteriota*, *Planctomycetota*, and *Spirochaetota* (**Supplemental Figure 2.4**). In contrast, NADH dehydrogenase and purine

biosynthesis were retained as predictors of uncultured status across multiple phyla (**Figure 2.5**). Conversely, purine biosynthesis emerged in our analyses as a strong genomic predictor of uncultured status across multiple phyla (**Figure 2.5**). Because purine biosynthesis is metabolically expensive (Chitty & Fraser, 2017; Moffatt & Ashihara, 2002), microbes that invest in it may be outcompeted in nutrient-rich laboratory media where purines are abundant. This suggests that uncultured microbes may favor strategies incompatible with current cultivation protocols Together, these analyses highlight overlapping but distinct predictors of culturability.

Clustering of phyla based on LASSO-derived predictors further revealed meaningful culturability patterns (**Figure 2.7**). Fast-growing phyla that have many cultured members, such as *Bacteroidetes* and *Firmicutes*, clustered tightly and shared a relatively uniform set of culturability-associated pathways. By contrast, slow-growing phyla, including *Spirochaetota*, *Desulfobacterota*, and *Acidobacteriota*, clustered separately and displayed greater pathway heterogeneity. These patterns likely reflect a broader range of growth strategies, ecological adaptations, and unrecognized metabolic dependencies. This also suggests that culturability-linked features are more conserved among fast-growing taxa.

Taken together, our results indicate that culturability is a complex, multifactorial trait shaped by biosynthetic independence, energy metabolism, and ecological interdependence. Cofactor biosynthesis emerges as a consistent feature of culturability, while pathways associated with ecological dependency or high energetic cost (e.g., purine biosynthesis) are more common among uncultured lineages.

These insights challenge the notion that uncultured microbes remain so due to oversight or insufficient effort. Instead, they highlight fundamental differences in metabolic and ecological strategies. Targeted, genome-informed cultivation, especially approaches that supplement key vitamins or simulate inter-species interactions, offers a promising roadmap for accessing the hidden functional potential of uncultured microbes.

In conclusion, our study provides robust evidence that uncultured microbes encode greater predicted functional novelty and contain distinctive genomic predictors of culturability. By identifying generalizable pathways associated with cultured status, we offer a framework for prioritizing taxa and traits for experimental cultivation. Unlocking these microbes will not only expand the known tree of life, but also reveal the vast and largely unexplored landscape of microbial functional diversity.

Methods

Description of data sets

GEM

The GEM (Genomes from Earth's Microbiomes) data comes from the GEM catalog, which consists of 52, 505 metagenome-assembled genomes (MAGs) assembled from a diverse set of environments sampled from around the world including continental soils, oceans, agricultural soils, animal and human hosts. Of these, about 17.4% of the total MAGs were classified as high-quality (≤5% contamination ≥90% completeness, ≥18/20 rRNA and tRNA genes), and the remaining MAGs were medium quality according to the minimum information about a Metagenome Assembled Genome (MIMAG) standard. Hence, each MAG has a completeness level greater than 50% and a

contamination level less than 5% as estimated with CheckM (v1.0.11). The GEM data have been classified taxonomically at higher ranks using GTDB and clustered on 95% ANI to derive the species-level Operational Taxonomic Unit (OTU). The species-level OTUs assignment of these MAGs was cross-validated with that of the GTDB taxonomy (Nayfach et al. 2021). GEM MAGs were annotated for Cluster of Orthologous Genes (COGs) pathways and genes in the anvio environment using the "anvi-run-ncbi-cogs" commands

Isolates

The isolates dataset comprises 7035 genomes sourced from the IMG/JGI website which only includes only those isolates filtered for "cultured Yes.

Cultured Status

To infer the cultured status of each GEM MAG, we first compared taxonomic annotations of each MAG to those RefSeq category (excluding any uncultured genomes such as single amplified genomes or metagenome-assembled genomes) of the same GTDB release (vs 89). MAGs were classified to the lowest taxonomic level at which they have a cultured representative, for instance if a MAG matched a RefSeq genome taxonomically ranked to an order level then the MAG has the corresponding cultured level. b) We next aligned the sequences of each MAG in the GEM dataset to NCBI isolate RefSeq genomes via Reference Seeker (Schwengers et al., 2020) to find if there are any closely related reference genomes. Reference Seeker first conducted a Kmer-based genome distance between the GEM MAGs and RefSeq with Mash (Ondov et al.,

2016) and then calculated the Average Nucleotide Identity (ANI) to determine the MAGs with ANI > 95% and conserved DNA > 69%, a commonly used threshold for delineating the prokaryotic species level. MAGs with species cultured representative or matched an isolate RefSeq with ANI > 95% are termed as “cultured MAGs” while MAGs that do not meet this criterion are referred to as “uncultured MAGs”

Comparing The Genes in Cultured Genomes To Uncultured Genomes

To determine the similarity of sequences from cultured and uncultured genomes to SwissProt, a database whose gene products have been rigorously characterized. We conducted alignments between each gene in the GEM dataset, translated into amino acid sequences, and to the genes in the SwissProt dataset using diamond's "blastp" command (Buchfink et al., 2021).

Diamond’s “blastp” command was also conducted on the cultured isolates group. We set the e-value cutoff to 100 instead of the default e-value because we are interested in any slightest similarity of the GEM sequences to Swissprot. This cutoff ensures many false positive alignments and that even minor similarities between GEM genes and SwissProt genes are identified. We deal with false positives separately by comparing the sequence similarity of random sequences simulated from the GEM database to the SwissProt. The BlastP reports were excluded to only the top hits of the query search. We found that 112,784,914 out of 122,978,191 protein sequences in the GEM MAGs aligned to genes in the SwissProt database, and 27,952,866 out of 30,840,483 protein sequences of the isolates aligned to the SwissProt database sequences. The bit scores were normalized by dividing them by the alignment length. We also

generated a random seq from the GEM MAGs dataset using the same e-value. We further compared the sequences in the GEM MAGs at each taxonomic cultured level to the Agnostos HMM profile database to quantify the genes with known and unknown functions (**Supplemental Figure 2.3**).

Fisher's enrichment test analysis

To identify the functional pathways and genes that could provide insights into potential factors contributing to cultivability, we analyzed the COGs enriched in cultured or uncultured MAGs. For each phylum, we did comparative statistical analysis of COGs presence or absence using Fisher's exact test and enriched COGs in either cultured MAGs or uncultured MAGs based on relative proportions. We specifically focused on COG pathways that are significantly over-represented or under-represented in each group ($P < 0.05$).

Description of the LASSO regression model pipeline for analyzing cultured status

To assess the distinction between cultured and uncultured GEM microbes, we employed LASSO regression to predict if an annotated pathway or gene can be associated with cultured status. We designed a machine learning-based pipeline specifically intended to predict metabolic pathways associated with cultured status while mitigating the effects of phylogeny. To create the custom dataset, we assigned a label of "cultured" to the cultured MAGs or "uncultured" if it belonged to any other taxonomic cultured level. The dataset contained both pathway features, which represent coarse-grained COG assignments (aggregated into higher biochemical pathways), and annotation features, a finer-grained measure of COG assignments. The value of each feature is the

approximate number of genes in a given genome/species with a specific COG assignment. We normalized these pathways by abundance counts to accurately represent the phylum. To partially control for phylogeny, we remove all metadata features from the feature set used in the model, as we have previously shown metadata to be overly predictive of culturability regarding this particular dataset (Babjac et al., 2022). These metadata include the cultured levels, taxonomy classes and taxonomic distances, completeness, contamination, and genome size. To finish controlling for phylogeny, we subset the data by phylum and run the machine learning part of our pipeline separately on each phylum (**Supplemental Figure 2.5**), effectively controlling for taxonomy. The machine learning section of our pipeline consists of the following steps: (i) preprocess the data, (ii) train a LASSO-based bagging regressor to predict cultured or uncultured status, and (iii) compute LASSO coefficient confidence intervals and feature importances. We will describe each of these steps in more detail in the following sections and the code can be found on Github (https://github.com/ababjac/GEM-feature-selection)

Details of necessary machine learning preprocessing

The preprocessing consists of standard machine learning practices: (i) stratify the data into training/testing sets, (ii) scale the data to meet model requirements, and (iii) rebalance by phylum to create a balanced training set. We split the data using the train_test_split function from scikit-learn, with 20% reserved for validation, and 80% utilized for training and stratification performed based on the cultured labels. We scale the data using the MinMax normalization technique with standard parameters. We then rebalance using SMOTE (Chawla et al., 2002) which is an oversampling technique that

creates new samples in a feature space by selecting along the path of two closely related real data points. We additionally note that any phyla that does not have a minimum of ten samples total and/or has less than three samples of either label are removed from the analysis (Supplemental Table 1); this is due to statistical constraints on the number of samples that are required to use the methodologies presented here.

Details of LASSO-based regression model for predicting cultured and uncultured status

We will employ LASSO (Least Absolute Shrinkage and Selection Operator) as the regression component of our analysis pipeline as it has been previously shown to be a robust feature selection tool (Queen & Emrich, 2021)and to be a comparable dimensionality reduction method to neural network-based architectures on these data (Babjac et al., 2022). LASSO is a modified form of regression that uses a cost function to tune the model based on an α hyperparameter (Fonti & Belitser, 2017). The lower the α parameter, the more LASSO begins to approximate standard regression, whereas higher α values tend to cause coefficients to more rapidly trend toward zero. When using LASSO, features are selected based on the coefficient, i.e., if the coefficient is non-zero despite the stringent α penalty, then we say it contributes to the model.

We incorporate LASSO as the underlying model in a bagged regressor -- an ensemble model that is the composite of fitting a base model (LASSO) on many random subsets of the data. We build our bagging regressor using scikit-learn's Bagging Regressor with estimators set to fifty and bootstrapping set to True. Each estimator in the bagged regressor is an instance of a logistic regression (scikit-learn's Logistic

Regression) model where the penalty is set to "l1" representing a LASSO penalty. We train a separately bagged regressor for each phylum using 5-fold cross-validation within the 80% training partition.

Details regarding calculating and interpreting coefficients, confidence intervals and permutation importance

A bagged regressor is advantageous because it mitigates the effect of randomness as LASSO (like many other machine learning methods) will return slightly different features based on the random seeding. This not only makes our predictions more verbose but also allows for calculating t-confidence intervals for each coefficient. For each feature, we calculate the mean and standard deviation of its associated coefficients from the underlying LASSO models in the trained bagged regressor. We then compute standard confidence intervals and p-values using the t-distribution. Confidence intervals and p-values are directly linked in that smaller p-values will have larger confidence intervals because we are more confident, that we contain the true value. Since we are looking for non-zero coefficients based on LASSO, we can select a feature if its confidence interval does not contain zero. We tested two cutoffs for selecting significant coefficients based on the p-values ($p < 0.05$ and $p < 0.1$) and found that $p < 0.05$ had a stronger signal. If a coefficient is significant, we then interpret it as follows: "For each (normalized) 1 unit increase of a sample's feature_X, we would expect an average coefficient_X increase/decrease in the probability that sample is cultured", where increase/decrease comes from the value of the coefficient.

A common misconception when interpreting coefficients of regression models is that larger coefficients are more “important” to model understanding. To determine the most important coefficients we additionally perform permutation importance testing (Fisher et al., 2019) using the 20% validation data which was held out from the original modeling. The permutation importance algorithm is an explainable AI method that calculates how much each feature contributes to a trained model by permuting over the fitted estimator (our LASSO bagged regressor in this case) and randomly shuffling one feature each time. When this feature is removed, we can get a measure of what the model is “missing” by randomizing it. Higher permutation importance scores indicate an increasingly positive influence on the model, while scores below zero suggest the feature may have negatively affected model performance. We then put this feature back into the model and remove the next, iterating through all features to get a full set of importances. By doing this on the validation data, we get a measure of each feature's generalizability to new data. We used Scikit-Learn’s permutation importance algorithm for these calculations (Pedregosa et al., 2011). This algorithm was chosen because it should be more robust to feature inflation than impurity-based feature selection methods (e.g. Gini).

Shuffling data to create a random baseline

As an additional sanity check, we aim to ensure the performance and coefficients from our analysis cannot occur by random chance. A standard technique for validating experimental results is to perform shuffling of the input data before repeating the analysis and comparing the results to those of the true data. We create a shuffled dataset using Scikit Learn’s shuffle algorithm with default parameters. We omit both the cultured

labels and metadata columns from being shuffled to preserve the integrity of the phyla and predictions when modeling. We permute five different partitions of shuffled data and average the results to produce our random baseline set. We replicate all steps of the analysis using the shuffled data as described previously.

Evaluating model performance

In order to evaluate our models, we use standard classification metrics: AUROC, accuracy, precision and recall. AUROC represents the area under the receiver operator curve which is a number between 0 and 1 showing a true vs false positive rate across all classification thresholds (Huang & Ling, 2005). For balanced data, we would expect a minimum AUROC of 0.5 if we predicted everything as true for a baseline comparison. Accuracy measures the proportion of correct predictions with respect to the total number of predictions (i.e., Accuracy = TP + TN / Total). Precision accounts for the proportion of correct positive identifications (i.e., Precision = TP / TP + FP). Finally, Recall identifies the proportion of correctly classified positive labels (i.e., Recall = TP / TP + FN). We compute each metric using the 20% validation set per phylum. We determine a minimum threshold for "good" performance based on the shuffled data, i.e., if the real performance is statistically greater than the shuffled performance, we say that the model is performing better than we would expect by random chance (**Supplemental Figure 2.6 through 2.9**). We also note that this dataset (GEM) was previously baselined against several other methods (Babjac et al., 2022) but not by individual phylum.

References

Alain, K., & Querellou, J. (2009). Cultivating the uncultured: Limits, advances and future challenges. *Extremophiles*, *13*(4), 583–594. https://doi.org/10.1007/s00792-009-0261-3

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., & Hugenholtz, P. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, *39*(1), 105–114.

Babjac, A., Royalty, T., Steen, A. D., & Emrich, S. J. (2022). A comparison of dimensionality reduction methods for large biological data. *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 1–7. https://doi.org/10.1145/3535508.3545536

Berdy, B., Spoering, A. L., Ling, L. L., & Epstein, S. S. (2017). In situ cultivation of previously uncultivable microorganisms using the ichip. *Nature Protocols*, *12*(10), 2232–2242.

Bernard, G., Pathmanathan, J. S., Lannes, R., Lopez, P., & Bapteste, E. (2018). Microbial dark matter investigations: How microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery. *Genome Biology and Evolution*, *10*(3), 707–715.

Berney, M., & Cook, G. M. (2010). Unique flexibility in energy metabolism allows mycobacteria to combat starvation and hypoxia. *PloS One*, *5*(1), e8614.

Boyd, E. S., Thomas, K. M., Dai, Y., Boyd, J. M., & Outten, F. W. (2014). Interplay between Oxygen and Fe–S Cluster Biogenesis: Insights from the Suf Pathway. *Biochemistry*, *53*(37), 5834–5847. https://doi.org/10.1021/bi500488r

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), 366–368.

Burstein, D., Sun, C. L., Brown, C. T., Sharon, I., Anantharaman, K., Probst, A. J., Thomas, B. C., & Banfield, J. F. (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nature Communications*, *7*(1), 10613.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chitty, J. L., & Fraser, J. A. (2017). Purine acquisition and synthesis by human fungal pathogens. *Microorganisms*, *5*(2), 33.

Chung, S. Y., & Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology. *Structure*, *4*(10), 1123–1127.

Davis, K. E. R., Joseph, S. J., & Janssen, P. H. (2005). Effects of Growth Medium, Inoculum Size, and Incubation Time on Culturability and Isolation of Soil Bacteria. *Applied and Environmental Microbiology*, *71*(2), 826–834. https://doi.org/10.1128/AEM.71.2.826-834.2005

D'Onofrio, A., Crawford, J. M., Stewart, E. J., Witt, K., Gavrish, E., Epstein, S., Clardy, J., & Lewis, K. (2010). Siderophores from neighboring organisms promote the growth of uncultured bacteria. *Chemistry & Biology*, *17*(3), 254–264.

Dunbrack Jr, R. L. (2006). Sequence comparison and protein structure prediction. *Current Opinion in Structural Biology*, *16*(3), 374–384.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

Fonti, V., & Belitser, E. (2017). Feature selection using lasso. *VU Amsterdam Research Paper in Business Analytics*, *30*, 1–25.

Garza, D. R., & Dutilh, B. E. (2015). From cultured to uncultured genome sequences: Metagenomics and modeling microbial ecosystems. *Cellular and Molecular Life Sciences*, *72*(22), 4287–4308. https://doi.org/10.1007/s00018-015-2004-1

Ghatak, S., King, Z. A., Sastry, A., & Palsson, B. O. (2019). The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. *Nucleic Acids Research*, *47*(5), 2446–2454.

Ghoul, M., & Mitri, S. (2016). The ecology and evolution of microbial competition. *Trends in Microbiology*, *24*(10), 833–845.

Godzik, A. (2011). Metagenomics and the protein universe. *Current Opinion in Structural Biology*, *21*(3), 398–403.

Guillen, M. N., Rosener, B., Sayin, S., & Mitchell, A. (2021). Assembling stable syntrophic Escherichia coli communities by comprehensively identifying beneficiaries of secreted goods. *Cell Systems*, *12*(11), 1064–1078.

Hamm, J. N., Erdmann, S., Eloe-Fadrosh, E. A., Angeloni, A., Zhong, L., Brownlee, C., Williams, T. J., Barton, K., Carswell, S., Smith, M. A., Brazendale, S., Hancock, A. M., Allen, M. A., Raftery, M. J., & Cavicchioli, R. (2019). Unexpected host dependency of Antarctic Nanohaloarchaeota. *Proceedings of the National Academy of Sciences*, *116*(29), 14661–14670. https://doi.org/10.1073/pnas.1905179116

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, *68*(4), 669–685. https://doi.org/10.1128/MMBR.68.4.669-685.2004

Hessler, T., Huddy, R. J., Sachdeva, R., Lei, S., Harrison, S. T., Diamond, S., & Banfield, J. F. (2023). Vitamin interdependencies predicted by metagenomics-informed network analyses and validated in microbial community microcosms. *Nature Communications*, *14*(1), 4768.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299–310.

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., Takano, Y., Uematsu, K., Ikuta, T., & Ito, M. (2020). Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature*, *577*(7791), 519–525.

Kamagata, Y., & Tamaki, H. (2005). Cultivation of uncultured fastidious microbes. *Microbes and Environments*, *20*(2), 85–91.

Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z., & Ettema, T. J. (2021). Innovations to culturing the uncultured microbial majority. *Nature Reviews Microbiology*, *19*(4), 225–240.

Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J., & Crosby, L. (2018). Phylogenetically novel uncultured microbial cells dominate earth microbiomes. *MSystems*, *3*(5), e00055-18.

Mand, T. D., & Metcalf, W. W. (2019). Energy Conservation and Hydrogenase Function in Methanogenic Archaea, in Particular the Genus*Methanosarcina*. *Microbiology and Molecular Biology Reviews*, *83*(4). https://doi.org/10.1128/mmbr.00020-19

Mee, M. T., Collins, J. J., Church, G. M., & Wang, H. H. (2014). Syntrophic exchange in synthetic microbial communities. *Proceedings of the National Academy of Sciences*, *111*(20). https://doi.org/10.1073/pnas.1405641111

Moffatt, B. A., & Ashihara, H. (2002). Purine and pyrimidine nucleotide synthesis and metabolism. *The Arabidopsis Book/American Society of Plant Biologists*, *1*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243375/

Mu, D.-S., Ouyang, Y., Chen, G.-J., & Du, Z.-J. (2021). Strategies for culturing active/dormant marine microbes. *Marine Life Science & Technology*, *3*(2), 121–131. https://doi.org/10.1007/s42995-020-00053-z

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., & Huntemann, M. (2021). A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, *39*(4), 499–509.

Norton, J. M., Klotz, M. G., Stein, L. Y., Arp, D. J., Bottomley, P. J., Chain, P. S. G., Hauser, L. J., Land, M. L., Larimer, F. W., Shin, M. W., & Starkenburg, S. R. (2008). Complete Genome Sequence of *Nitrosospira multiformis* , an Ammonia-Oxidizing Bacterium from the Soil Environment. *Applied and Environmental Microbiology*, *74*(11), 3559–3572. https://doi.org/10.1128/aem.02722-07

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 132. https://doi.org/10.1186/s13059-016-0997-x

Overmann, J. (2006). Principles of enrichment, isolation, cultivation and preservation of prokaryotes. *Prokaryotes*, *1*, 80–136.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Prakash, O., Parmar, M., Vaijanapurkar, M., Rale, V., & Shouche, Y. S. (2021). Recent trend, biases and limitations of cultivation-based diversity studies of microbes. *FEMS Microbiology Letters*, *368*(17), fnab118.

Prakash, O., Shouche, Y., Jangid, K., & Kostka, J. E. (2013). Microbial cultivation and the role of microbial resource centers in the omics era. *Applied Microbiology and Biotechnology*, *97*, 51–62.

Queen, O., & Emrich, S. J. (2021). LASSO-based feature selection for improved microbial and microbiome classification. *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2301–2308. https://ieeexplore.ieee.org/abstract/document/9669485/

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., & Gies, E. A. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–437.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, *12*(2), 85–94.

Salam, N., Xian, W.-D., Asem, M. D., Xiao, M., & Li, W.-J. (2021). From ecophysiology to cultivation methodology: Filling the knowledge gap between uncultured and cultured microbes. *Marine Life Science & Technology*, *3*(2), 132–147. https://doi.org/10.1007/s42995-020-00064-w

Schwengers, O., Hain, T., Chakraborty, T., & Goesmann, A. (2019). ReferenceSeeker: Rapid determination of appropriate reference genomes. *BioRxiv*, 863621.

Solden, L., Lloyd, K., & Wrighton, K. (2016). The bright side of microbial dark matter: Lessons learned from the uncultivated majority. *Current Opinion in Microbiology*, *31*, 217–226.

Steen, A. D., Crits-Christoph, A., Carini, P., DeAngelis, K. M., Fierer, N., Lloyd, K. G., & Cameron Thrash, J. (2019). High proportions of bacteria and archaea across most biomes remain uncultured. *The ISME Journal*, *13*(12), 3126–3130.

Stewart, E. J. (2012). Growing Unculturable Bacteria. *Journal of Bacteriology*, *194*(16), 4151–4160. https://doi.org/10.1128/JB.00345-12

Szenk, M., Dill, K. A., & de Graff, A. M. (2017). Why do fast-growing bacteria enter overflow metabolism? Testing the membrane real estate hypothesis. *Cell Systems*, *5*(2), 95–104.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

Vartoukian, S. R., Palmer, R. M., & Wade, W. G. (2010). Strategies for culture of ‘unculturable’bacteria. *FEMS Microbiology Letters*, *309*(1), 1–7.

Winkler, A. M., Webster, M. A., Vidaurre, D., Nichols, T. E., & Smith, S. M. (2015). Multi-level block permutation. *Neuroimage*, *123*, 253–268.

Xavier, J. C., Patil, K. R., & Rocha, I. (2017). Integration of biomass formulations of genome-scale metabolic models with experimental data reveals universally essential cofactors in prokaryotes. *Metabolic Engineering*, *39*, 200–208.

Zengler, K., & Zaramela, L. S. (2018). The social network of microorganisms—How auxotrophies shape complex communities. *Nature Reviews Microbiology*, *16*(6), 383–390.

Zheng, M., Wen, L., He, C., Chen, X., Si, L., Li, H., Liang, Y., Zheng, W., & Guo, F. (2024). Sequencing-guided re-estimation and promotion of cultivability for environmental bacteria. *Nature Communications*, *15*(1), 9051.

Appendix II

Figures

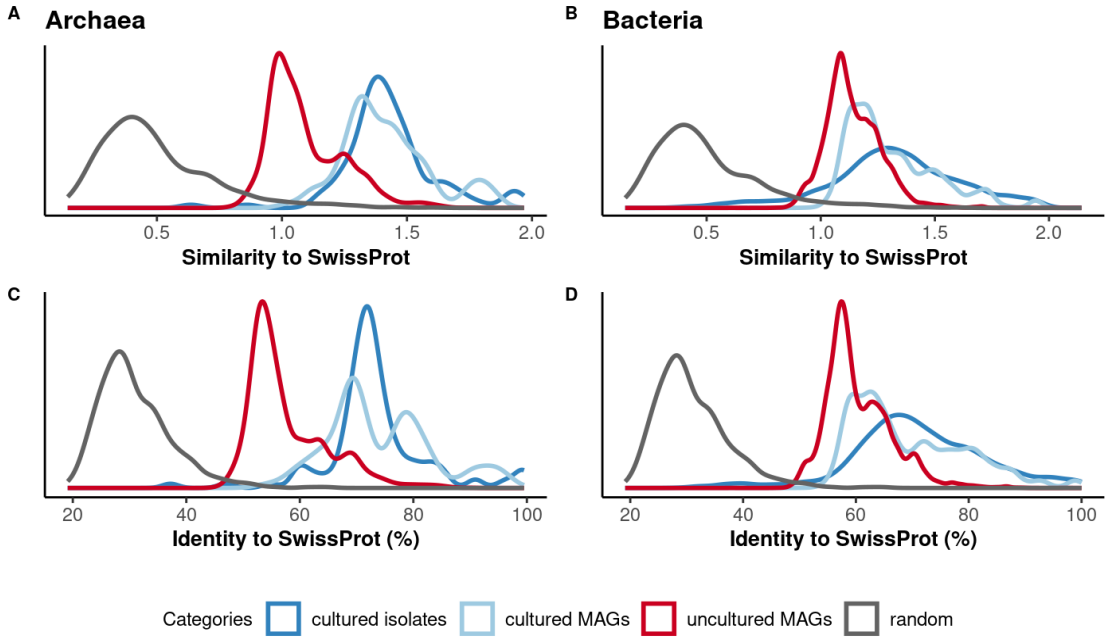


Figure 2.1: Similarity of predicted protein sequences in MAGs and isolates to SwissProt entries. A and C show normalized bitscores and percent identities, respectively, for archaeal genomes; B and D show the same metrics for bacterial genomes. The distributions represent the average similarity of all predicted protein-coding genes to their closest match in the SwissProt. Random sequences (gray) represent simulated protein sequences with no known homology and serve as a negative control.

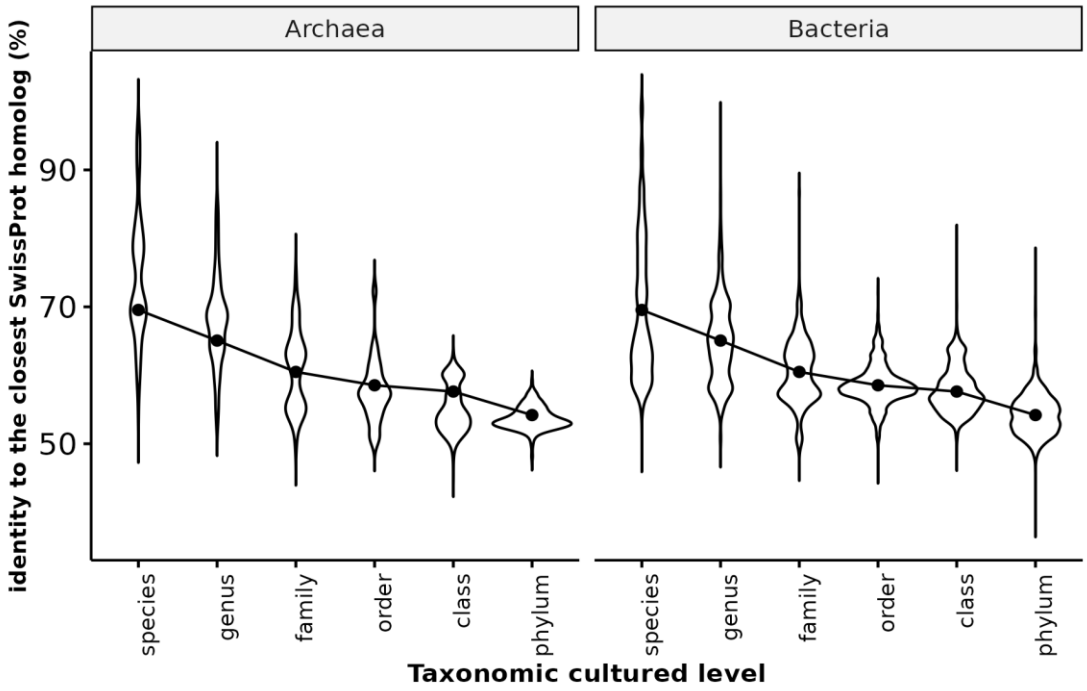


Figure 2.2: Similarity of predicted proteins in Archaea (left) and Bacteria (right) to their closest SwissProt homolog, stratified by taxonomic proximity to cultured isolates. For each metagenome-assembled genome (MAG) in the GEM catalog, the lowest shared taxonomic rank with a cultured isolate was determined. Protein similarity—measured as percent identity to the nearest SwissProt match—is shown as a function of this taxonomic proximity. Taxonomic rank assignments for GEM MAGs are based on the Genome Taxonomy Database (GTDB), as implemented by Nayfach et al. Cultured lineages were identified based on RefSeq annotations in the same GTDB release, excluding genomes

derived from single-cell amplification or other metagenome-assembled genomes (MAGs or SAGs).

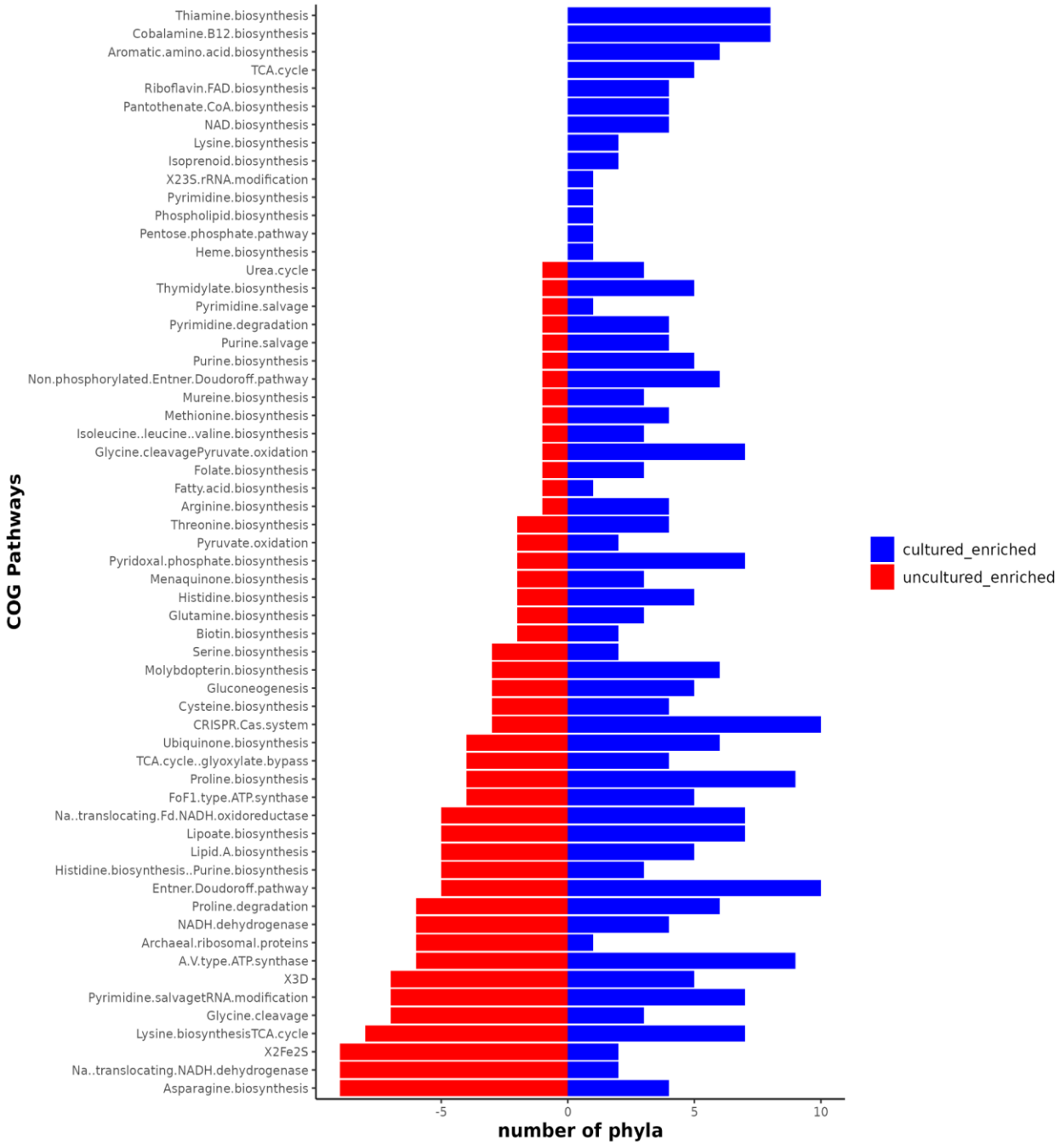


Figure 2.3: COG pathways that are enriched in cultured and uncultured MAGs (P < 0.005, fisher's exact test). Pathways related to co-factors, vitamins, energy production, and CRISPR Cas are over-represented in cultured MAGs.

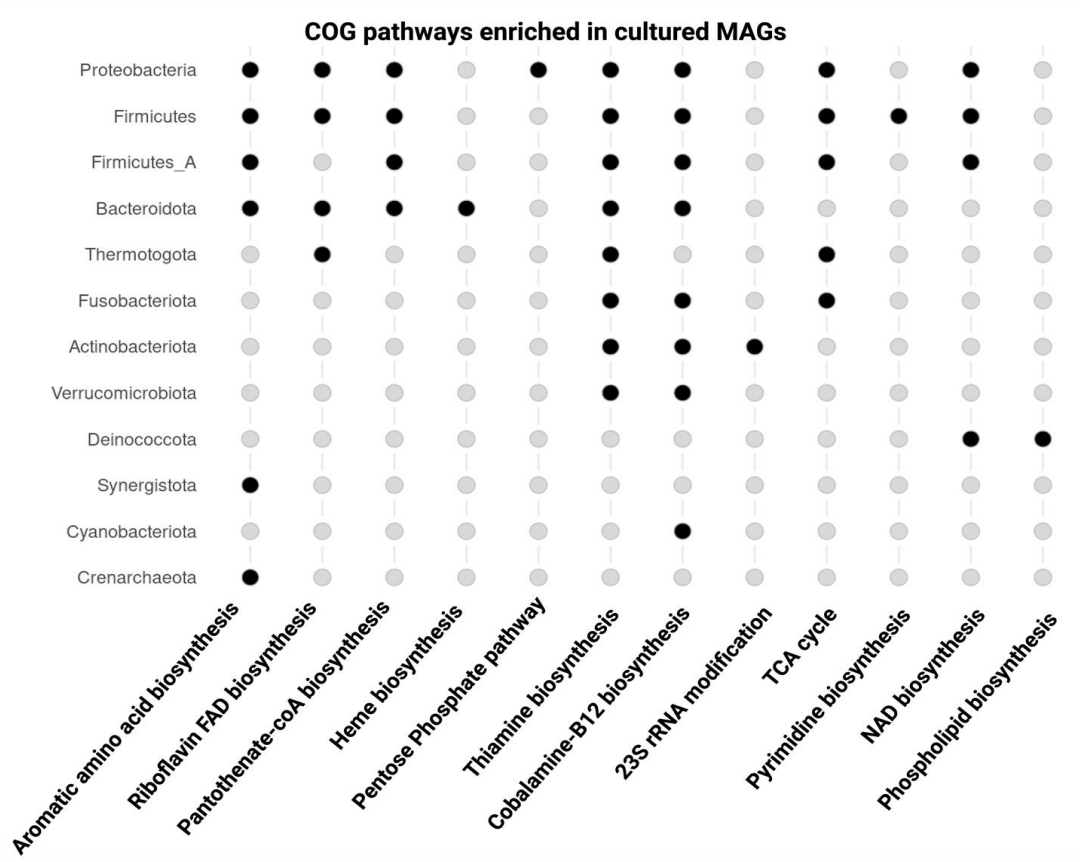


Figure 2.4: Clusters of Orthologous Groups (COG) pathways that are significantly enriched in cultured MAGs in at least one phylum and not enriched in uncultured MAGs in any phylum. Each column in the plot represents a COG pathway, and each row represents a phylum. Solid circles indicate phyla where the corresponding pathway is significantly enriched in cultured MAGs (Fisher's exact test, FDR-adjusted $P < 0.05$).

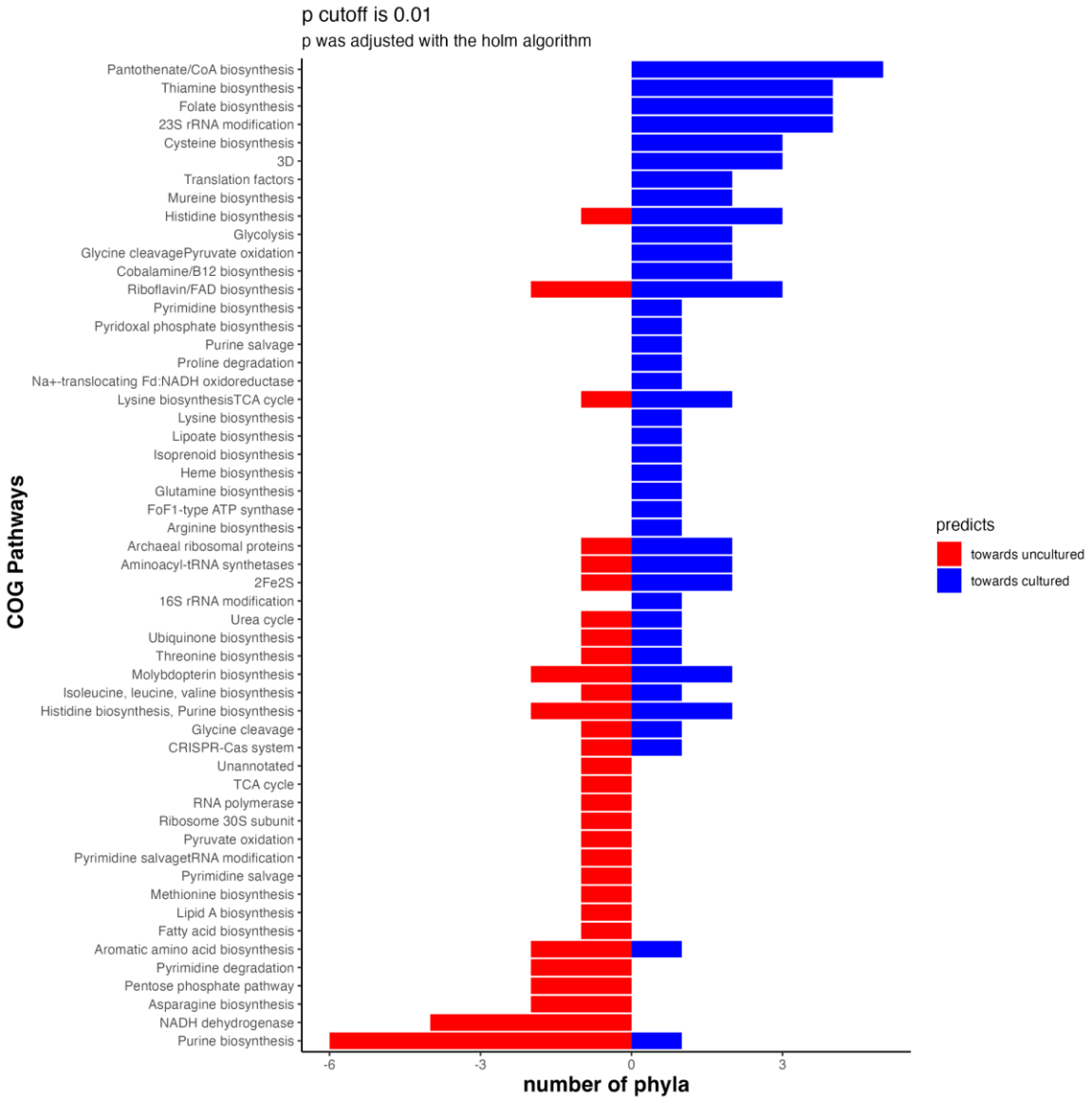


Figure 2.5: Machine learning (LASSO) predicts important COG pathways that are predictive of microbial cultured or uncultured status. Pathways are filtered based on the

permutation importance. Only pathways with permutation importance (predictive power>0.1) are included in the plot.

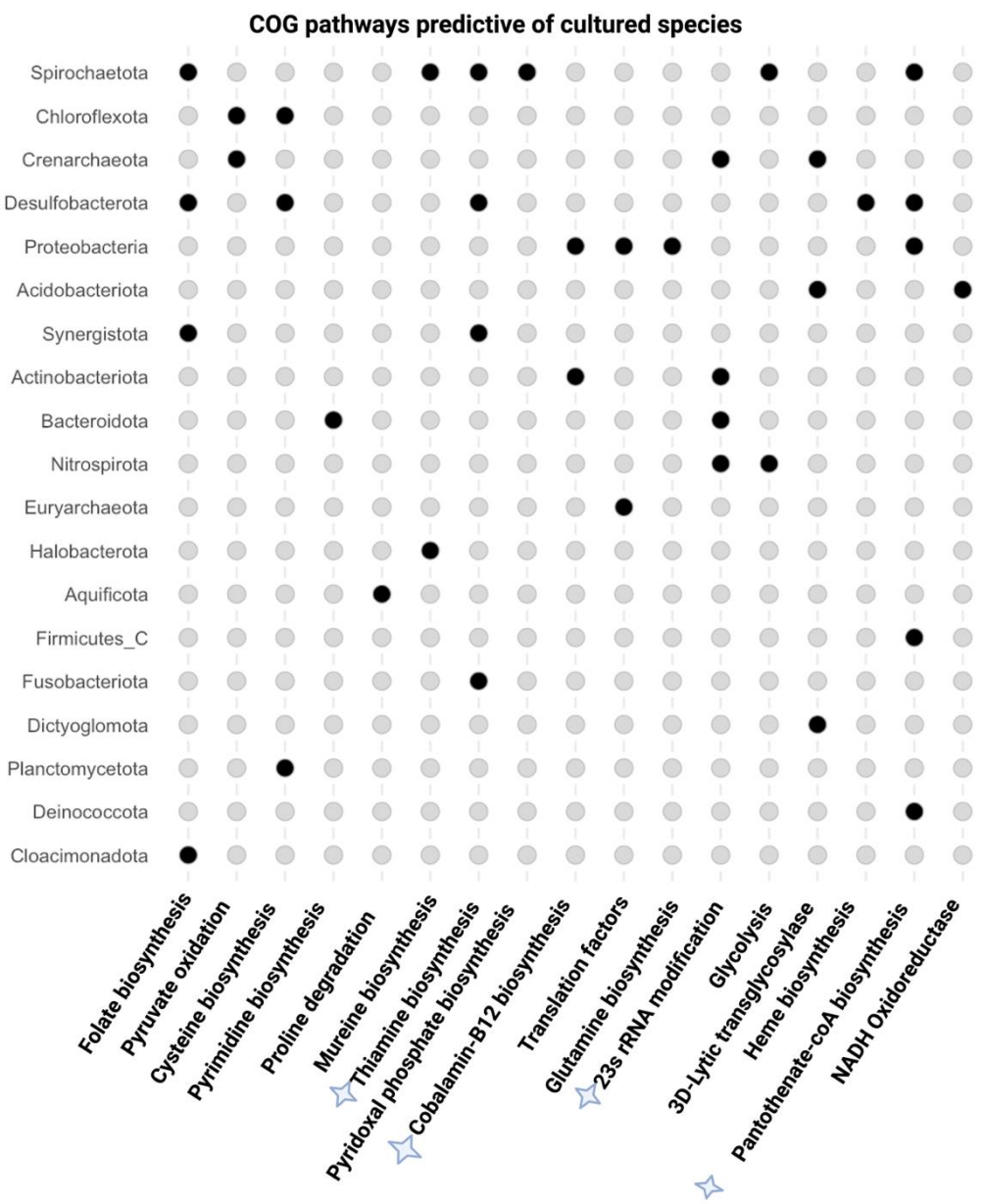


Figure 2.6: Machine learning (LASSO) predicts COG pathways that are strongly predictive of cultured status by phylum. Only COG pathways that predict cultured status with high permutation importance (>0.1) are included in the plot. Solid circles indicate the phyla for which this COG pathway is significantly predicts cultured status. COG

pathways that are both enriched in cultured MAGs and predict microbial cultured status are starred and are predominantly vitamin biosynthetic pathways.

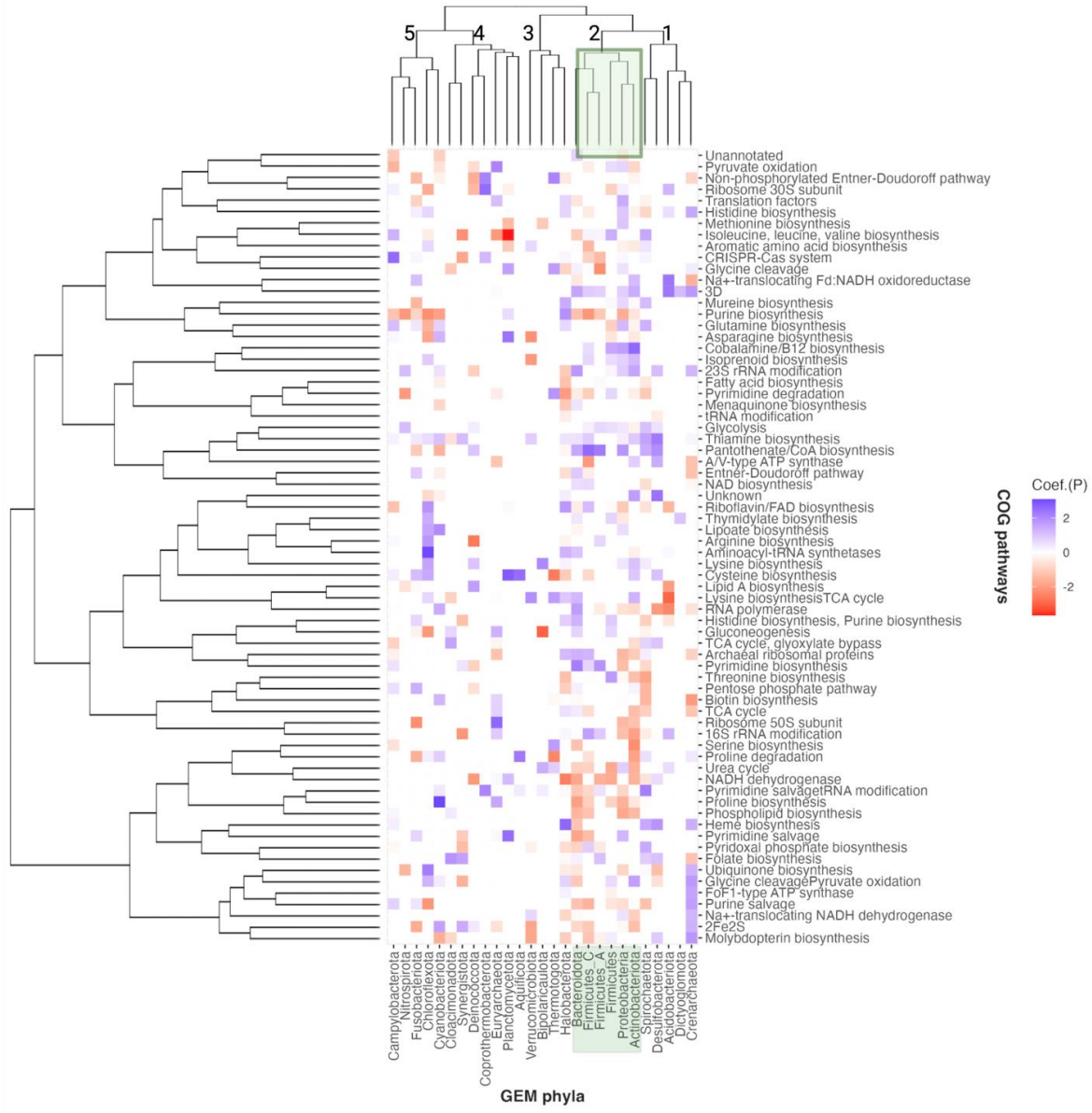


Figure 2.7: Heat map showing COG pathways predicting cultured (blue) and uncultured (red) species within each phylum. The intensity of the color represents the strength of the prediction (see methods). Various clusters are numbered and phyla in which fast growers are known to be abundant are highlighted in green. Microbes that are known to be fast

growers and have higher cultured representatives cluster together while slow growers show greater variability in clustering.

Supplemental Results

Classifying GEM data set into taxonomic cultured levels

Although in theory any species can be assigned a taxonomy to the species level, in practice, microbes with no close relation in the Genome Taxonomy Database (GTDB) will not be assigned to lower specific taxonomic ranks (species, genus…). Thus, the level of assigned taxonomic rank is a proxy for the frequency of an organism's close relatives in GTDB. A larger fraction of bacterial GEM MAGs were assigned to lower taxonomic ranks and more frequently to taxa with cultured representatives than Archaea, which suggests that bacteria tend to be more cultured (**Supplemental Figure 2.2**). For instance, 52.4% of bacteria GEM MAGs were assigned to a species vs. 43.3% for archaeal MAGs. 22.2% of bacterial MAGs were classified as a species with a cultured representative, whereas only 10.5% of archaeal MAGs were classified as the same species-cultured group. At the other end of the taxonomic rank spectrum, 47.6% of bacterial GEM MAGs were taxonomically ranked only to the genus level while for the archaeal MAGs, 56.7% were taxonomically ranked only to the genus level. (**Supplemental Figure 2.2B**). For the taxonomic cultured groups, 73.7% of Bacterial MAGs were taxonomically classified up to the same genus level as a cultured organism, whereas 78.7% of archaeal MAGs were classified as being cultured (**Supplemental Figure 2.2A**.).

4.1% of the bacterial MAGs were not taxonomically classified to any cultured organism, which is a lesser percentage than the 10.8% proportion of the archaeal MAGs. Since GTDB includes sequences from both cultured and uncultured organisms for taxonomic assignment, this indicates that a reasonable percentage of the microbial taxonomic information was derived from uncultured groups.

Machine learning model to predict microbial cultured vs. uncultured status differences using COGs pathways and genes

To confirm our LASSO performance is better than what would occur by random chance, we repeated our analysis on shuffled data, i.e., data for which the null hypothesis of no genes or gene clusters being associated with culturability was known to be true (see Methods). **Supplemental Figure 2.6** through 2.9 show the performance comparison for evaluating the real data (orange) and shuffled data (blue), where the real data outperforms the shuffled data on average across all the performance metrics considered. For both annotations and pathways, Accuracy and AUROC are both greater than 0.8 for most phyla, whereas shuffled performance sits around 0.4-0.7 and 0.4-0.6 respectively (**Supplemental Figure 2.6 and 2.7**). The Recall is similarly greater than 0.8, and the Precision is greater than 0.7 for nearly all phyla (**Supplemental Figure 2.8 and 2.9**). The difference is more striking with Precision and Recall as most shuffled phyla are well below a score of 0.5 for these metrics which suggests a high level of misprediction in our random baseline that does not occur in our real models. There are only a few phyla where the shuffled performance was within a margin of error of the real performance based on Accuracy (Acidobacteriodata, Coprothermobacterota, and Planctomycetata). After further

inspection, these phyla are highly unbalanced with few cultured/uncultured representatives (Supplemental Table 1) which explains the similar Accuracy for both real and shuffled data. When looking at other metrics for these phyla that better handle imbalance (AUROC, precision, recall), we see that these phyla do outperform random chance in terms of false positives/negatives (**Supplemental Figure 2.6 to 2.9**). We conclude based on these results that the performance of all phyla could not occur by random chance. This confirms that our LASSO models are capturing biologically important signals from both the pathway and annotation features that can be interpreted in the next sections.

Supplemental Tables

Supplemental table 1.1: Counts of cultured and uncultured MAGs with average gene identity < 70% to SwissProt

	Fraction with average gene identity < 70% identity
Cultured MAGs (14620)	8520
Uncultured MAGs (37847)	35112
Cultured Isolates (7035)	3439

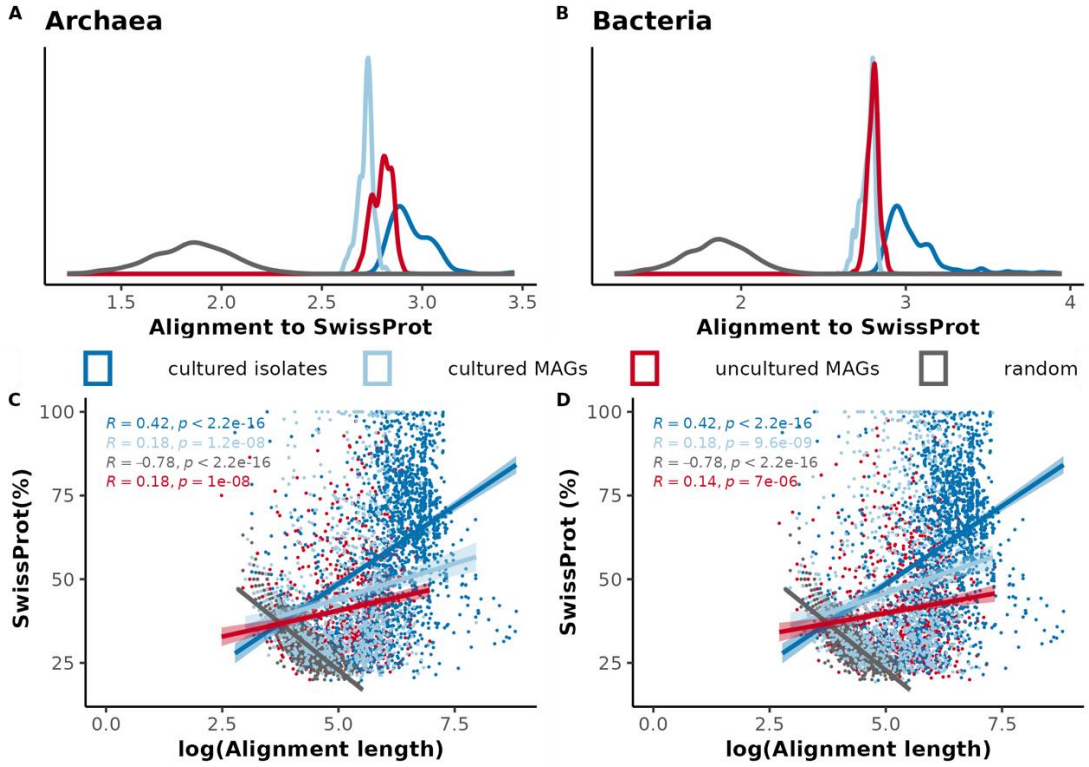
Supplemental Table 2.2: Counts of cultured and uncultured samples per phylum based on the taxonomy of cultured species used in LASSO regression analysis. Phyla are excluded from analysis (and therefore also from this table) if either category contains less than 5 samples (see Methods).

Phylum	**Cultured Counts**	**Uncultured Counts**
Fusobacteriota	81	134
Planctomycetota	8	894
Deinococcota	145	55
Firmicutes_A	1962	6853
Cyanobacteriota	10	478
Firmicutes	658	1301
Firmicutes_C	599	654
Coprothermobacterota	10	6
Chloroflexota	40	1112
Verrucomicrobiota	208	1202
Thermotogota	25	176

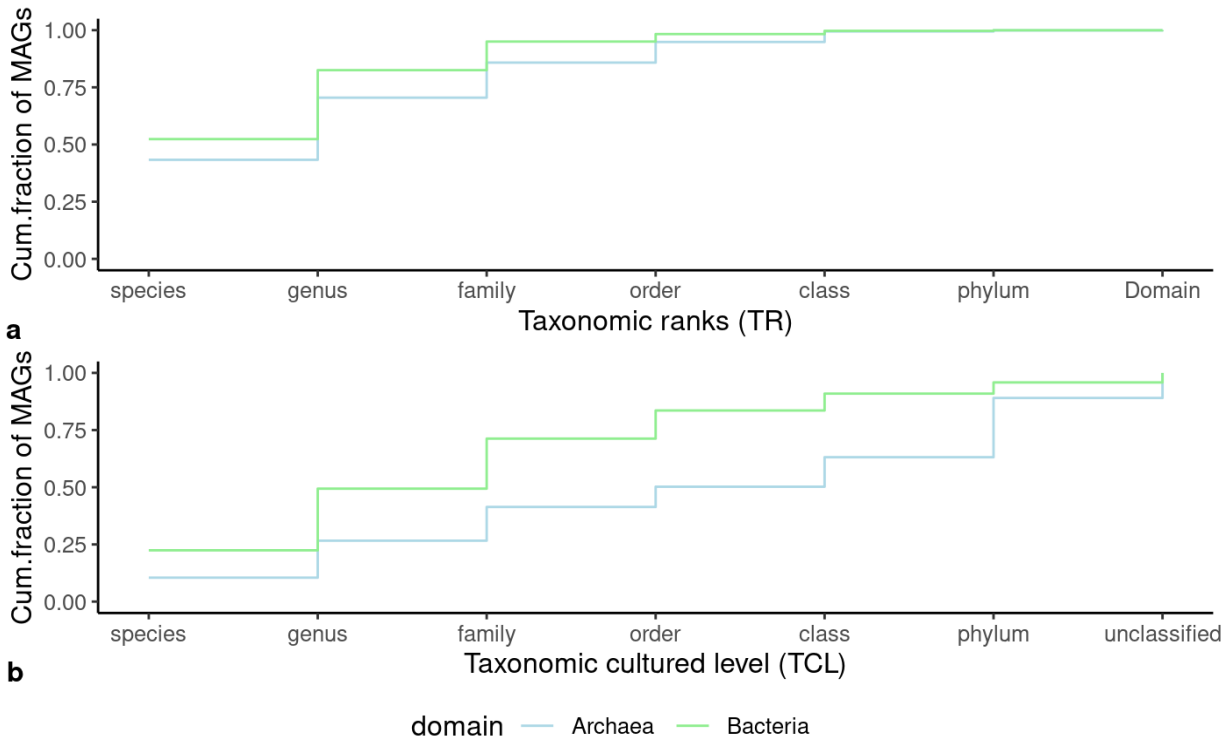
Supplemental Table 2.2 Continued

Acidobacteriota	5	739
Proteobacteria	2190	8460
Cloacimonadota	11	107
Campylobacterota	85	279
Bacteroidota	2776	6266
Actinobacteriota	841	3210
Desulfobacterota	72	910
Aquificota	9	44
Synergistota	18	135
Nitrospirota	16	153

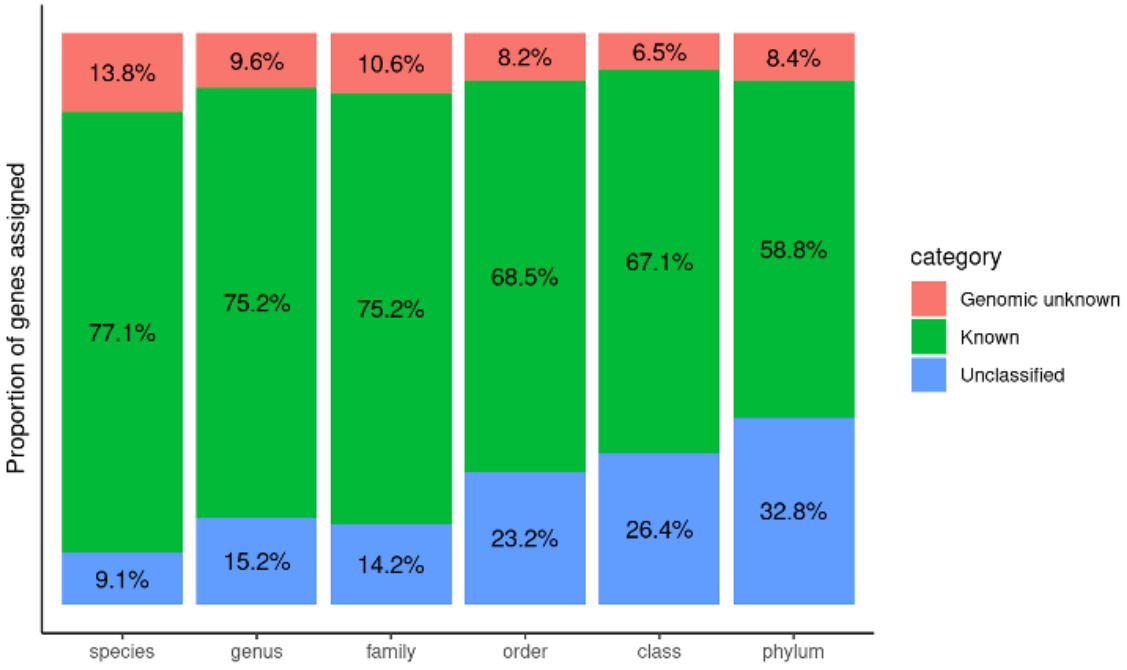
Supplemental figures



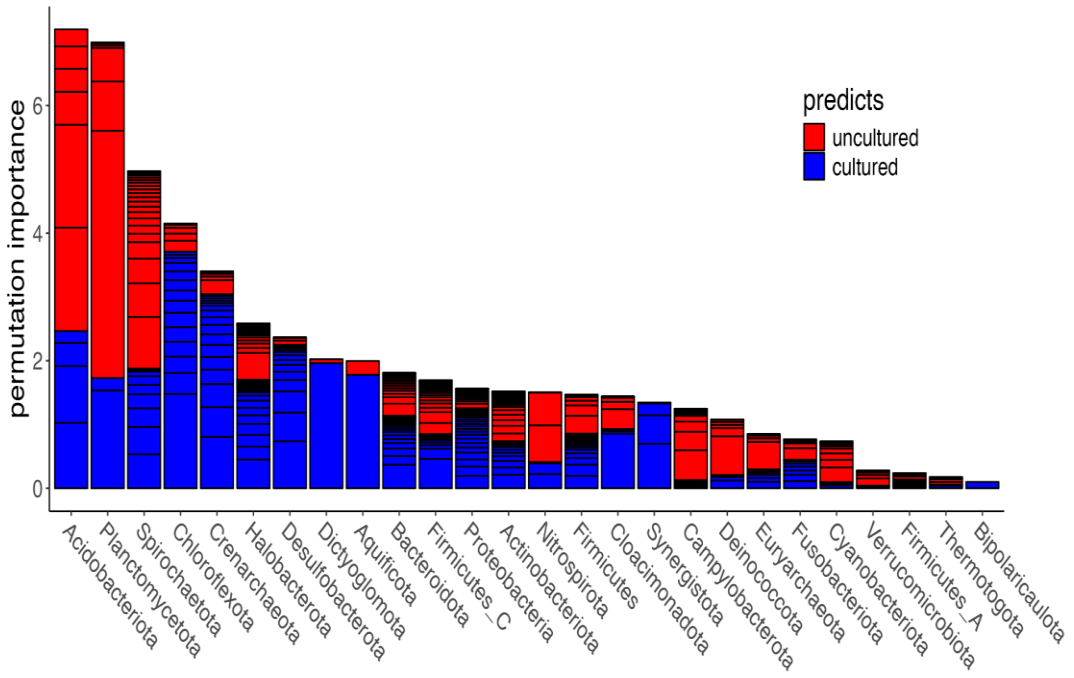
Supplemental Figure 2.1: Distribution of mean of alignment length of MAGs, isolates, and random sequences to SwissProt in A) Archaea and in B) Bacteria. C and D represent the relationship between alignment length and identity to SwissProt of genes (1000 samples) from cultured isolates, uncultured MAGs, cultured MAGs Archaea and Bacteria respectively.



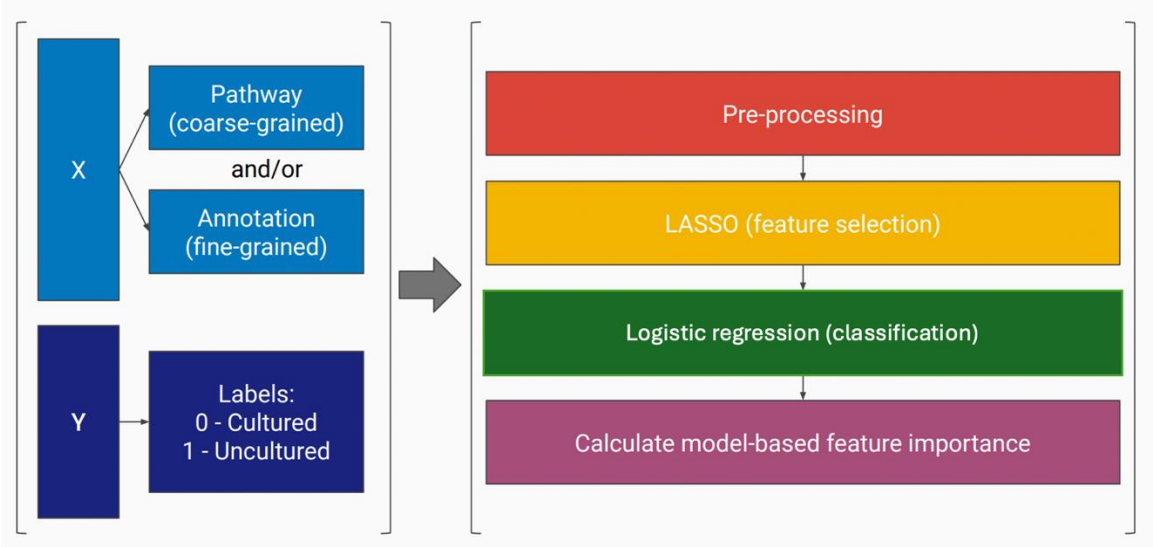
Supplemental Figure 2.2: Cumulative fractions of taxonomic cultured representatives in GEM data sets. The GEM MAG lowest taxonomic rank was assigned by Nayfach et al. (2021) using all Genome Taxonomy Database (GTDB) reference genomes (Chaumeil et al., 2020) while the taxonomic cultured levels were assigned based on the RefSeq category of the same GTDB release excluding single cell amplified genomes and the assembled metagenomes.



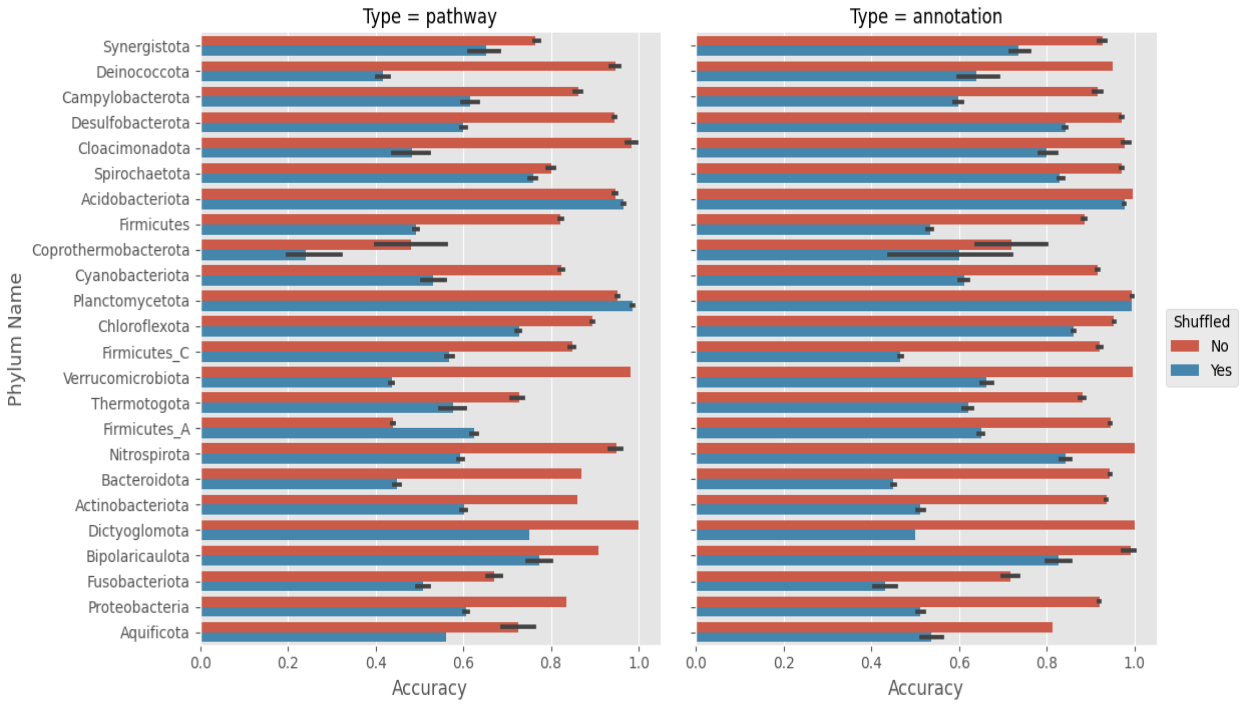
Supplemental Figure 2.3: Proportion of genes assigned by the Agnostos DB clusters profile search database at different GEM taxonomic cultured levels.



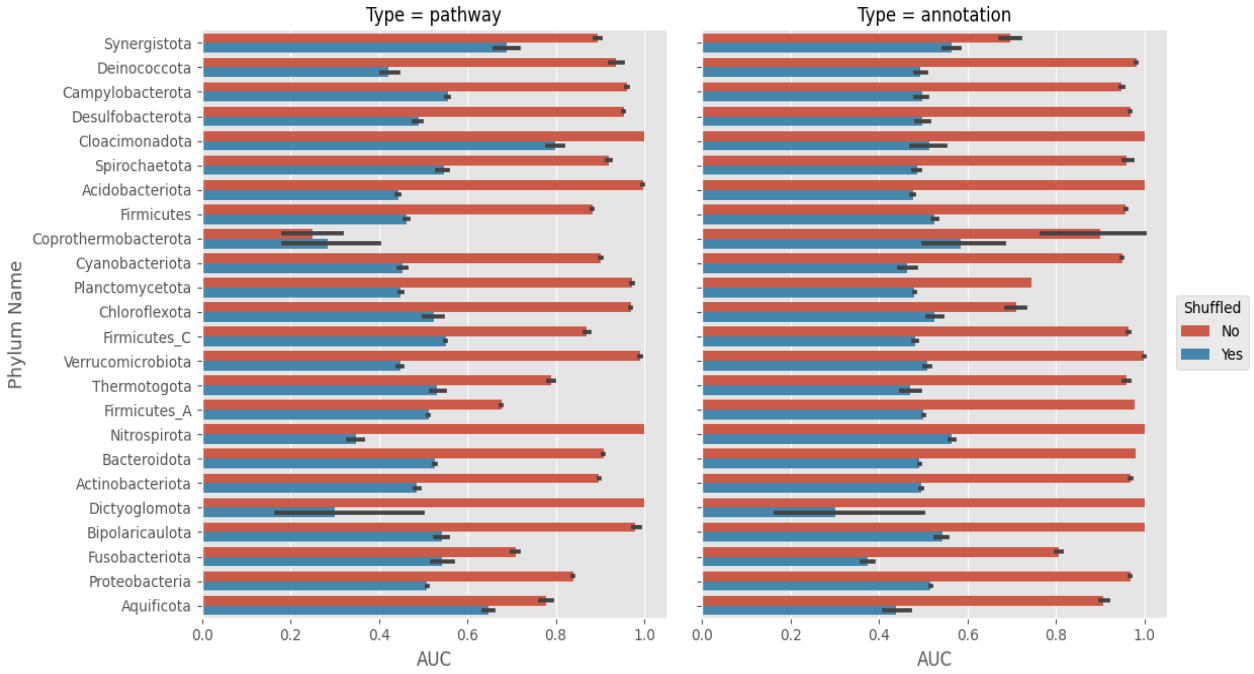
Supplemental Figure 2.4: Permutation importance by GEM phyla. Pathways from less cultured phyla tend to have COG pathways with higher permutation importance.



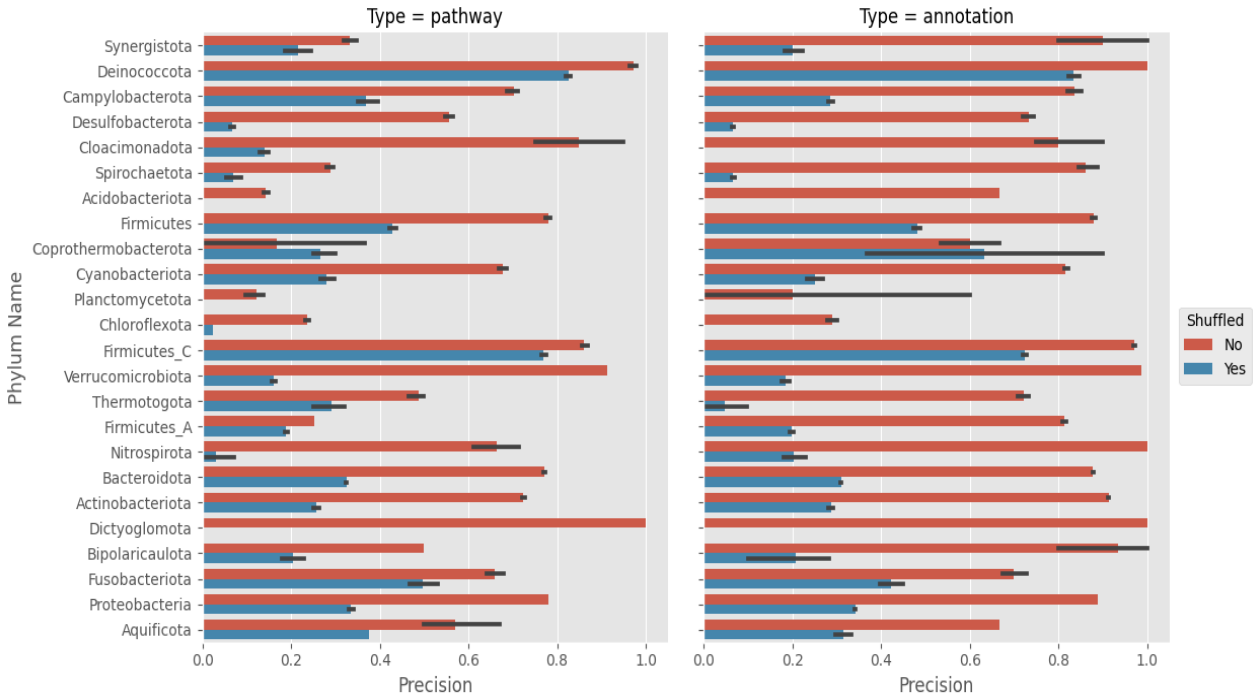
Supplemental Figure 2.3: Machine learning workflow for LASSO regression analysis. We created a custom dataset containing metadata, pathway features, annotation features, and cultured labels (See Methods). Each of these data were then split into subsets representing their individual phyla and the metadata were removed from the subsets. We removed any subsets that did not meet the minimum data requirements (see Methods). We lastly ran our LASSO pipeline 5 times on each dataset and agglomerate the counts of significant phyla into a final output file.



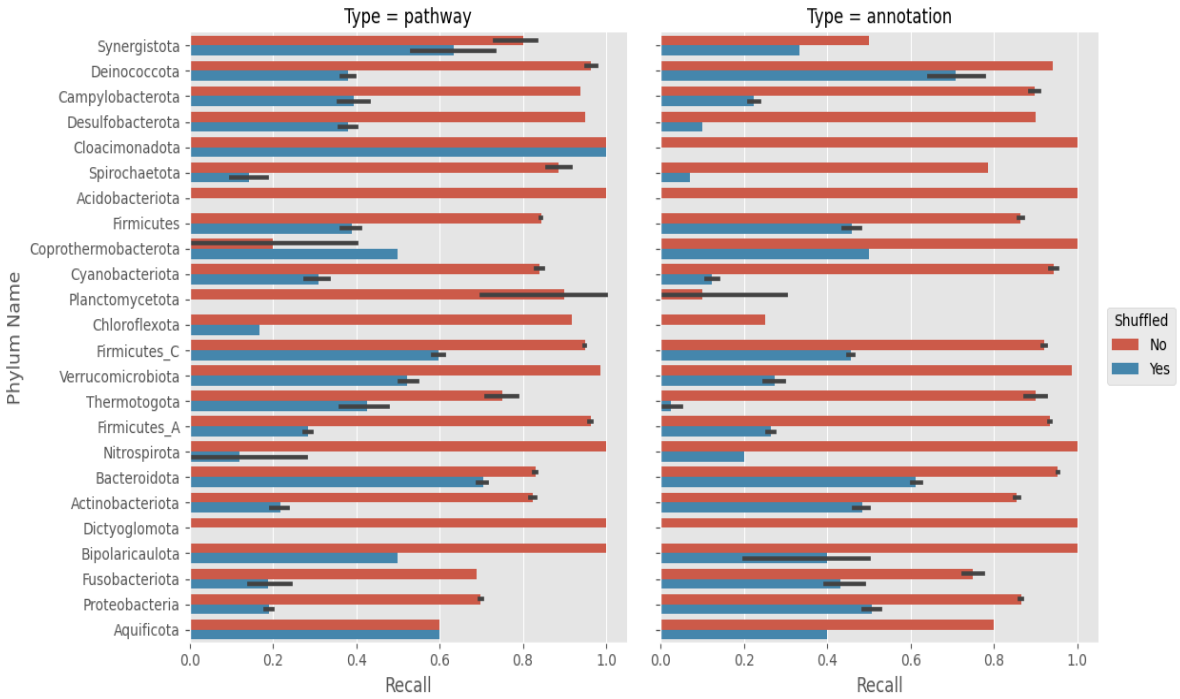
Supplemental Figure 2.6: Performance metrics (Accuracy) for GEM annotations and pathways. A value of 0.0 is the worst and 1.0 is the best. The orange represents real data (i.e. shuffled=" No") and the blue represents our random baseline (i.e. shuffled="Yes"). Phyla are omitted from interpretation if the margin of error of shuffled performance is within our actual predictions across all metrics.



Supplemental Figure 2.7: Performance metrics (AUC) for GEM annotations and pathways. A value of 0.5 is worst for AUC. The orange represents real data (i.e. shuffled=" No") and the blue represents our random baseline (i.e. shuffled="Yes"). Phyla are omitted from interpretation if the margin of error of shuffled performance is within our actual predictions across all metrics.



Supplemental Figure 2.8: Precision for GEM annotations and pathways. A value of 0.0 is worst while a value of 1.0 is the best. The orange represents real data (i.e. shuffled=" No") and the blue represents our random baseline (i.e. shuffled="Yes"). Phyla are omitted from interpretation if the margin of error of shuffled performance is within our actual predictions across all metrics.



Supplemental Figure 2.9: Recall for GEM annotations and pathways. A value of 0.0 is worst while a value of 1.0 is the best. The orange represents real data (i.e. shuffled=" No") and the blue represents our random baseline (i.e. shuffled="Yes"). Phyla are omitted from interpretation if the margin of error of shuffled performance is within our actual predictions across all metrics.

CHAPTER III

THERMOPHILICITY CORRELATES WITH SLOW GROWTH POTENTIAL IN COLD SUBSURFACE

Abstract

Permafrost soil is characterized by prolonged freezing conditions. Thermophilic microbes have been discovered in various permanently cold environments, including permafrost, where they can persist for extended periods. The reason for this apparent mismatch between microbial adaptations and environmental conditions is unclear. Here, we test the hypothesis that thermophilic traits provide selective advantage to extremely slow-growing microbes, even in cold temperatures. We used a computational approach to predict optimal growth rates and several measures of thermophilicity in metagenome-assembled genomes (MAGs) from permafrost and active layer soils in diverse cold regions. We find that in permafrost, where available energy is always low, measures of thermophilicity correlate positively with minimum doubling time, indicating that slow growers in permafrost have more thermophilic traits. This trend is reversed in microbes in active layer soil, in which seasonal thawing, temperature changes, and episodic rain events allow periodic fast growth. Similar trends were observed in the relationship between optimal growth rates and the optimal temperature of nucleoside diphosphate kinase (NDPK), an enzyme whose temperature optimum is known to be correlated to optimal growth temperatures of the host organism. Thermophilic traits within slow growers appear to be environmentally rather than phylogenetically constrained, and thermophilic slow growers share few horizontal gene transfers with other permafrost microbes. These findings suggest that the presence of thermophilic traits in slow-growers appears to be an adaptation to extreme slow growth in a persistently low-energy environment.

Introduction

Permafrost soils, by definition, remain frozen for at least two consecutive years. They are often overlain by the active layer, which consists of the soils that freeze and thaw seasonally (Dobinski, 2011). Despite the inhospitable conditions, a diverse community of microbes remains active in both permafrost and active layer soils (Gilichinsky et al., 2008; Hultman et al., 2015; Mackelprang et al., 2011). Oddly, some of these microbes appear to be thermophilic, even though these environments do not experience high temperatures.

Several studies have reported the presence of thermophiles (organisms that grow at temperatures above 40°C) in permafrost, seasonally frozen active layer soils and other permanently cold environments such as marine sediments with temperature in range of -2° and <15°C (De Rezende et al., 2013; Gilichinsky et al., 2007; Hubert et al., 2009; Vishnivetskaya et al., 2022). Many previous studies have focused on spore-forming thermophiles living in hot environments, and the ability of those microorganisms to survive in dormant state in the cold (De Rezende et al., 2013; Hubert et al., 2009, 2010). However, recent 16S rRNA gene studies indicate the presence of non-spore-forming thermophiles in permanently cold environments (Bulat et al., 2004; Vishnivetskaya et al., 2022). Simultaneous incubation of permafrost samples at a variety of temperatures and in diverse media showed the presence of multiple thermophilic microorganisms. These collectively represented 1% to ~47% of the total cultivable microbial community for both aerobic heterotrophic bacteria, which grew at 55°C, and anaerobic microorganisms, which were able to grow at 75°C (Vishnivetskaya et al., 2022). The latter results provide

evidence that thermophiles might survive in permafrost over long timescales and become active when conditions are favorable.

Understanding how these apparent thermophiles survive in permafrost is challenging, because, in general, thermophiles do not grow in the cold (Marchant et al., 2008; Milojevic et al., 2020, 2022). Previous studies have shown that hyperthermophiles stored in cold sea water can survive up to nine months under laboratory conditions (Mora et al., 2014), suggesting that thermophilic microbes may be transported from hot environments to cold soils. However, permafrost represents a uniquely challenging environment, including sub-freezing temperatures, low water activity, elevated salinity, and limited nutrient availability (Ponder et al., 2005) which demand specific adaptations for microbes to persist (Ayala-del-Río et al., 2010; Bakermans et al., 2003; Bergholz et al., 2009). The presence of apparent thermophiles therefore seems unusual. We hypothesize that the thermophily, or thermophilic-like traits, of certain permafrost microbes represent an adaptation that facilitates an ecological strategy of extremely slow growth at consistently low temperatures.

Computational tools can successfully predict microbial physiological parameters, including optimal growth temperature (OGT) and minimum doubling time, from microbial genomes (Li et al., 2019; Wang et al., 2020; Weissman et al., 2021). To test whether thermophilic-like traits are an adaptation to a slow-growth life strategy, we computationally predicted traits associated with thermophiles, including optimal growth temperature, thermophilic index in genome-wide amino acid composition, and optimal temperatures of the enzyme nucleoside diphosphate kinase (NDPK), for metagenome-

assembled genomes (MAGs) obtained from various permafrost and active layer environments. We also predicted organismal minimum doubling time (i.e., doubling time under optimal growth conditions in pure culture), since long minimum doubling times are a proxy for microbes with an evolutionary strategy oriented towards extreme slow growth (Lloyd & Steen, 2025). We compared these to microbes in the seasonally thawed layer of soils, so called active layer, overlying permafrost in which microbes are periodically able to metabolize and grow rapidly (Allison & Martiny, 2008; Graham et al., 2012; Monteux et al., 2020; Wu et al., 2021). Specifically, we aim to: (1) test whether there is a correlation between minimum doubling time and optimal growth temperature, and whether any correlation differs between permafrost and active layer; (2) ascertain whether phylogeny or environmental factors shape these traits; and (3) investigate the role of horizontal gene transfer in conferring traits to apparent thermophiles in permafrost soils. This study enhances our understanding of the evolutionary and adaptive strategies of permafrost microbes by investigating the thermophilic traits as a potential mechanism for sustaining slow growth in cold environment.

Results

Slow growers in permafrost and active layer

We used gRodon (Weissman et al., 2021) to computationally predict the maximal growth rate - i.e., the fastest growth rate possible under optimal growth conditions calculated as the minimum doubling time - for metagenome-assembled genomes (MAGs) binned from permafrost and active-layer soils from the same alpine permafrost site on Tibetan Plateau (Kang et al., 2024). This data set had a relatively low number of MAGs

(35 and 24, respectively), and we were unable to identify other sites at which MAGs had been binned from both permafrost and overlying active layer soils, so we also selected two active layer sites with numerous, high quality MAGs (Ny-Ålesund, Svalbard and Stordalen Mire palsa, Sweden) plus one permafrost site with comparable quality MAGs (Kolyma-Indigirka Lowland, Siberia).

Perhaps unsurprisingly, MAGs in both permafrost and active layer soils were predicted to grow significantly more slowly than human gut MAGs (**Supplemental Figure 3.1**). There were differences in predicted growth rate among all permafrost and active layer sites, but these differences were small and unlikely to be meaningful in the real world. gRodon is not able to quantitatively predict maximal growth rates above a doubling time of ~5 hours, and in all soil samples, the predicted median growth rate was above 5.33 hours, so these results do not necessarily support actual differences in the maximal growth rates of microbes among these sites (**Supplemental Figure 3.1, Figure 3.2A**).

Thermophilicity is positively correlated with doubling time in permafrost but not in active layer soils

Optimal growth temperature (as predicted by TOME, (G. Li et al., 2019)) and "thermophilic index" (i.e., the fraction of codons in the genome coding for isoleucine, valine, tyrosine, tryptophan, arginine, glutamic acid, and leucine (Zeldovich et al., 2007) were positively correlated to the natural log of predicted minimum doubling time in permafrost MAGs, indicating that slower growers were adapted to higher temperatures, and negatively correlated in the active layer. Only one of these correlations was

statistically significant at α=0.05: the positive correlation between predicted growth temperature and minimum doubling time in alpine permafrost (p=0.0024, R=0.50, n=35) (**Figure 3.1A**).

Statistically significant correlations in the same directions were observed in all other permafrost and active layer soils (**Figure 3.2**). In permafrost, faster-growing microbes (lower predicted minimum doubling times) tended to have lower optimal growth temperatures and thermophilic index, whereas in both active layer soils, faster growers tended to have higher predicted optimal growth temperatures and thermophilic index (regression statistics are noted in **Figure 3.2B**). We predicted the minimum doubling times of 125 MAGs from Siberian permafrost. Because data sets with substantial numbers of MAGs from corresponding permafrost soils and active layers are rare, we compared these predictions to 676 MAGs from two active layer soils: 167 MAGs from Svalbard and 507 MAGs from Stordalen Mire palsa. For these samples, we also predicted the optimum temperature of nucleoside diphosphate kinases (NDPK), a ubiquitous enzyme that catalyzes phosphate transfer from nucleoside triphosphate to nucleoside diphosphate (**Figure 3.2D**). The same correlations were observed for all three proxies for growth temperature.

Thermophilicity of functional genes is correlated to slow growth in permafrost

We hypothesized that functional genes, especially those related to carbon cycling, would reflect more thermophilic traits, particularly in slow-growing microbes from permafrost. Our analysis focused on genes involved in energy production and conversion (COG category 'C') and carbohydrate transport and metabolism (COG category 'G'), as

these play critical roles in soil carbon cycling. Minimum doubling times and thermophilicity were positively correlated for both gene categories (COG category C: $R = 0.35$, $p < 0.05$; COG category G: $R = 0.25$, $p < 0.05$; **Figure 3.3**), suggesting that enhanced stability of these functional genes may support microbial carbon cycling at slow growth rates. Furthermore, there was a negative correlation between minimum doubling time and the abundance of these genes (COG category C: $R = -0.45$, $p < 0.05$; COG category G: $R = -0.35$, $p < 0.05$; **Supplemental Figure 3.2.**) This indicates that slower-growing microbes have fewer genes associated with energy production and carbohydrate metabolism, yet the genes they possess code for enzymes that are more thermostable.

Thermophilic traits and minimum doubling time are independent of phylogeny in permafrost

To assess whether traits such as predicted optimal growth temperature and predicted minimum doubling time are influenced more by evolutionary lineage than environmental conditions, we measured the phylogenetic signal of predicted optimal growth temperature and predicted minimum doubling time using Pagel's lambda (λ) (Pagel, 1999). The phylogenetic signal evaluates whether a trait correlates with phylogenetic relatedness, indicating if shared evolutionary history affects its distribution. As shown in **Table 3.1**, λ values in permafrost were close to 0, indicating weak phylogenetic signal, while those in the active layer were close to 1, indicating a strong phylogenetic signal. This suggests that evolutionary lineage has a minimal effect on trait distribution in permafrost, potentially because traits shaped by environmental adaptation

are under selective pressure that can override phylogenetic influence. In contrast, traits in the active layer show stronger phylogenetic influence, reflecting closer alignment with ancestral lineage.

To further explore this, we applied phylogenetic generalized least squares (PGLS) regression (Symonds & Blomberg, 2014) under a Brownian motion model to examine relationships between predicted optimal growth temperature and predicted minimum doubling time (**Table 3.2**). The slope from this model represents the phylogenetic rate of change in predicted optimal growth temperature as predicted minimum doubling time varies. In permafrost, a non-significant positive correlation was observed between predicted minimum doubling time and predicted optimal growth temperature, consistent with the weak phylogenetic signal. Conversely, in the active layer, predicted minimum doubling time showed a significant negative correlation with predicted optimal growth temperature, suggesting stronger phylogenetic influence on this trait. With nearly balanced sample sizes for permafrost (n = 26) and active layer (n = 27), our results suggest that traits in permafrost are likely shaped more by environmental adaptation rather than ancestral lineage, in contrast to traits in the active layer.

Thermophilic slow growers share few horizontally transferred genes with mesophilic permafrost microbes

Horizontal gene transfer (HGT) plays a significant role in the sharing of adaptive traits, particularly in permafrost environments where microbes have longer generation times (Mackelprang et al., 2017). To investigate whether thermophilic slow growers engage in genetic interactions with other microbes in the permafrost, we categorized

MAGs into four groups based on predicted optimal growth temperature and predicted minimum doubling time data: thermophilic fast growers, thermophilic slow growers, mesophilic fast growers, and mesophilic slow growers. For these purposes, "thermophilic" indicates a predicted optimum growth temperature above 40°C and "slow-grower" indicates a predicted minimum doubling time longer than five hours. In Siberian permafrost, 52 MAGs were identified as mesophilic fast growers, 50 MAGs as mesophilic slow growers, and 12 MAGs as thermophilic slow growers. In the Svalbard active layer, 23 MAGs were classified as mesophilic fast growers, 106 MAGs as mesophilic slow growers, and 4 MAGs as thermophilic fast growers. HGT predictions were made using MetaCHIP (Song et al., 2019), which employs both phylogenetic and best-match approaches.

In permafrost communities, 703 genes were involved in HGT transfers, whereas in the active layer, only 63 genes were identified as HGT genes. In permafrost, the majority of HGT transfers occurred between mesophilic slow and fast growers, with only 7 HGT genes transferred within thermophilic slow growers. In contrast, within the active layer, thermophilic fast growers were more involved in HGT events, with 22 out of 63 HGT genes transferred within this group (**Figure 3.4**). The fewer HGT events observed within the permafrost thermophilic group suggest that thermophilic traits may have evolved independently within this group.

Discussion

Permafrost shows thermophilic intrinsic adaptation to slow growth

Permafrost soils represent low-energy environments. Even though organic matter may be plentiful, cold temperatures and low water availability limit the rate at which metabolic reactions can proceed (Morgalev et al., 2017; Ping et al., 2015). Growth rate predictions from gRodon suggest that the microbes that dominate both permafrost and overlying active layer soils grow slowly compared to microbes in a high-energy environment like the human gut (**Supplemental Figure 3.1**). While this is perhaps unsurprising, there has been disagreement in the literature about whether active-layer soil microbes are predominantly fast- or slow-growing (Li et al., 2024; Scheel et al., 2023). Slow growth in permafrost is consistent with the low rates of metabolic processes, which reflects adaptation to the low energy environment (Bergholz et al., 2009; Jansson & Taş, 2014; Mykytczuk et al., 2013; Price & Sowers, 2004; Sipes et al., 2021).

It is likely that, for the taxa that are predicted to be slow growers in this analysis, predicted minimum doubling time is an underestimate of the true minimum doubling time, which itself is an underestimate of the *in situ* doubling time. gRodon uses codon usage bias to predict growth rate. Codon usage bias is mainly used by microbes to optimize translational efficiency, particularly in highly expressed genes (Frumkin et al., 2018; Gouy & Gautier, 1982; Shah & Gilchrist, 2011; Yannai et al., 2018). Thus, gRodon underestimates the minimum doubling times of taxa with minimum doubling times above ~5 hrs (Weissman et al., 2021). Despite gRodon's tendency to underestimate growth rates of slow growers, it remains superior to alternative tools (Vieira-Silva & Rocha, 2010).

Previous work has shown that the true minimum doubling times of environmental microorganisms (which might be measured in pure cultures growing exponentially) are in turn substantially shorter than the *in situ* doubling times (Gibson et al., 2018), since *in situ* doubling times are a function of a host of environmental factors, in addition to an organism's biology. Thus, genomically predicted optimal growth rates may be a better indication of an organism's preferred growth strategy. Despite the similar abundances of slow-growers between permafrost and active layers, the slow-growers in permafrost differed from those in the active layer in their apparent adaptation to temperature. In the active layer, in which episodic thawing and rainfall events cause microbes to episodically experience warmer temperatures and receive pulses of organic matter (Graham et al., 2012; Pautler et al., 2010; Walz et al., 2017), faster-growing microbes were predicted via both TOME and the thermophilic index to be adapted to higher temperatures. This is consistent with a split between r-strategists, which capitalize on conditions in which fast growth is possible (warm temperature, substrate availability, water availability) and k-strategists, which persist effectively when fast growth rates are not possible, for instance between episodic warm or wet events (Winkler et al., 2017).

The pattern in the permafrost was the opposite: taxa predicted to be slow growers were predicted to be adapted to higher temperatures. Furthermore, many microbes in permafrost had predicted optimal growth temperatures well above temperatures that they would ever experience *in situ*. It is common for microbes to have predicted optimal temperatures on the order of ~10°C above typical *in situ* environmental temperatures (Bárcenas-Moreno et al., 2009; Nottingham et al., 2019; Panikov, 1999), but many

microbes in permafrost had optimal temperatures above 30°C, whereas *in situ* temperatures by definition never exceed ~0°C (Boike et al., 2018).

The trends observed with the microbial thermophilic index and that of the carbohydrate and energy related genes were similar (**Figure 3.1C** & **Figure 3.2**). Also, the predicted optimal temperature for NDPK, the temperature optimum of which has previously been shown to reflect the environmental temperatures experienced by the host organism (Akanuma et al., 2013), match the same trend (**Figure 3.2D**). What would be the value of adaptation to high temperatures in microbes that will never experience those temperatures?

We propose that the apparent high temperature optimum of slow-growing permafrost microbes is not an adaptation to high temperature *per se*, but rather an adaptation to extreme slow growth rates, the class of organism for which the term aeonophile has recently been proposed (Lloyd & Steen, 2025). Theoretical work indicates that as microbial metabolic rates approach zero, the energetic cost of protein repair increases as a fraction of a cell's total energy requirements (Kempes et al., 2017). All else being equal, a cell that turns its biomass carbon over more slowly must either have a lower fraction of active enzymes relative to inactive enzymes, or must have more stable enzymes, than a cell that turns its carbon over faster. The soil carbon turnover time in the permafrost is generally longer than that of the active layer across all ecosystems (J. Li et al., 2023). Adaptations for efficient energy acquisition and substrate utilization reflect strong selective pressures in permafrost, driving microbial community structure and function (Waldrop et al., 2023). Preliminary evidence from deep marine sediments, in

which microbial metabolisms are also remarkably slow, suggests that secreted hydrolases are in fact unusually thermostable (Schmidt et al., 2021).

We note that we are agnostic as to whether the predicted thermophiles identified in these data sets are true thermophiles (or thermotolerant microbes), in the sense of being capable of growth or persistence at high temperatures. Rather, we argue that these microbes' genomes code for the same traits that TOME (Li et al., 2019) identifies as thermophilic, which primarily reflect protein stability. We suggest that enhanced protein stability also facilitates an ultra-slow-growth lifestyle at cold temperatures.

Weak phylogenetic signal in traits correlation associated with permafrost.

Our results indicate that the predicted optimal growth temperature and predicted minimum doubling time in permafrost communities exhibit weak phylogenetic signals. A weak phylogenetic signal indicates that closely related species may not share similar traits, implying that evolutionary history alone cannot account for the observed trait patterns. In such cases, ecological factors, particularly environmental adaptations, likely play a more critical role in shaping community composition (Kassaw et al., 2023; Srivastava et al., 2012). Phylogenetic signal strength can vary across traits and scales. For instance, traits related to methane oxidation kinetics in methane-oxidizing bacteria display weak phylogenetic signals, highlighting the dominant influence of ecological adaptations (Krause et al., 2014). Similarly, weak phylogenetic signals observed in microbial traits related to resource use strategies suggest that environmental factors impact certain traits more than others (Gravuer & Eskelinen, 2017; Malik et al., 2020; McGuinness et al., 2024).

This weak phylogenetic signal in permafrost microbes may also be linked to a high rate of horizontal gene transfer. Our analysis indicates that horizontal gene transfer occurs more frequently in permafrost than in the active layer (**Figure 3.4**). Horizontal gene transfer can facilitate the spread of adaptive traits, especially in extreme environments like permafrost, where microbes have prolonged generation times (Mackelprang et al., 2017). Moreover, horizontal gene transfer events in nature are linked to environmental dynamics and can contribute to the co-existence of microbial populations which further increase the microbial complexity (Babajanyan et al., 2024; Goyal et al., 2020). However, in thermophilic slow growers within permafrost, we observed fewer genes that had been horizontally transferred both to and from other groups, suggesting an isolated evolutionary pathway (**Figure 3.4**). It is unclear whether these thermophilic microbes are indigenous to permafrost or originated from thermophilic environments. Scarce hot springs and fumaroles within discrete permafrost-affected regions can support microbial communities, suggesting possible migration or gene flow (Kochetkova et al., 2020; Vishnivetskaya et al., 2022).

In permafrost or other cold environments, adaptation to slow growth may allow thermophilic microbes to be viable for hundreds to thousands of years (Henry J. Sun, E. Imre Friedmann, 1999; Rivkina et al., 2000). This suggests at low temperatures, thermophilic traits may include nutrient utilization strategies that help counteract cold stress and low energy availability. Furthermore, slow growers may have acquired thermophilic traits to increase their resistance to the cold environments. The unique combination of low temperature and metabolic flexibility in permafrost may thus have

shaped a microbial community that integrates traits for both thermophily and stability, reflecting the complex interplay of ecological and evolutionary processes in this extreme environment.

Methods

Environmental description

The permafrost samples analyzed in this study were collected from various locations with distinct thermal profiles. Siberian permafrost, a "cold" perennially frozen soil, maintains subzero temperatures between -12°C and -8°C at depths of 3-25 meters (Sipes et al., 2021). Permafrost core samples were from borehole AL1_15 with total depth of 25.8 m (Wu et al., 2023), borehole Ch1_17 with total depth of 21.5 m (Liang et al., 2021), and borehole AL3_15 with total depth of 20 m (Sipes et al., 2021), which were collected within the Northeastern Kolyma-Indigirka Lowland Region in Siberia, Russia between 2015-2017 and described in previous studies(Liang et al., 2021; Sipes et al., 2021; Wu et al., 2023). The active layer samples originated from Svalbard and Stordalen Mire. By contrast, Svalbard permafrost has an average temperature of -2.8^{o}C and is classified as "warm" permafrost (Boike et al., 2018). Svalbard active layer samples were collected at Bayelva, Ny Ålesund, Svalbard, from depths of 21–58 cm (Sipes et al., 2022). Stordalen Mire, a discontinuous permafrost region with active layer soil temperatures ranging from 3.5°C to 11°C (Lupascu et al., 2012), soil samples were collected from 1-51 cm depth (Woodcroft et al., 2018). Alpine permafrost and active layers were collected from the Tibetan Plateau, featuring a mean annual temperature of -

4.5°C to 1.8°C; active layer samples were taken from 0-10 cm, and permafrost samples from 150-200 cm (L. Kang et al., 2024).

Metagenome data sets

The permafrost datasets included 125 metagenome-assembled genomes (MAGs) from Siberia (60 from Wu et al., 52 from Liang et al., and 35 from Sipes et al.)(Liang et al., 2021; Sipes et al., 2021; X. Wu et al., 2023) and 35 MAGs from alpine region of the Tibetan plateau. Active layer samples included 167 MAGs from Svalbard (Sipes et al., 2021) and 507 from Stordalen Mire (Woodcroft et al., 2018) and 24 MAGs from the Alpine region. Additionally, we included 957 MAGs from human gut samples (Nayfach et al., 2019). For Alpine MAGs, 66 metagenomic sequence data from permafrost and active layer samples were downloaded from PRJNA1037019 (L. Kang et al., 2024). The raw metagenomic reads were quality-filtered using Trimmomatic (Bolger et al., 2014), assembled with MEGAHIT (Li et al., 2015), and binned into MAGs using MaxBin2 (Y.-W. Wu et al., 2016), CONCOCT (Alneberg et al., 2013), and MetaBAT2 (D. D. Kang et al., 2019), followed by dereplication with DAS Tool (Sieber et al., 2018), checked for completeness and contamination with CheckM (Parks et al., 2015), and taxonomical classified with GTDB-TK (Chaumeil et al., 2020). All these analyses were done on the KBase platform and are available under these narratives (https://narrative.kbase.us/narrative/195818, https://narrative.kbase.us/narrative/191544 and https://narrative.kbase.us/narrative/193190). All MAGs have >50% completeness and <10% contamination. Protein-coding open reading frames (ORFs) in the MAGs were annotated using Prokka 2.6.3 (Seemann, 2014). We searched and filtered Nucleoside

diphosphate kinase sequences from the genome and confirmed the annotation with Pfam. Genes for energy production, and conversion, and carbohydrate transport/metabolism genes were annotated using eggNOG-mapper (Cantalapiedra et al., 2021).

Growth rate Prediction

The bacterial minimum doubling time (a proxy for growth rate) was estimated with gRodon (Weissman et al., 2021), an R package that estimate maximum growth rate from genomic data using codon usage bias. For the Prokka, annotated ribosomal proteins were filtered using the search “ribosomal protein”. Ribosomal proteins are often used as the set of highly expressed genes in gRodon prediction because they are known to have higher translational efficiency and universal across the tree of life (Higgs et al., 2007; Petibon et al., 2021). Because our study is an extreme environment, the Madin et al. dataset (Madin et al., 2020) was selected as the training set of gRodon prediction. This phenotypic dataset included more extremophiles than that complied by Vierira-Silva and Rocha used originally in GrowthPred (Vieira-Silva & Rocha, 2010). The default minimum gene length of 240 bp was used. Despite gRodon's tendency to underestimate slow growers' rates due to transcription/translation efficiency biases, it has higher accuracy with the slow growers compared to other growth rate predictors (Xu et al., 2024) We expect a reliable estimation across permafrost samples since it reflects the overall microbial life style.

Optimal growth temperature and optimal temperature of enzymes and thermophilicity prediction

Optimal growth temperature (OGT) was predicted using TOME (Temperature Optimal for Microorganisms and Enzymes) (G. Li et al., 2019), which is a tool developed by training a machine learning model on amino acid composition and dipeptide frequencies of 21498 microorganisms and their OGT. TOME's suitability for predicting thermophiles makes it ideal for this dataset. For the optimal temperature of Nucleoside Diphosphate Kinase (NDPK), the Tomer, an extended version of TOME, with improved accuracy from resampling was used (Gado et al., 2020). The thermophilic index was computed by taking the sum fraction the hydrophobic amino acid set consisting of Ile, Val, Tyr, Trp, Arg, Glu, Leu ((IVYWREL) in the amino acid sequences of each MAG (Zeldovich et al., 2007). The same analysis was also conducted for COG functional genes related to energy production and conversion (COG C) and carbohydrate metabolism and transport (COG G) of each MAG.

Phylogenetic signals and regression

To explore phylogenetic influences on predicted OGT and predicted minimum doubling time, a phylogenetic tree in IQTREE-2 (Minh et al., 2020) was generated using 16S rRNA sequences extracted with Barrnap (https://github.com/tseemann/barrnap). The phylogenetic signal (λ) was computed for each trait using the phylosig function in package phytools (version 0.7-70) (Revell, 2012) and conducted phylogenetic generalized least squares (PGLS) regression with the phylolm package (version 2.6.2) using the default

parameters (Tung Ho & Ané, 2014). P-values were adjusted for false positives using the Benjamini-Hochberg (BH) method with the p.adjust function in R.

Horizontal gene transfer prediction

MAGs were classified into four categories based on the predicted OGT and minimum doubling time: thermophilic slow growers (TS; OGT > 40°C, doubling time > 5 hrs), thermophilic fast growers (TF; OGT > 40°C, doubling time < 5 hrs), mesophilic slow growers (MS; OGT 20-40°C, doubling time > 5 hrs), and mesophilic fast growers (MF; OGT 20-40°C, doubling time < 5 hrs). The 5 hrs minimum doubling time cut off for slow and fast growers was based on gRodon (Weissman et al., 2021). Putative horizontal gene transfer (HGT) events were predicted using MetaCHIP (Song et al., 2019), employing phylogenetic and best match approaches. Gene flow between categories was visualized using the R circlize package.

Data availability

All metagenomic data sets are available in Supplemental Table 1, and as described by their original publications

References

Akanuma, S., Nakajima, Y., Yokobori, S., Kimura, M., Nemoto, N., Mase, T., Miyazono, K., Tanokura, M., & Yamagishi, A. (2013). Experimental evidence for the thermophilicity of ancestral life. *Proceedings of the National Academy of Sciences*, *110*(27), 11067–11072. https://doi.org/10.1073/pnas.1308215110

Allison, S. D., & Martiny, J. B. H. (2008). Resistance, resilience, and redundancy in microbial communities. *Proceedings of the National Academy of Sciences*, *105*(supplement_1), 11512–11519. https://doi.org/10.1073/pnas.0801925105

Alneberg, J., Bjarnason, B. S., Bruijn, I. de, Schirmer, M., Quick, J., Ijaz, U. Z., Loman, N. J., Andersson, A. F., & Quince, C. (2013). *CONCOCT: Clustering cONtigs on COverage and ComposiTion* (arXiv:1312.4038). arXiv. http://arxiv.org/abs/1312.4038

Ayala-del-Río, H. L., Chain, P. S., Grzymski, J. J., Ponder, M. A., Ivanova, N., Bergholz, P. W., Di Bartolo, G., Hauser, L., Land, M., & Bakermans, C. (2010). The genome sequence of Psychrobacter arcticus 273-4, a psychroactive Siberian permafrost bacterium, reveals mechanisms for adaptation to low-temperature growth. *Applied and Environmental Microbiology*, *76*(7), 2304–2312.

Babajanyan, S. G., Garushyants, S. K., Wolf, Y. I., & Koonin, E. V. (2024). Microbial diversity and ecological complexity emerging from environmental variation and horizontal gene transfer in a simple mathematical model. *bioRxiv*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10836074/

Bakermans, C., Tsapin, A. I., Souza-Egipsy, V., Gilichinsky, D. A., & Nealson, K. H. (2003). Reproduction and metabolism at − 10°C of bacteria isolated from Siberian permafrost. *Environmental Microbiology*, *5*(4), 321–326. https://doi.org/10.1046/j.1462-2920.2003.00419.x

Bárcenas-Moreno, G., Gómez-Brandón, M., Rousk, J., & Bååth, E. (2009). Adaptation of soil microbial communities to temperature: Comparison of fungi and bacteria in a laboratory experiment. *Global Change Biology*, *15*(12), 2950–2957. https://doi.org/10.1111/j.1365-2486.2009.01882.x

Bergholz, P. W., Bakermans, C., & Tiedje, J. M. (2009). Psychrobacter arcticus 273-4 uses resource efficiency and molecular motion adaptations for subzero temperature growth. *Journal of Bacteriology*, *191*(7), 2340–2352.

Boike, J., Juszak, I., Lange, S., Chadburn, S., Burke, E., Overduin, P. P., Roth, K., Ippisch, O., Bornemann, N., & Stern, L. (2018). A 20-year record (1998–2017) of permafrost, active layer and meteorological conditions at a high Arctic permafrost research site (Bayelva, Spitsbergen). *Earth System Science Data*, *10*(1), 355–390.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Bulat, S. A., Alekhina, I. A., Blot, M., Petit, J.-R., De Angelis, M., Wagenbach, D., Lipenkov, V. Y., Vasilyeva, L. P., Wloch, D. M., & Raynaud, D. (2004). DNA signature of thermophilic bacteria from the aged accretion ice of Lake Vostok, Antarctica: Implications for searching for life in extreme icy environments. *International Journal of Astrobiology*, *3*(1), 1–12.

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829.

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). *GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database*. Oxford University Press.

De Rezende, J. R., Kjeldsen, K. U., Hubert, C. R., Finster, K., Loy, A., & Jørgensen, B. B. (2013). Dispersal of thermophilic Desulfotomaculum endospores into Baltic Sea sediments over thousands of years. *The ISME Journal*, *7*(1), 72–84.

Dobinski, W. (2011). Permafrost. *Earth-Science Reviews*, *108*(3–4), 158–169.

Frumkin, I., Lajoie, M. J., Gregg, C. J., Hornung, G., Church, G. M., & Pilpel, Y. (2018). Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proceedings of the National Academy of Sciences*, *115*(21). https://doi.org/10.1073/pnas.1719375115

Gado, J. E., Beckham, G. T., & Payne, C. M. (2020). Improving enzyme optimum temperature prediction with resampling strategies and ensemble learning. *Journal of Chemical Information and Modeling*, *60*(8), 4098–4107.

Gibson, B., Wilson, D. J., Feil, E., & Eyre-Walker, A. (2018). The distribution of bacterial doubling times in the wild. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1880), 20180789. https://doi.org/10.1098/rspb.2018.0789

Gilichinsky, D. A., Wilson, G. S., Friedmann, E. I., Mckay, C. P., Sletten, R. S., Rivkina, E. M., Vishnivetskaya, T. A., Erokhina, L. G., Ivanushkina, N. E., Kochkina, G. A., Shcherbakova, V. A., Soina, V. S., Spirina, E. V., Vorobyova, E. A., Fyodorov-Davydov, D. G., Hallet, B., Ozerskaya, S. M., Sorokovikov, V. A., Laurinavichyus, K. S., … Tiedje, J. M. (2007). Microbial Populations in Antarctic Permafrost: Biodiversity, State, Age, and Implication for Astrobiology. *Astrobiology*, *7*(2), 275–311. https://doi.org/10.1089/ast.2006.0012

Gilichinsky, D., Vishnivetskaya, T., Petrova, M., Spirina, E., Mamykin, V., & Rivkina, E. (2008). Bacteria in Permafrost. In R. Margesin, F. Schinner, J.-C. Marx, & C. Gerday (Eds.), *Psychrophiles: From Biodiversity to Biotechnology* (pp. 83–102). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74335-4_6

Gouy, M., & Gautier, C. (1982). Codon usage in bacteria: Correlation with gene expressivity. *Nucleic Acids Research*, *10*(22), 7055–7074.

Goyal, A., Gelbwaser-Klimovsky, D., & Gore, J. (2020). Horizontal gene transfer becomes disadvantageous in rapidly fluctuating environments. *bioRxiv*, 2020–08.

Graham, D. E., Wallenstein, M. D., Vishnivetskaya, T. A., Waldrop, M. P., Phelps, T. J., Pfiffner, S. M., Onstott, T. C., Whyte, L. G., Rivkina, E. M., & Gilichinsky, D. A. (2012). Microbes in thawing permafrost: The unknown variable in the climate change equation. *The ISME Journal*, *6*(4), 709–712.

Gravuer, K., & Eskelinen, A. (2017). Nutrient and rainfall additions shift phylogenetically estimated traits of soil microbial communities. *Frontiers in Microbiology*, *8*, 1271.

Henry J. Sun, E. Imre Friedmann. (1999). Growth on Geological Time Scales in the Antarctic Cryptoendolithic Microbial Community. *Geomicrobiology Journal*, *16*(2), 193–202. https://doi.org/10.1080/014904599270686

Higgs, P. G., Hao, W., & Golding, G. B. (2007). Identification of Conflicting Selective Effects on Highly Expressed Genes. *Evolutionary Bioinformatics*, *3*, 117693430700300015. https://doi.org/10.1177/117693430700300015

Hubert, C., Arnosti, C., Brüchert, V., Loy, A., Vandieken, V., & Jørgensen, B. B. (2010). Thermophilic anaerobes in Arctic marine sediments induced to mineralize complex organic matter at high temperature. *Environmental Microbiology*, *12*(4), 1089–1104. https://doi.org/10.1111/j.1462-2920.2010.02161.x

Hubert, C., Loy, A., Nickel, M., Arnosti, C., Baranyi, C., Brüchert, V., Ferdelman, T., Finster, K., Christensen, F. M., & Rosa de Rezende, J. (2009). A constant flux of diverse thermophilic bacteria into the cold Arctic seabed. *Science*, *325*(5947), 1541–1544.

Hultman, J., Waldrop, M. P., Mackelprang, R., David, M. M., McFarland, J., Blazewicz, S. J., Harden, J., Turetsky, M. R., McGuire, A. D., & Shah, M. B. (2015). Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature*, *521*(7551), 208–212.

Jansson, J. K., & Taş, N. (2014). The microbial ecology of permafrost. *Nature Reviews Microbiology*, *12*(6), 414–425.

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ, 7*, e7359.

Kang, L., Song, Y., Mackelprang, R., Zhang, D., Qin, S., Chen, L., Wu, L., Peng, Y., & Yang, Y. (2024). Metagenomic insights into microbial community structure and metabolism in alpine permafrost on the Tibetan Plateau. *Nature Communications*, *15*(1), 5920.

Kassaw, E., Masresha, G., & Asefa, M. (2023). Environment differentially affects the functional and phylogenetic structures of plant communities in a dry evergreen Afromontane tropical forest. *Biotropica*, *55*(6), 1203–1214. https://doi.org/10.1111/btp.13260

Kempes, C. P., van Bodegom, P. M., Wolpert, D., Libby, E., Amend, J., & Hoehler, T. (2017). Drivers of bacterial maintenance and minimal energy requirements. *Frontiers in Microbiology*, *8*, 31.

Kochetkova, T. V., Toshchakov, S. V., Zayulina, K. S., Elcheninov, A. G., Zavarzina, D. G., Lavrushin, V. Y., Bonch-Osmolovskaya, E. A., & Kublanov, I. V. (2020). Hot in cold: Microbial life in the hottest springs in permafrost. *Microorganisms*, *8*(9), 1308.

Krause, S., van Bodegom, P. M., Cornwell, W. K., & Bodelier, P. L. (2014). Weak phylogenetic signal in physiological traits of methane-oxidizing bacteria. *Journal of Evolutionary Biology*, *27*(6), 1240–1247.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674–1676.

Li, G., Rabe, K. S., Nielsen, J., & Engqvist, M. K. (2019). Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. *ACS Synthetic Biology*, *8*(6), 1411–1420.

Li, J., Ding, J., Yang, S., Zhao, L., Li, J., Huo, H., Wang, M., Tan, J., Cao, Y., & Ren, S. (2023). Depth-dependent driver of global soil carbon turnover times. *Soil Biology and Biochemistry*, *185*, 109149.

Li, Y., Xue, Y., Roy Chowdhury, T., Graham, D. E., Tringe, S. G., Jansson, J. K., & Taş, N. (2024). Genomic insights into redox-driven microbial processes for carbon decomposition in thawing Arctic soils and permafrost. *mSphere*, *9*(7), e00259-24. https://doi.org/10.1128/msphere.00259-24

Liang, R., Li, Z., Lau Vetter, M. C. Y., Vishnivetskaya, T. A., Zanina, O. G., Lloyd, K. G., Pfiffner, S. M., Rivkina, E. M., Wang, W., Wiggins, J., Miller, J., Hettich, R. L., & Onstott, T. C. (2021). Genomic reconstruction of fossil and living microorganisms in ancient Siberian permafrost. *Microbiome*, *9*(1), 110. https://doi.org/10.1186/s40168-021-01057-2

Lloyd, K. G., & Steen, A. D. (2025). Defining ultra-slow-growing extremophilic microorganisms as aeonophiles. *Nature Microbiology*, 1–3.

Lupascu, M., Wadham, J. L., Hornibrook, E. R. C., & Pancost, R. D. (2012). Temperature Sensitivity of Methane Production in the Permafrost Active Layer at

Stordalen, Sweden: A Comparison with Non-permafrost Northern Wetlands. *Arctic, Antarctic, and Alpine Research*, *44*(4), 469–482. https://doi.org/10.1657/1938-4246-44.4.469

Mackelprang, R., Burkert, A., Haw, M., Mahendrarajah, T., Conaway, C. H., Douglas, T. A., & Waldrop, M. P. (2017). Microbial survival strategies in ancient permafrost: Insights from metagenomics. *The ISME Journal*, *11*(10), 2305–2318.

Mackelprang, R., Waldrop, M. P., DeAngelis, K. M., David, M. M., Chavarria, K. L., Blazewicz, S. J., Rubin, E. M., & Jansson, J. K. (2011). Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature*, *480*(7377), 368–371.

Madin, J. S., Nielsen, D. A., Brbic, M., Corkrey, R., Danko, D., Edwards, K., Engqvist, M. K., Fierer, N., Geoghegan, J. L., & Gillings, M. (2020). A synthesis of bacterial and archaeal phenotypic trait data. *Scientific Data*, *7*(1), 170.

Malik, A. A., Martiny, J. B., Brodie, E. L., Martiny, A. C., Treseder, K. K., & Allison, S. D. (2020). Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. *The ISME Journal*, *14*(1), 1–9.

Marchant, R., Franzetti, A., Pavlostathis, S. G., Tas, D. O., Erdbr\Hugger, I., \HUnyayar, A., Mazmanci, M. A., & Banat, I. M. (2008). Thermophilic bacteria in cool temperate soils: Are they metabolically active or continually added by global atmospheric transport? *Applied Microbiology and Biotechnology*, *78*, 841–852.

McGuinness, B., Weber, S. C., & Guichard, F. (2024). Resource-use plasticity governs the causal relationship between traits and community structure in model microbial communities. *bioRxiv*, 2024–08.

Milojevic, T., Cramm, M. A., Hubert, C. R., & Westall, F. (2022). "Freezing" Thermophiles: From One Temperature Extreme to Another. *Microorganisms*, *10*(12), 2417.

Milojevic, T., Zebec, Z., & Schimak, M. P. (2020). Cultivation with powdered meteorite (NWA 1172) as the substrate enhances low-temperature preservation of the extreme thermoacidophile Metallosphaera sedula. *Frontiers in Astronomy and Space Sciences*, *7*, 37.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*(5), 1530–1534.

Monteux, S., Keuper, F., Fontaine, S., Gavazov, K., Hallin, S., Juhanson, J., Krab, E. J., Revaillot, S., Verbruggen, E., & Walz, J. (2020). Carbon and nitrogen cycling in Yedoma permafrost controlled by microbial functional limitations. *Nature Geoscience*, *13*(12), 794–798.

Mora, M., Bellack, A., Ugele, M., Hopf, J., & Wirth, R. (2014). The Temperature Gradient-Forming Device, an Accessory Unit for Normal Light Microscopes To Study the Biology of Hyperthermophilic Microorganisms. *Applied and*

Environmental Microbiology, *80*(15), 4764–4770. https://doi.org/10.1128/AEM.00984-14

Morgalev, Y. N., Lushchaeva, I. V., Morgaleva, T. G., Kolesnichenko, L. G., Loiko, S. V., Krickov, I. V., Lim, A., Raudina, T. V., Volkova, I. I., & Shirokova, L. S. (2017). Bacteria primarily metabolize at the active layer/permafrost border in the peat core from a permafrost region in western Siberia. *Polar Biology*, *40*, 1645–1659.

Mykytczuk, N., Foote, S. J., Omelon, C. R., Southam, G., Greer, C. W., & Whyte, L. G. (2013). Bacterial growth at- 15 C; molecular insights from the permafrost bacterium Planococcus halocryophilus Or1. *The ISME Journal*, *7*(6), 1211–1226.

Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S., & Kyrpides, N. C. (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature*, *568*(7753), 505–510.

Nottingham, A. T., Bååth, E., Reischke, S., Salinas, N., & Meir, P. (2019). Adaptation of soil microbial growth to temperature: Using a tropical elevation gradient to predict future changes. *Global Change Biology*, *25*(3), 827–838. https://doi.org/10.1111/gcb.14502

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*(6756), 877–884.

Panikov, N. S. (1999). Understanding and prediction of soil microbial community dynamics under global change. *Applied Soil Ecology*, *11*(2–3), 161–176.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055.

Pautler, B. G., Simpson, A. J., Mcnally, D. J., Lamoureux, S. F., & Simpson, M. J. (2010). Arctic Permafrost Active Layer Detachments Stimulate Microbial Activity and Degradation of Soil Organic Matter. *Environmental Science & Technology*, *44*(11), 4076–4082. https://doi.org/10.1021/es903685j

Petibon, C., Malik Ghulam, M., Catala, M., & Abou Elela, S. (2021). Regulation of ribosomal protein genes: An ordered anarchy. *WIREs RNA*, *12*(3), e1632. https://doi.org/10.1002/wrna.1632

Ping, C. L., Jastrow, J. D., Jorgenson, M. T., Michaelson, G. J., & Shur, Y. L. (2015). Permafrost soils and carbon cycling. *Soil*, *1*(1), 147–171.

Ponder, M. A., Gilmour, S. J., Bergholz, P. W., Mindock, C. A., Hollingsworth, R., Thomashow, M. F., & Tiedje, J. M. (2005). Characterization of potential stress responses in ancient Siberian permafrost psychroactive bacteria. *FEMS Microbiology Ecology*, *53*(1), 103–115.

Price, P. B., & Sowers, T. (2004). Temperature dependence of metabolic rates for microbial growth, maintenance, and survival. *Proceedings of the National Academy of Sciences*, *101*(13), 4631–4636. https://doi.org/10.1073/pnas.0400522101

Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *2*, 217–223.

Rivkina, E. M., Friedmann, E. I., McKay, C. P., & Gilichinsky, D. A. (2000). Metabolic Activity of Permafrost Bacteria below the Freezing Point. *Applied and Environmental Microbiology*, *66*(8), 3230–3233. https://doi.org/10.1128/AEM.66.8.3230-3233.2000

Scheel, M., Zervas, A., Rijkers, R., Tveit, A. T., Ekelund, F., Campuzano Jiménez, F., Christensen, T. R., & Jacobsen, C. S. (2023). Abrupt permafrost thaw triggers activity of copiotrophs and microbiome predators. *FEMS Microbiology Ecology*, *99*(11), fiad123.

Schmidt, J. M., Royalty, T. M., Lloyd, K. G., & Steen, A. D. (2021). Potential activities and long lifetimes of organic carbon-degrading extracellular enzymes in deep subsurface sediments of the Baltic sea. *Frontiers in Microbiology*, *12*, 702015.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069.

Shah, P., & Gilchrist, M. A. (2011). Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proceedings of the National Academy of Sciences*, *108*(25), 10231–10236. https://doi.org/10.1073/pnas.1016719108

Sieber, C. M., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield, J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, *3*(7), 836–843.

Sipes, K., Almatari, A., Eddie, A., Williams, D., Spirina, E., Rivkina, E., Liang, R., Onstott, T. C., Vishnivetskaya, T. A., & Lloyd, K. G. (2021). Eight metagenome-assembled genomes provide evidence for microbial adaptation in 20,000-to 1,000,000-year-old Siberian permafrost. *Applied and Environmental Microbiology*, *87*(19), e00972-21.

Sipes, K., Paul, R., Fine, A., Li, P., Liang, R., Boike, J., Onstott, T. C., Vishnivetskaya, T. A., Schaeffer, S., & Lloyd, K. G. (2022). Permafrost active layer microbes from Ny Ålesund, Svalbard (79° N) show autotrophic and heterotrophic metabolisms with diverse carbon-degrading enzymes. *Frontiers in Microbiology*, *12*, 757812.

Song, W., Wemheuer, B., Zhang, S., Steensen, K., & Thomas, T. (2019). MetaCHIP: Community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. *Microbiome*, *7*(1), 36. https://doi.org/10.1186/s40168-019-0649-y

Srivastava, D. S., Cadotte, M. W., MacDonald, A. A. M., Marushia, R. G., & Mirotchnick, N. (2012). Phylogenetic diversity and the functioning of ecosystems. *Ecology Letters*, *15*(7), 637–648. https://doi.org/10.1111/j.1461-0248.2012.01795.x

Symonds, M. R. E., & Blomberg, S. P. (2014). A Primer on Phylogenetic Generalised Least Squares. In L. Z. Garamszegi (Ed.), *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology* (pp. 105–130). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-43550-2_5

Tung Ho, L. si, & Ané, C. (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology*, *63*(3), 397–408.

Vieira-Silva, S., & Rocha, E. P. (2010). The systemic imprint of growth and its uses in ecological (meta) genomics. *PLoS Genetics*, *6*(1), e1000808.

Vishnivetskaya, T. A., Mironov, V. A., Abramov, A. A., Shcherbakova, V. A., & Rivkina, E. M. (2022). Biogeochemical Characteristics of Earth's Volcanic Permafrost: An Analog of Extraterrestrial Environments. *Astrobiology*.

Waldrop, M. P., Chabot, C. L., Liebner, S., Holm, S., Snyder, M. W., Dillon, M., Dudgeon, S. R., Douglas, T. A., Leewis, M.-C., & Walter Anthony, K. M. (2023). Permafrost microbial communities and functional genes are structured by latitudinal and soil geochemical gradients. *The ISME Journal*, *17*(8), 1224–1235.

Walz, J., Knoblauch, C., Böhme, L., & Pfeiffer, E.-M. (2017). Regulation of soil organic matter decomposition in permafrost-affected Siberian tundra soils-Impact of oxygen availability, freezing and thawing, temperature, and labile organic matter. *Soil Biology and Biochemistry*, *110*, 34–43.

Wang, X.-F., Gao, P., Liu, Y.-F., Li, H.-F., & Lu, F. (2020). Predicting thermophilic proteins by machine learning. *Current Bioinformatics*, *15*(5), 493–502.

Weissman, J. L., Hou, S., & Fuhrman, J. A. (2021). Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *Proceedings of the National Academy of Sciences*, *118*(12), e2016810118. https://doi.org/10.1073/pnas.2016810118

Winkler, M.-K., Boets, P., Hahne, B., Goethals, P., & Volcke, E. I. (2017). Effect of the dilution rate on microbial competition: R-strategist can win over k-strategist at low substrate concentration. *PLoS One*, *12*(3), e0172785.

Woodcroft, B. J., Singleton, C. M., Boyd, J. A., Evans, P. N., Emerson, J. B., Zayed, A. A., Hoelzle, R. D., Lamberton, T. O., McCalley, C. K., & Hodgkins, S. B. (2018). Genome-centric view of carbon processing in thawing permafrost. *Nature*, *560*(7716), 49–54.

Wu, X., Almatari, A. L., Cyr, W. A., Williams, D. E., Pfiffner, S. M., Rivkina, E. M., Lloyd, K. G., & Vishnivetskaya, T. A. (2023). Microbial life in 25-m-deep boreholes in ancient permafrost illuminated by metagenomics. *Environmental Microbiome*, *18*(1), 33. https://doi.org/10.1186/s40793-023-00487-9

Wu, X., Chauhan, A., Layton, A. C., Lau Vetter, M. C. Y., Stackhouse, B. T., Williams, D. E., Whyte, L., Pfiffner, S. M., Onstott, T. C., & Vishnivetskaya, T. A. (2021). Comparative Metagenomics of the Active Layer and Permafrost from Low-Carbon Soil in the Canadian High Arctic. *Environmental Science & Technology*, *55*(18), 12683–12693. https://doi.org/10.1021/acs.est.1c00802

Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, *32*(4), 605–607.

Xu, L., Zakem, E., & Weissman, J. L. (2024). Improved maximum growth rate prediction from microbial genomes by integrating phylogenetic information. *bioRxiv*, 2024–10.

Yannai, A., Katz, S., & Hershberg, R. (2018). The codon usage of lowly expressed genes is subject to natural selection. *Genome Biology and Evolution*, *10*(5), 1237–1246.

Zeldovich, K. B., Berezovsky, I. N., & Shakhnovich, E. I. (2007). Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Computational Biology*, *3*(1), e5.

Appendix III

Tables

Table 3.1: The phylogenetic signals of traits in permafrost (SIB_PF) and active layer (SVA_AL). PMDT indicated predicted minimum doubling time; POGT indicates predicted optimal growth temperature.

SIB_PF				**SVA_AL**		
Traits	n	Pagel′s λ	p	n	Pagel′s λ	p
PMDT	26	< 0.001	>0.05	27	0.999	1.337×10^{-8}
POGT	26	< 0.001	>0.05	27	0.999	3.114×10^{-9}

Table 3.2: PGLS regression of PMDT and POGT in permafrost (SIB_PF) and active layer (SVA_AL)

Traits	**SIB_PF**					**SVA_AL**				
	Slope	p	PBH	R^2	Adj. R^2	Slope	p	PBH	R^2	Adj.R^2
PMDT vs POGT	2.990	0.112	0.223	0.128	0.082	-5.190	4.274×10^{-5}	8.551×10^{-5}	0.495	0.475

Figures

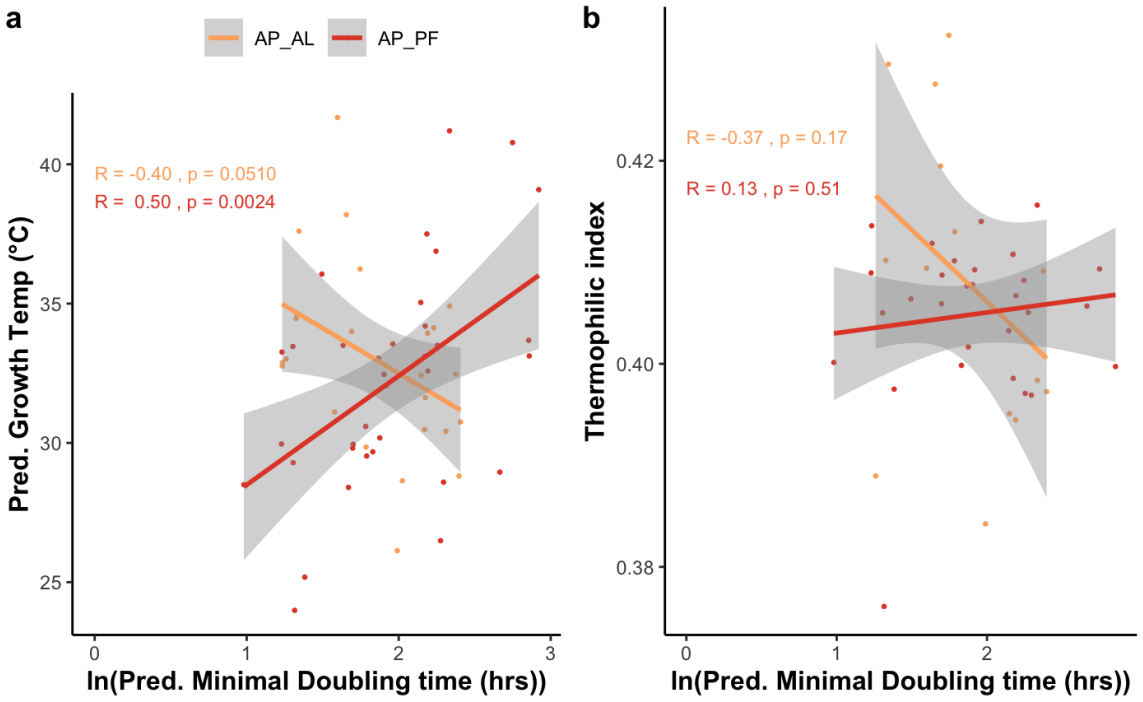


Figure 3.1: Correlations between predicted minimum doubling time and either (a) predicted growth temperature by TOME; or (b) thermophilic index in permafrost and active layer from the Tibetan Plateau alpine region. Thermophilic index is the fraction amount of amino acid set: Ile, Val, Tyr, Trp, Arg, Glu, Leu (IVYWREL) which serves as indicator for thermophilicity and thermostability Abbreviations: AP_AL - alpine active layer; AP_PF - alpine permafrost.

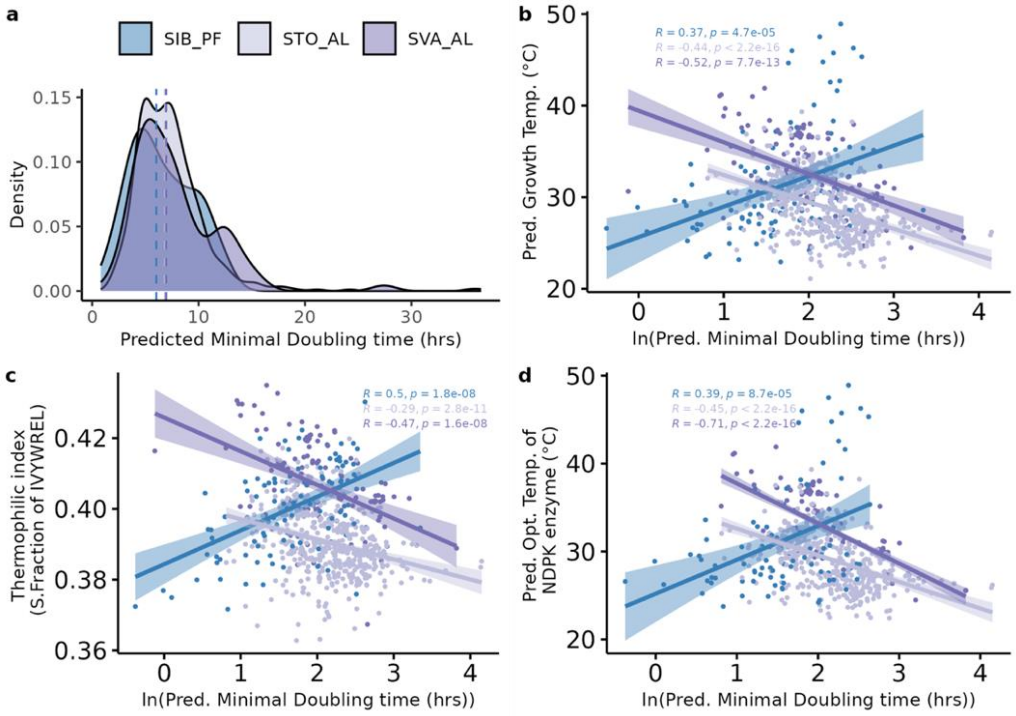


Figure 3.2: Thermophilicity is correlated with minimum doubling time. (a) Predicted minimum doubling time (hrs) of microbes from permafrost and active layer using gRodon (Weissman et al., 2021). The vertical dashed lines represent the median of the predicted minimum doubling time. Abbreviation as follows: SVA_AL - Svalbard active layer; STO_AL - Stordalen Mire palsa active layer; SIB_PF - Siberian permafrost. (b-d) Relationship between thermophilicity and growth rate in permafrost and active layer soils. Thermophilicity was measured by predicted optimal growth temperature as predicted by TOME (G. Li et al., 2019) (b), by the optimal activity temperature for NDPK as predicted by TOMER (Gado et al., 2020), and as the thermophilic index. All

correlation analyses were two tailed and the shaded areas around the regression line represent the 95% confidence interval.

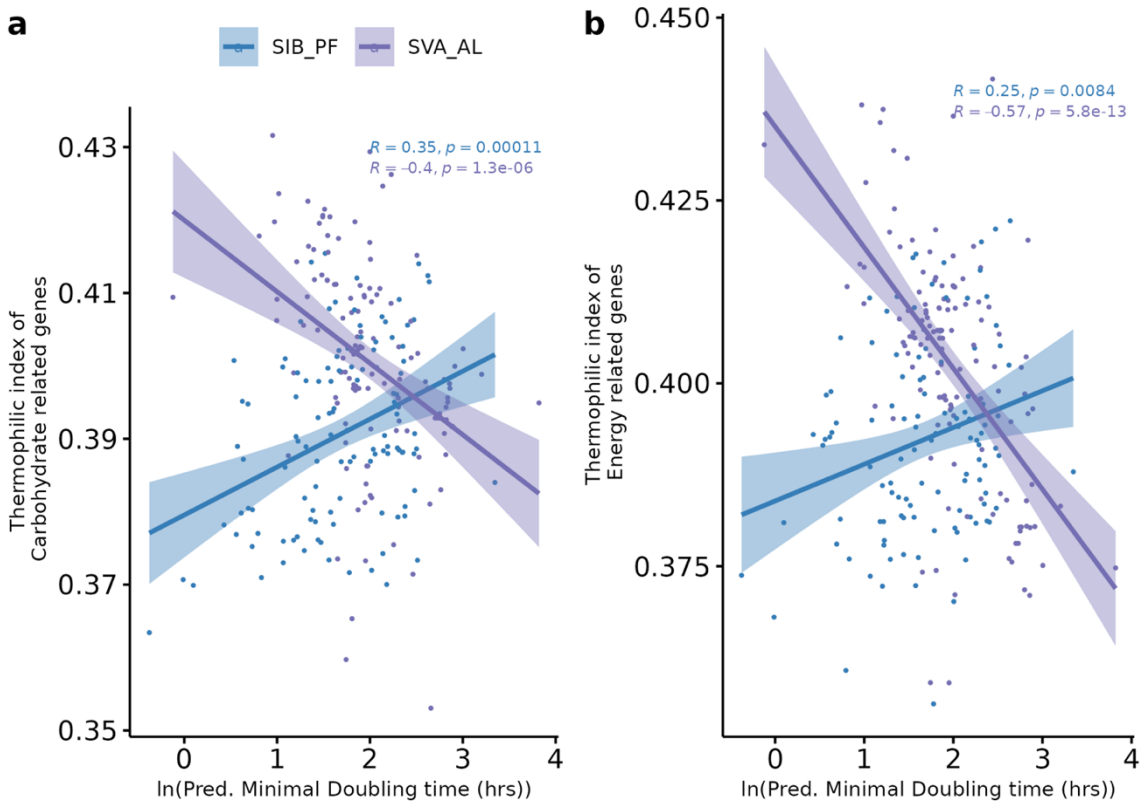


Figure 3.3: Correlations between predicted minimum doubling time and thermophilic index either specific to (a) annotated carbohydrate transport and metabolism genes (NCBI COG category 'G') or specific to (b) energy production and conversion genes (NCBI COG category 'C'). The correlation analyses in a and b are two tailed. The shaded areas around the regression line show the 95% confidence interval.

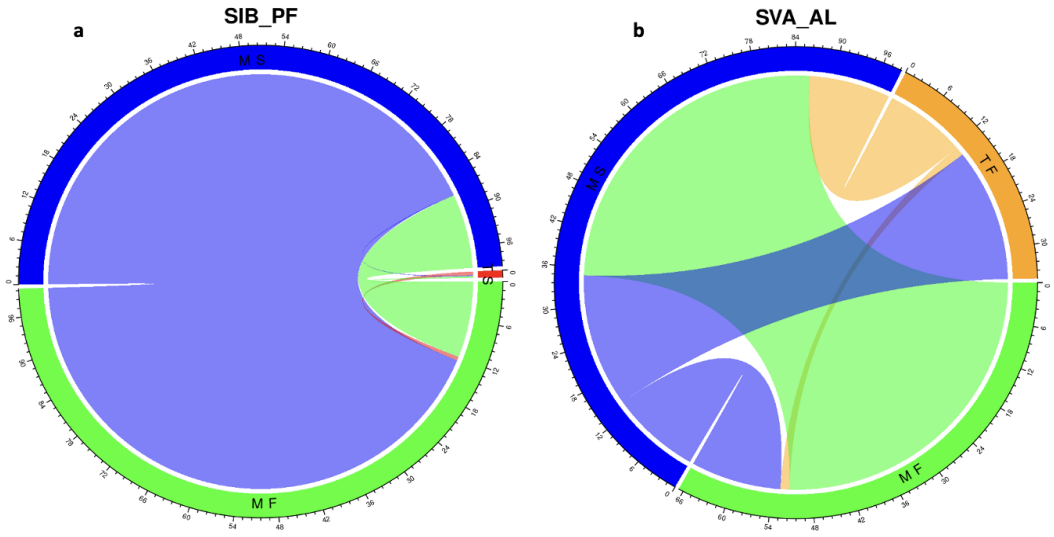
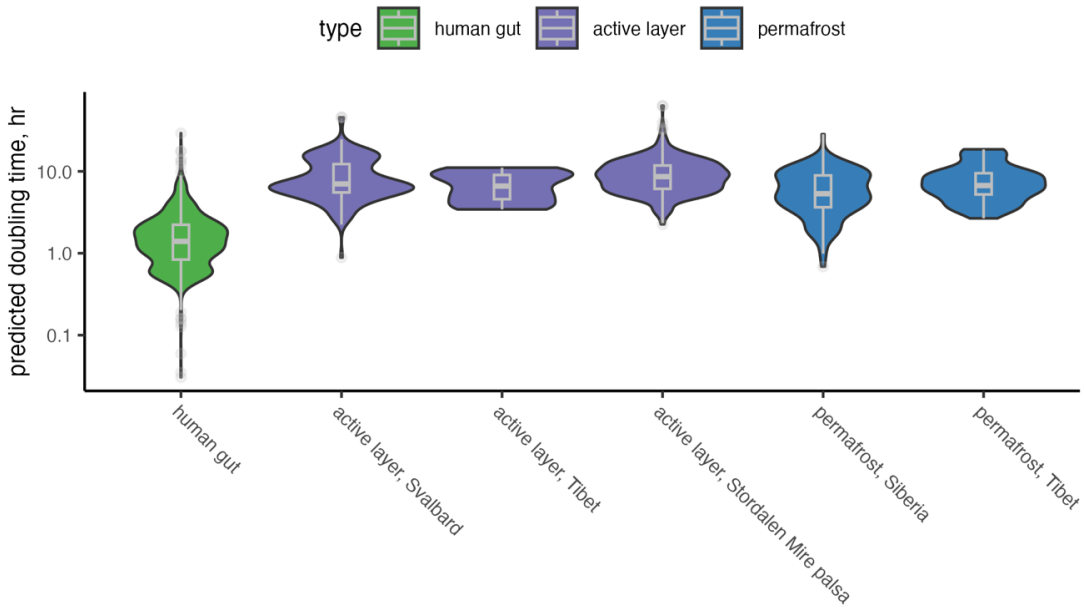


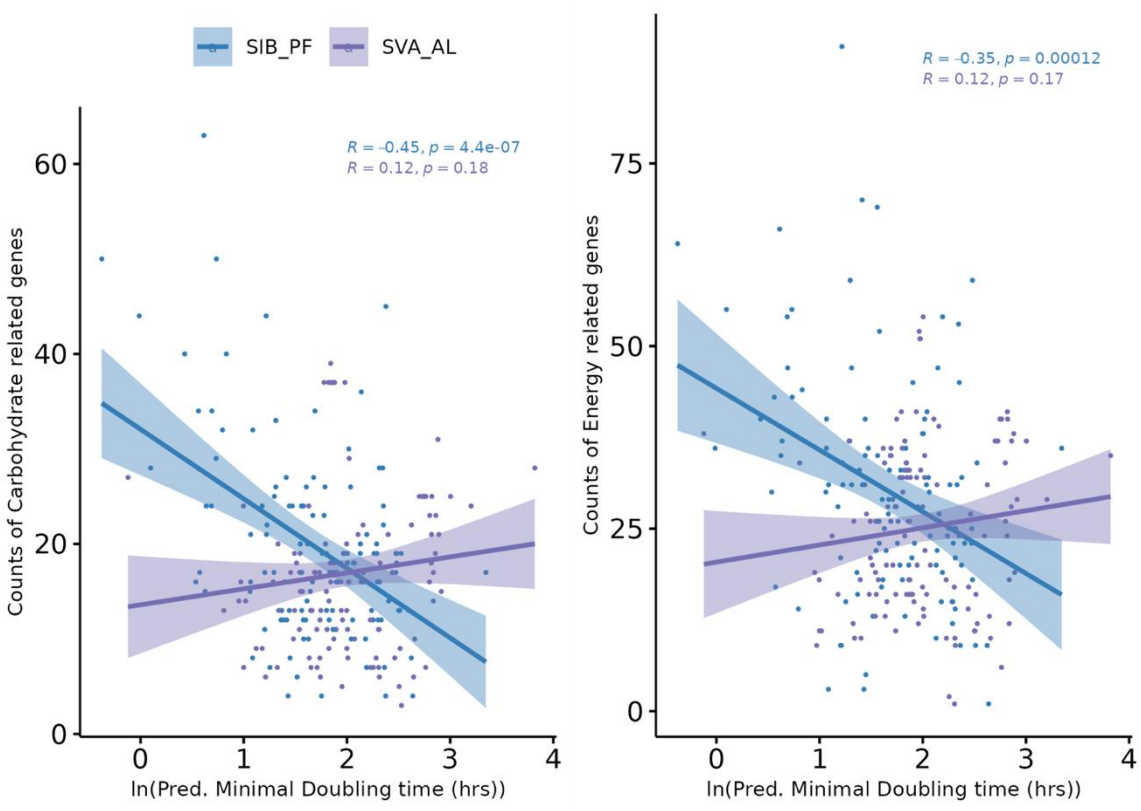
Figure 3.4: Predicted horizontal gene transfer events within the permafrost and active layer microbial communities. Bands connect the donors and recipients. The band's color corresponds to the donors and the band width represents the number of HGTs. The customized groups include Thermophilic slow growers, TS (predicted optimal growth temperature > 40°C, predicted minimum doubling time > 5 hrs), Thermophilic fast growers, TF (predicted optimal growth temperature > 40°C, predicted minimum doubling time < 5 hrs), Mesophilic slow growers, MS (predicted optimal growth temperature between 20°C and 40°C, predicted minimum doubling time > 5 hrs), Mesophilic fast growers, MF (predicted optimal growth temperature between 20°C and 40°C, predicted minimum doubling time < 5 hrs). The 5 hrs cut off is based on gRodon classification of

slow and fast growers (Weissman et al., 2021). The number outside the grid represents the matrix percentage of each donor.

Supplemental figures



Supplemental Figure 3.1: Comparing the predicted minimum doubling time of all the active layers and permafrost to human gut. MAGs in the active layer and permafrost have higher minimum doubling time (slower growth rate) than that of the human gut.



Supplemental Figure 3.2: Correlation of minimum doubling time with the counts of annotated genes relating to energy production and conversion (a) and carbohydrate transports and metabolism (b). MAGs with predicted slower growth rates have lesser number of genes in energy production and carbohydrate transports.

Supplemental Table

Supplemental Table 3.1: Data sources for permafrost and active layer MAGs

Site	# MAGs	Median completeness	Median contamination	BioProject accession no	Original reference
Tibetan plateau (Active layer)	24	73.67	3.86	PRJNA1037019	Kang, L., Song, Y., Mackelprang, R., Zhang, D., Qin, S., Chen, L., ... & Yang, Y. (2024). Metagenomic insights into microbial community structure and metabolism in alpine permafrost on the Tibetan Plateau. *Nature Communications*, *15*(1), 5920.
Tibetan plateau (Permafrost)	35	69.03	5.00	PRJNA1037019	Kang, L., Song, Y., Mackelprang, R., Zhang, D., Qin, S., Chen, L., ... & Yang, Y. (2024). Metagenomic insights into microbial community structure and metabolism in alpine permafrost on the Tibetan Plateau. *Nature Communications*, *15*(1), 5920.
Svalbard (active layer)	167	80.42	3.42	PRJNA971479	Sipes, K., Buongiorno, J., Steen, A. D., Abramov, A. A., Abuah, C., Peters, S. L., ... & Lloyd, K. G. (2024). Depth-specific distribution of bacterial MAGs in permafrost active layer in Ny Ålesund, Svalbard (79° N). *Systematic and Applied Microbiology*, *47*(6), 126544.
Stordalen (active layer)	507	88.34	2.24	PRJNA386568	Woodcroft, B. J., Singleton, C. M., Boyd, J. A., Evans, P. N., Emerson, J. B., Zayed, A. A., ... & Tyson, G. W. (2018). Genome-centric view of carbon processing in thawing permafrost. *Nature*, *560*(7716), 49-54.
Siberia (permafrost)	125	80.61	3.01	PRJNA596250, PRJNA680161, PRJNA601698	Sipes, K., Almatari, A., Eddie, A., Williams, D., Spirina, E., Rivkina, E., ... & Lloyd, K. G. (2021). Eight metagenome-assembled genomes provide evidence for microbial adaptation in 20,000-to 1,000,000-year-old Siberian permafrost. *Applied and Environmental Microbiology*, *87*(19), e00972-21 Liang, R., Li, Z., Lau Vetter, M. C., Vishnivetskaya, T. A., Zanina, O. G., Lloyd, K. G., ... & Onstott, T. C. (2021). Genomic reconstruction of fossil and living microorganisms in ancient Siberian permafrost. *Microbiome*, *9*, 1-20. Wu, X., Almatari, A. L., Cyr, W. A., Williams, D. E., Pfiffner, S. M., Rivkina, E. M., ... & Vishnivetskaya, T. A. (2023). Microbial life in 25-m-deep boreholes in ancient permafrost illuminated by metagenomics. *Environmental Microbiome*, *18*(1), 33.

CHAPTER IV

TRANSFORMER-BASED PREDICTION OF MICROBIAL GROWTH RATES FROM GENOMIC DATA

Abstract

Accurate estimation of microbial growth rates is essential for understanding microbial life strategies, their ecological roles, and cultivation potential. Existing computational methods that infer growth rates---particularly minimum doubling times---using codon usage bias (CUB) often struggle with low accuracy for slow-growing microbes. Here, we present a novel transformer-based deep learning model called LookingGlass2 tailored for microbial genomes to predict doubling times directly from genomic features. Fine-tuned on ribosomal protein-coding genes, our model consistently outperforms the widely used CUB-based estimator gRodon, achieving higher predictive accuracy and stronger correlations with experimentally measured growth rates. Analysis of 25,000+ microbial genomes from the "Genomes from Earth's Microbiomes" (GEM) catalog reveals persistent cultivation biases toward fast-growing taxa across environments and lineages. Our model highlights the potential of transformer architectures for genome-based growth rate prediction and provides a powerful new tool for identifying microbes with high or low cultivation potential in metagenomic datasets.

Introduction

Microbial growth rate is a fundamental physiological trait that shapes microbial life strategies, ecological roles, and evolutionary dynamics. Minimum doubling times which serves as a proxy for maximum growth rates vary widely—from less than 10 minutes in some laboratory-cultured species (Eagon 1962) to several days for oligotrophic marine microbes (Azam and Malfatti 2007; Imachi et al. 2020, 2025; Jogler, Glöckner, and Kolter 2011; Rappé et al. 2002). Even under optimal conditions, microbes exhibit striking differences in growth potential, largely driven by genome-encoded variation in DNA replication and biosynthetic capacities.

Despite this diversity, an estimated 81% of Earth's microbes remain uncultured (Lloyd et al. 2018), limiting direct measurement of their physiological traits. Cultivation-based methods are often biased against slow-growing species, which are important because they have unique metabolisms which contribute to overall ecosystem function and are often miss out due to methodological constraints (Alain and Querellou 2009; Prakash et al. 2013). As a result, genome-based prediction of microbial growth potential has emerged as a key objective in microbial ecology and genomics.

Most existing computational methods estimate microbial growth rates using codon usage bias (CUB) in ribosomal and other highly expressed genes (Vieira-Silva and Rocha 2010; Weissman et al., 2021). CUB correlates with translational efficiency, which in turn is associated with growth potential (Frumkin et al. 2018; Long et al. 2021; Sharp et al. 2005). Tools like *gRodon* leverage these associations to estimate minimum doubling times. However, these approaches often underperform for slow-growing organisms,

whose genomes may lack strong or canonical CUB patterns (Weissman et al., 2021). Moreover, environmental factors such as temperature can influence growth rate independently of genomic features, complicating predictions for extremophiles and other niche-adapted microbes (Birgander et al. 2013; Price and Sowers 2004).

Deep learning methods offer a promising alternative to CUB-based predictors. Biological language models (LLMs), particularly those using transformer architectures, have greatly expanded our ability to capture more complex genomic signatures from large-scale nucleotide and protein sequences (Babjac et al., 2023; Madani et al. 2023; Outeiral and Deane 2024; Rives et al. 2021; Schmirler, Heinzinger, and Rost 2024). Transformers can also model long-range dependencies in sequence data, making them well-suited to uncover complex biological patterns beyond local codon usage (Dotan et al. 2024; Settles and Craven 2008). One such model, *LookingGlass,* was pretrained on diverse prokaryotic genomes and has shown promising results on downstream tasks such as enzyme function prediction and optimal growth temperature inference (Hoarfrost et al. 2022).

To our knowledge, transformer-based models have not yet been applied to predict microbial growth rates. In this study, we fine-tuned *LookingGlass* using multiple genomic training inputs, including the same phenotypic trait dataset (Madin et al. 2020) previously used by *gRodon* (Weissman et al., 2021), enabling direct performance comparisons. We then apply our model to predict doubling times for 25,103 metagenome-assembled genomes (MAGs) from the “Genomes from Earth’s Microbiomes” (GEM) dataset (Nayfach et al. 2021). Our findings provide new insights

into the growth potential of uncultured microbes and demonstrate how transformer-based models can advance genome-based growth rate prediction. As an example, this approach could help identify slow-growing yet ecologically critical organisms from current and future metagenomic datasets—informing cultivation strategies and ecosystem modeling efforts.

Results

LookingGlass models indicate that growth signals are stronger in ribosomal related genes

To evaluate which functionally annotated genomic sequences are most informative for predicting microbial growth rates, we fine-tuned *LookingGlass* v2 on various subsets of coding sequences categorized by functional COG (Clusters of Orthologous Groups) annotations which include metabolism-related genes (energy production [C], amino acid metabolism [E], nucleotide metabolism [F], carbohydrate metabolism [G], coenzyme metabolism [H], lipid metabolism [I], inorganic ion metabolism [P], secondary metabolite biosynthesis [Q]), cellular processing and signaling genes (cell cycle control [D], signal transduction [T], intracellular trafficking [U], defense mechanisms [V], membrane biogenesis [M], cell motility [N], protein modification and chaperones [O]), and information storage and processing genes (translation and ribosome structure [J], transcription [K], replication and repair [L]). Among all tested subsets, the model trained exclusively on ribosomal protein sequences performed best, achieving the lowest root mean square error (RMSE = 1.013) and mean absolute error (MAE = 0.764), indicating the most accurate predictions of minimum doubling time (**Table 4.1**). This

model also yielded the highest predictive power, with a coefficient of determination (R^2) of 0.560 (**Table 4.1**).

Models trained on coding sequences associated with information storage and processing functions also performed relatively well (RMSE = 1.269; MAE = 0.994), though they were less accurate than the ribosomal subset (**Table 4.1**). By contrast, the model trained on the full set of coding sequences performed the worst across all metrics, with the highest RMSE (1.336), highest MAE (1.082), and a negative R^2 value (–0.027) (**Table 4.1**). A negative R^2 indicates that the model performs worse than simply predicting the mean doubling time for all samples. These findings are consistent with previous studies linking ribosomal protein gene composition and codon usage to translational efficiency and growth potential.

LookingGlass model improves minimum doubling time prediction

We evaluated the predictive performance of the *LookingGlass* model fine-tuned on ribosomal sequences (LGM_rib) against two versions of the gRodon model using a held-out test dataset (see Methods). Pearson correlation analysis between predicted and observed minimum doubling times (**Figure 4.1**) show that LGM_rib achieved the strongest correlation ($R = 0.75$, $p < 0.005$), outperforming both the original *gRodon* model ($R = 0.40$, $p < 0.005$) and the temperature-corrected version (gRodon_tmp; $R = 0.60$, $p < 0.005$). A correlation of 0.75 indicates a strong linear relationship between predicted and observed doubling times. In addition, *LGM_rib* has the lowest root mean square error (RMSE=1.013), indicating the highest overall accuracy among the models tested.

To further assess model performance across different microbial growth categories, we divided the dataset into fast and slow growers using a 5-hour minimum doubling time threshold, consistent with the default setting in *gRodon*. For fast-growing organisms, *gRodon_tmp* performed slightly better (RMSE = 0.906) than both LGM_rib (RMSE = 0.974) and gRodon (RMSE = 1.197). However, for slow growers, LGM_rib (RMSE=1.082) substantially outperformed both *gRodon_tmp* (1.637) and *gRodon* (1.751) (**Figure 4.1**).

LookingGlass model embeddings differentiate fast growers from slow growers

We assessed how different minimum doubling time thresholds affect the classification of microbes as fast or slow growers using LookingGlass-derived sequence embeddings. Specifically, we compared three thresholds: 5 hours (the default used by gRodon), 2.6 hours (the empirical median of the training set), and 8 hours (the 75th percentile of the doubling time distribution).

Among these thresholds, the 5-hour cutoff provided the best overall classification performance, yielding the highest area under the receiver operating characteristic curve (AUC-ROC = 0.87), compared to 2.6-hours (AUC-ROC = 0.78) and 8-hours (AUC-ROC = 0.81). At this threshold, AUC and precision-recall metrics plateaued, indicating that 5 hours serves as a robust dividing line between fast and slow growers (**Supplemental Figure 4.1**).

As expected, our model was more effective at identifying fast growers than slow growers (**Figure 4.2**), suggesting class imbalance or weaker genomic signals in slow-growing taxa. Embeddings generated using the 5-hour threshold exhibited clearer

clustering by growth category in t-SNE plots (**Figure 4.2E**), demonstrating that this threshold may reflect a meaningful biological division, e.g., between organisms adapted for rapid resource exploitation versus those with slower life histories.

LookingGlass model reveals culturable fast and slow growers across earth environments

We applied the fine-tuned LookingGlass model to predict minimum doubling times for 25,103 microbial genomes from the "Genome from Earth's Microbiomes" (GEM) dataset, focusing on genomes containing at least 300 ribosomal coding sequences. Across most ecosystems, cultured microbes exhibited a narrower distribution of predicted doubling times, skewed toward faster growth (<5 hours), compared to uncultured microbes (**Figure 4.3**; **Supplemental Figure 4.2**). Environments such as city subways, ruminant hosts, plant-associated habitats, lab enrichments, and algae-associated systems were particularly enriched in cultured microbes with short predicted doubling times (Fig. 3a), suggesting a cultivation bias toward fast growers in these settings.

An exception was observed in human-associated environments, where cultured microbes spanned a broader range of predicted doubling times (**Figure 4.3A**; **Supplemental Figure 4.2A**), indicating more inclusive cultivation success across growth rates in anthropogenic environments.

These environment-specific trends are mirrored at the phylum level, where growth potential appears to influence culturability. *Proteobacteria*, *Firmicutes*, and *Deinococcota*—phyla with generally short doubling times—are largely represented by cultured genomes. In contrast, within *Cyanobacteria,* only fast-growing representatives

appear to be cultured, while slow-growing members remain uncultured. Some phyla, including *Bacteroidota*, *Fusobacteriota*, *Actinobacteriota*, and *Campylobacteriota*, contain both fast- and slow-growing cultured strains (**Figure 4.3B**). While predicted doubling times provide a useful proxy, they may not fully capture the influence of growth-limiting environmental conditions.

Discussion

To our knowledge, this is the first application of a transformer-based language model for predicting microbial growth rates, providing improved resolution for slow-growing taxa. The model performed best when trained on ribosomal protein-coding sequences (**Figure 4.1**), consistent with prior work showing that codon usage bias (CUB) in ribosomal and other highly expressed genes serves as a reliable proxy for growth rate (Vieira-Silva and Rocha 2010; Weissman et al., 2021). CUB in these genes is strongly associated with translational efficiency and rapid growth (Frumkin et al. 2018; Shah and Gilchrist 2011). Moreover, the conservation and ubiquity of ribosomal genes across microbial taxa (Lim, Furuta, and Kobayashi 2012; Santamaria et al. 2012) make them well-suited for sequence-based prediction across isolates and metagenomes. Compared to gRodon, a widely used tool for predicting minimum doubling time, our ribosomal sequence–trained model (*LGM_rib*) demonstrated substantially higher prediction accuracy—especially for slow-growing organisms (**Figure 4.2**) — and required no auxiliary metadata such as optimal growth temperature. This streamlines prediction workflows and enables broader application in contexts where environmental metadata are missing or uncertain.

However, like other sequence-based models, LGM_rib showed reduced accuracy for extremely slow-growing organisms (**Figure 4.1**). This limitation is likely due to evolutionary factors—such as weak selection and increased genetic drift—that obscure codon optimization signals in these taxa (Dethlefsen and Schmidt 2007; Lynch 2010). Nonetheless, the model's relative performance for slow growers marks a significant improvement over prior approaches, offering a new avenue for identifying ecologically important microbes that have been historically overlooked due to cultivation bias.
Our findings also support prior observations that codon optimization signals plateau at minimum doubling times >5 hours (Weissman et al. 2021). Classification based on this 5-hour threshold yielded the best performance (**Figure 4.2**), outperforming thresholds based on the empirical median (2.6 hours) or upper quartile (8 hours). This reflects distinct codon usage patterns that correlate with microbial life-history strategies (Kirchman 2016; Vieira-Silva and Rocha 2010).

When applied to over 25,000 metagenome-assembled genomes (MAGs) from the "Genomes from Earth's Microbiomes" (GEM) dataset, LGM_rib revealed striking differences in predicted doubling times between cultured and uncultured microbes across diverse environments (**Figure 4.3**). Cultured microbes were typically predicted to grow faster, particularly in nutrient-rich or anthropogenic environments such as subways, laboratory enrichments, plant rhizospheres, and ruminant digestive systems. These trends mirror longstanding cultivation biases that favor copiotrophic taxa, i.e., microbes that thrive in environments rich in nutrients (Couso et al. 2023; J. Li et al. 2019; Steen et al. 2019). Human-associated microbes were an exception, with successful cultivation

spanning a broad range of growth rates—likely reflecting focused efforts to isolate pathogenic and commensal taxa (Lagier et al. 2016) .

At the phylum level, we observed that culturability was often associated with predicted growth rate. *Proteobacteria*, *Firmicutes*, and *Deinococcota* were dominated by fast-growing, cultured representatives. In *Cyanobacteria*, only fast-growing lineages had cultured representatives, likely due to the complexity of phototrophic growth and the longer division cycles of slower-growing strains (Reimers et al. 2017). In contrast, phyla such as *Bacteroidota*, *Fusobacteriota*, *Firmicutes*, *Actinobacteriota*, and *Campylobacteriota* included both fast- and slow-growing cultured strains, reflecting their ecological and clinical significance (Bereded et al. 2022; Dong et al. 2018).

A key limitation of our approach is the scarcity of microbial genomes with experimentally verified minimum doubling times, especially among slow growers. This restricts training diversity and model generalizability. However, ongoing efforts to isolate previously uncultured microbes (Cross et al. 2019; Lewis et al. 2021; Mu et al. 2021; Pulschen et al. 2017), along with the expansion of phenotypic trait databases such as those compiled by Madin et al. (2020), are expected to alleviate this bottleneck.

In conclusion, our results demonstrate that transformer-based models offer a scalable and metadata-light solution for predicting microbial growth rates directly from genomic sequence data. We anticipate that deep learning will play a central role in linking microbial genotypes to ecological phenotypes, aiding both cultivation efforts and ecosystem modeling. Future work could incorporate attention-based interpretability methods to identify specific genomic features contributing to predicted growth rates.

Methods

Isolates trait dataset preparation

We utilized the Madin et al. dataset, which consists of 581 microbial isolates with reported minimum doubling times. Previous studies have shown that growth rates exhibit minimal phylogenetic conservatism, with phylogeny predicting only 38% of growth rate variation (Walkup et al. 2023). Furthermore, Xu et al. 2024 reported that this dataset has a low phylogenetic signal (Pagel's $\lambda = 0.106$, $K = 0.137$) for minimum doubling time, indicating that it is relatively balanced with respect to phylogeny (Xu et al., 2024). To further minimize phylogenetic biases, we down sampled the dataset by including only one genome per species, resulting in a total dataset to 475 genomes with reported minimum doubling times. Due to small sample size of the dataset and to further minimize overlap across taxonomy, we split the data into 70% training and 15% validation using stratified sampling at the class level. The 15% of the remaining dataset serves as test dataset which comprises of taxonomic classes with 5 or fewer genomes that are not present in the training and validation datasets.

COGs and Ribosomal sequences annotation

We annotated the isolate genomes using Prokka (Seemann 2014) and extracted the genes relating to each COG categories. Prokka primarily conducts blastp and hmmsearch against NCBI COGs reference database to annotate genes with annotations further subjected to manual curation. Genes annotated with COG functions such as replication, recombination and repair (L), translation ribosomal structure and biogenesis (J) and transcription (K) were concatenated as information storage and processing

dataset. COG functions related to cell cycle and cell division (D), signal transduction mechanism (T), defense mechanism (V), cell motility (N), post translational modification (O), intracellular trafficking and secretion (U), cell wall membrane biogenesis (M) was concatenated as cellular processing and signaling sequences. COG functions related energy production and conversion (C), amino acid (E), carbohydrate (G), lipids (I), co-enzymes (H), secondary metabolites metabolisms and transports (Q) were concatenated as metabolism sequences. Sequences related to ribosomal genes were annotated with barranap (https://github.com/tseemann/barrnap) and were also concatenated together.

GEM dataset preparation

The GEM data set contains 52515 MAGs of medium and high-quality MAGS assembled from 10450 metagenomes from various environment (Nayfach et al. 2021). The GEM metadata includes the taxonomic and environmental information. The MAGs were inferred to be cultured and uncultured by taxonomically searching each MAGs genomic sequence against the RefSeq of the corresponding GTDB release (v89) and aligning the sequences against the NCBI RefSeq using RefSeeker (Schwengers et al. 2019). MAGs were classified as cultured if they taxonomically matched that of the RefSeq genome or aligned with at least 95% ANI to a RefSeq isolates. Ribosomal sequences were extracted from the MAGs using barranap and MAGs with short ribosomal sequences (< 300 nts) were removed from the dataset.

gRodon prediction of minimum doubling time

We estimated the minimum doubling time of the test data using the gRodon package in R. gRodon also uses ribosomal proteins in the genomes for doubling time

prediction. Madin et al dataset was used as the training dataset for these predictions. Additionally, we included the optimal growth temperature as a variable in the temperature corrected version.

Description of Lookingglass model

We used the LookingGlass-2 biological foundation model to fine tune our growth rate prediction models. LookingGlass-2 is an update to LookingGlass v1 (Hoarfrost et al. 2022), and is a larger transformer based biological foundational model pretrained on a larger corpus of read-length DNA sequences selected uniformly from taxonomic classes distributed across the prokaryotic tree of life. The current LookingGlass-2 model is adapted from RoBERTa (Liu et al. 2019), a bidirectional transformer-based masked language model, to learn fundamental biological features from DNA sequences. It uses 24 layers, 16 attention heads, an intermediate size of 4096, and a hidden size of 1024, with a custom DNA nucleotide tokenizer resulting in a model with 86,050,569 parameters total. LookingGlass-2 was trained on >11M read-length sequences (average sequence length 128 nucleotides, total training set ~1B tokens) derived from representative genomes sampled uniformly from each taxonomic order in the GTDB taxonomy (Parks et al. 2018). LookingGlass-2 was pretrained in a self-supervised manner on this dataset with a masked language modeling task, where 15% of nucleotides on an input DNA sequence are masked at random and predicted from surrounding context.

Lookingglass Model-2 Fine-tuning Procedure

We fine-tuned the LookingGlass model to predict the minimum doubling time using a supervised regression approach. For the LGM_rib model, the input were the

nucleotide sequences for all the ribosomal genes, which were tokenized using the LookingGlass custom tokenizer and split into several chunks of maximum length of 128 nucleotides. The training data consisted of over 2 million ribosomal sequences in total with their corresponding experimentally determined minimum doubling time labels. The model was finetuned on the training data using the Hugging Face Trainer API with default Adam optimizer. The model was trained for 8 epochs with a batch size of 8 and selected hyperparameters of weight decay 0.01, base learning rate of $2e^{-5}$ and 500 warmup steps. We extracted the embeddings from the final transformer layer and used them for the regression task. To incorporate the growth temperature in the prediction, the embeddings generated by the final layer were concatenated with corresponding growth temperature values. During training, the model performance of the validation set was evaluated for each epoch based on the Root Mean Squared Error (RMSE) to prevent overfitting while the test set was used for the final model evaluation After finetuning, the best-performing model was selected based on the lowest RMSE. The predicted values from the test set are compared to actual values for model evaluation and also compared to that of the gRodon prediction. The finetuned and evaluation procedures were also repeated for other gene sets.

Assessing model performance

Model performances were computed for the validation and test sets and evaluated using the standard regression metrics which include the RMSE, Mean Absolute Error (MAE), R2 and Pearson correlation. gRodon predictions were evaluated using the same metrics.

Classification into fast and slow growers

Microbial isolates were classified into fast and slow growers based on the predicted doubling time threshold of 5 hrs. The 5hrs cutoff classification was compared to using the median (2.6hrs) as threshold for classification. The embeddings of the final transformer layer from the finetuned LGM_rib were extracted and together with the predicted minimal doubling times were used as feature representations for fitting a logistic regression.

Assessing the classification performance

The classification performance was evaluated using the standard binary classification metrics which include the AUC-ROC which measures the model's ability to distinguish between fast and slow growers, Precision-Recall AUC (PR-AUC) computed separately for fast and slow growers, as well as precision, recall and F1-score

Data availability

The Madin dataset is available at the original source (https://figshare.com/collections/A_synthesis_of_bacterial_and_archaeal_phenotypic_trait_data/4843290/3)

Code availability

Codes and data are available at https://github.com/oduwoleiyanu/microbial-growth-rate-transformer

References

Alain, Karine, And Joël Querellou. 2009. “Cultivating The Uncultured: Limits, Advances And Future Challenges.” *Extremophiles* 13(4): 583–94. Doi:10.1007/S00792-009-0261-3.

Azam, Farooq, And Francesca Malfatti. 2007. “Microbial Structuring Of Marine Ecosystems.” *Nature Reviews Microbiology* 5(10): 782–91.

Babjac, Ashley Nicole, Zhixiu Lu, And Scott J Emrich. 2023. “Codonbert: Using BERT For Sentiment Analysis To Better Predict Genes With Low Expression.” In *Proceedings Of The 14th ACM International Conference On Bioinformatics, Computational Biology, And Health Informatics*, Houston TX USA: ACM, 1–6. Doi:10.1145/3584371.3613013.

Bereded, Negash Kabtimer, Getachew Beneberu Abebe, Solomon Workneh Fanta, Manuel Curto, Herwig Waidbacher, Harald Meimberg, And Konrad J. Domig. 2022. “The Gut Bacterial Microbiome Of Nile Tilapia (Oreochromis Niloticus) From Lakes Across An Altitudinal Gradient.” *BMC Microbiology* 22(1): 87. Doi:10.1186/S12866-022-02496-Z.

Birgander, Johanna, Stephanie Reischke, Davey L. Jones, And Johannes Rousk. 2013. “Temperature Adaptation Of Bacterial Growth And 14C-Glucose Mineralisation In A Laboratory Study.” *Soil Biology And Biochemistry* 65: 294–303.

Couso, Luciana L., Alfonso Soler-Bistué, Ariel A. Aptekmann, And Ignacio E. Sánchez. 2023. “Ecology Theory Disentangles Microbial Dichotomies.” *Environmental Microbiology*.

Cross, Karissa L., James H. Campbell, Manasi Balachandran, Alisha G. Campbell, Connor J. Cooper, Ann Griffen, Matthew Heaton, Et Al. 2019. “Targeted Isolation And Cultivation Of Uncultivated Bacteria By Reverse Genomics.” *Nature Biotechnology* 37(11): 1314–21.

Dong, Li, Jian Yin, Jing Zhao, Shan-Rui Ma, Hai-Rui Wang, Meng Wang, Wen Chen, And Wen-Qiang Wei. 2018. “Microbial Similarity And Preference For Specific Sites In Healthy Oral Cavity And Esophagus.” *Frontiers In Microbiology* 9: 1603.

Dotan, Edo, Gal Jaschek, Tal Pupko, And Yonatan Belinkov. 2024. “Effect Of Tokenization On Transformers For Biological Sequences.” *Bioinformatics* 40(4): Btae196.

Eagon, R. G. 1962. “*PSEUDOMONAS NATRIEGENS* , A MARINE BACTERIUM WITH A GENERATION TIME OF LESS THAN 10 MINUTES.” *Journal Of Bacteriology* 83(4): 736–37. Doi:10.1128/Jb.83.4.736-737.1962.

Frumkin, Idan, Marc J. Lajoie, Christopher J. Gregg, Gil Hornung, George M. Church, And Yitzhak Pilpel. 2018. “Codon Usage Of Highly Expressed Genes Affects Proteome-Wide Translation Efficiency.” *Proceedings Of The National Academy Of Sciences* 115(21). Doi:10.1073/Pnas.1719375115.

Gouy, Manolo, And Christian Gautier. 1982. “Codon Usage In Bacteria: Correlation With Gene Expressivity.” *Nucleic Acids Research* 10(22): 7055–74.

Ho, Adrian, D. Paolo Di Lonardo, And Paul LE Bodelier. 2017. “Revisiting Life Strategy Concepts In Environmental Microbial Ecology.” *FEMS Microbiology Ecology* 93(3): Fix006.

Hoarfrost, A., A. Aptekmann, G. Farfañuk, And Y. Bromberg. 2022. “Deep Learning Of A Bacterial And Archaeal Universal Language Of Life Enables Transfer Learning And Illuminates Microbial Dark Matter.” *Nature Communications* 13(1): 1–12.

Hug, Laura A. 2018. “Sizing Up The Uncultured Microbial Majority.” *Msystems* 3(5): E00185-18.

Imachi, Hiroyuki, Masaru K. Nobu, Shun’ichi Ishii, Yuga Hirakata, Tetsuro Ikuta, Yuta Isaji, Makoto Miyata, Et Al. 2025. “Eukaryotes’ Closest Relatives Are Internally Simple Syntrophic Archaea.” *Biorxiv*: 2025–02.

Imachi, Hiroyuki, Masaru K. Nobu, Nozomi Nakahara, Yuki Morono, Miyuki Ogawara, Yoshihiro Takaki, Yoshinori Takano, Et Al. 2020. “Isolation Of An Archaeon At The Prokaryote–Eukaryote Interface.” *Nature* 577(7791): 519–25.

Jogler, Christian, Frank Oliver Glöckner, And Roberto Kolter. 2011. “Characterization Of Planctomyces Limnophilus And Development Of Genetic Tools For Its Manipulation Establish It As A Model Species For The Phylum Planctomycetes.” *Applied And Environmental Microbiology* 77(16): 5826–29. Doi:10.1128/AEM.05132-11.

Kirchman, David L. 2016. “Growth Rates Of Microbes In The Oceans.” *Annual Review Of Marine Science* 8(1): 285–309. Doi:10.1146/Annurev-Marine-122414-033938.

Lagier, Jean-Christophe, Saber Khelaifia, Maryam Tidjani Alou, Sokhna Ndongo, Niokhor Dione, Perrine Hugon, Aurelia Caputo, Et Al. 2016. “Culture Of Previously Uncultured Members Of The Human Gut Microbiota By

Culturomics." *Nature Microbiology* 1(12). Https://Www.Nature.Com/Articles/Nmicrobiol2016203 (April 13, 2025).

Lewis, William H., Guillaume Tahon, Patricia Geesink, Diana Z. Sousa, And Thijs JG Ettema. 2021. "Innovations To Culturing The Uncultured Microbial Majority." *Nature Reviews Microbiology* 19(4): 225–40.

Li, Junhui, Rebecca L. Mau, Paul Dijkstra, Benjamin J. Koch, Egbert Schwartz, Xiao-Jun Allen Liu, Ember M. Morrissey, Et Al. 2019. "Predictive Genomic Traits For Bacterial Growth In Culture Versus Actual Growth In Soil." *The ISME Journal* 13(9): 2162–72.

Lim, Kyungtaek, Yoshikazu Furuta, And Ichizo Kobayashi. 2012. "Large Variations In Bacterial Ribosomal RNA Genes." *Molecular Biology And Evolution* 29(10): 2937–48.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Et Al. 2019. "Roberta: A Robustly Optimized BERT Pretraining Approach." Doi:10.48550/Arxiv.1907.11692.

Lloyd, Karen G., Andrew D. Steen, Joshua Ladau, Junqi Yin, And Lonnie Crosby. 2018. "Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes." *Msystems* 3(5): E00055-18.

Long, Andrew M., Shengwei Hou, J. Cesar Ignacio-Espinoza, And Jed A. Fuhrman. 2021. "Benchmarking Microbial Growth Rate Predictions From Metagenomes." *The ISME Journal* 15(1): 183–95.

Madani, Ali, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos Jr, Et Al. 2023. “Large Language Models Generate Functional Protein Sequences Across Diverse Families.” *Nature Biotechnology* 41(8): 1099–1106.

Madin, Joshua S., Daniel A. Nielsen, Maria Brbic, Ross Corkrey, David Danko, Kyle Edwards, Martin KM Engqvist, Et Al. 2020. “A Synthesis Of Bacterial And Archaeal Phenotypic Trait Data.” *Scientific Data* 7(1): 170.

Mahowald, Michael A., Federico E. Rey, Henning Seedorf, Peter J. Turnbaugh, Robert S. Fulton, Aye Wollam, Neha Shah, Et Al. 2009. “Characterizing A Model Human Gut Microbiota Composed Of Members Of Its Two Dominant Bacterial Phyla.” *Proceedings Of The National Academy Of Sciences* 106(14): 5859–64. Doi:10.1073/Pnas.0901529106.

Mu, Da-Shuai, Yang Ouyang, Guan-Jun Chen, And Zong-Jun Du. 2021. “Strategies For Culturing Active/Dormant Marine Microbes.” *Marine Life Science & Technology* 3(2): 121–31. Doi:10.1007/S42995-020-00053-Z.

Nayfach, Stephen, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, Et Al. 2021. “A Genomic Catalog Of Earth’s Microbiomes.” *Nature Biotechnology* 39(4): 499–509.

Outeiral, Carlos, And Charlotte M. Deane. 2024. “Codon Language Embeddings Provide Strong Signals For Use In Protein Engineering.” *Nature Machine Intelligence* 6(2): 170–79.

Parks, Donovan H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, And Philip Hugenholtz. 2018. “A Standardized Bacterial Taxonomy Based On Genome Phylogeny Substantially Revises The Tree Of Life.” *Nature Biotechnology* 36(10): 996–1004.

Prakash, Om, Yogesh Shouche, Kamlesh Jangid, And Joel E. Kostka. 2013. “Microbial Cultivation And The Role Of Microbial Resource Centers In The Omics Era.” *Applied Microbiology And Biotechnology* 97: 51–62.

Price, P. Buford, And Todd Sowers. 2004. “Temperature Dependence Of Metabolic Rates For Microbial Growth, Maintenance, And Survival.” *Proceedings Of The National Academy Of Sciences* 101(13): 4631–36. Doi:10.1073/Pnas.0400522101.

Pulschen, Andre A., Amanda G. Bendia, Ashwana D. Fricker, Vivian H. Pellizari, Douglas Galante, And Fabio Rodrigues. 2017. “Isolation Of Uncultured Bacteria From Antarctica Using Long Incubation Periods And Low Nutritional Media.” *Frontiers In Microbiology* 8: 1346.

Rappé, Michael S., Stephanie A. Connon, Kevin L. Vergin, And Stephen J. Giovannoni. 2002. “Cultivation Of The Ubiquitous SAR11 Marine Bacterioplankton Clade.” *Nature* 418(6898): 630–33.

Reimers, Alexandra-M., Henning Knoop, Alexander Bockmayr, And Ralf Steuer. 2017. “Cellular Trade-Offs And Optimal Resource Allocation During Cyanobacterial Diurnal Growth.” *Proceedings Of The National Academy Of Sciences* 114(31). Doi:10.1073/Pnas.1617508114.

Rives, Alexander, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Et Al. 2021. “Biological Structure And Function Emerge From Scaling Unsupervised Learning To 250 Million Protein Sequences.” *Proceedings Of The National Academy Of Sciences* 118(15): E2016239118. Doi:10.1073/Pnas.2016239118.

Santamaria, Monica, Bruno Fosso, Arianna Consiglio, Giorgio De Caro, Giorgio Grillo, Flavio Licciulli, Sabino Liuni, Et Al. 2012. “Reference Databases For Taxonomic Assignment In Metagenomics.” *Briefings In Bioinformatics* 13(6): 682–95.

Schmirler, Robert, Michael Heinzinger, And Burkhard Rost. 2024. “Fine-Tuning Protein Language Models Boosts Predictions Across Diverse Tasks.” *Nature Communications* 15(1): 7407.

Schwengers, Oliver, Torsten Hain, Trinad Chakraborty, And Alexander Goesmann. 2019. “Referenceseeker: Rapid Determination Of Appropriate Reference Genomes.” *Biorxiv*: 863621.

Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30(14): 2068–69.

Settles, Burr, And Mark Craven. 2008. “An Analysis Of Active Learning Strategies For Sequence Labeling Tasks.” In *Proceedings Of The 2008 Conference On Empirical Methods In Natural Language Processing*, , 1070–79. Https://Aclanthology.Org/D08-1112.Pdf (April 18, 2025).

Shah, Premal, And Michael A. Gilchrist. 2011. “Explaining Complex Codon Usage Patterns With Selection For Translational Efficiency, Mutation Bias, And Genetic

Drift." *Proceedings Of The National Academy Of Sciences* 108(25): 10231–36. Doi:10.1073/Pnas.1016719108.

Sharp, Paul M., Elizabeth Bailes, Russell J. Grocock, John F. Peden, And R. Elizabeth Sockett. 2005. "Variation In The Strength Of Selected Codon Usage Bias Among Bacteria." *Nucleic Acids Research* 33(4): 1141–53.

Steen, Andrew D., Alexander Crits-Christoph, Paul Carini, Kristen M. Deangelis, Noah Fierer, Karen G. Lloyd, And J. Cameron Thrash. 2019. "High Proportions Of Bacteria And Archaea Across Most Biomes Remain Uncultured." *The ISME Journal* 13(12): 3126–30.

Verma, Digvijay, Pankaj Kumar Garg, And Ashok Kumar Dubey. 2018. "Insights Into The Human Oral Microbiome." *Archives Of Microbiology* 200(4): 525–40. Doi:10.1007/S00203-018-1505-3.

Vieira-Silva, Sara, And Eduardo PC Rocha. 2010. "The Systemic Imprint Of Growth And Its Uses In Ecological (Meta) Genomics." *Plos Genetics* 6(1): E1000808.

Walkup, Jeth, Chansotheary Dang, Rebecca L. Mau, Michaela Hayer, Egbert Schwartz, Bram W. Stone, Kirsten S. Hofmockel, Et Al. 2023. "The Predictive Power Of Phylogeny On Growth Rates In Soil Bacterial Communities." *ISME Communications* 3(1): 73.

Weissman, J. L., Shengwei Hou, And Jed A. Fuhrman. 2021. "Estimating Maximal Microbial Growth Rates From Cultures, Metagenomes, And Single Cells Via Codon Usage Patterns." *Proceedings Of The National Academy Of Sciences* 118(12): E2016810118. Doi:10.1073/Pnas.2016810118.

Xu, Liang, Emily Zakem, And J. L. Weissman. 2024. “Improved Maximum Growth Rate Prediction From Microbial Genomes By Integrating Phylogenetic Information.” *Biorxiv*: 2024–10.

Yannai, Adi, Sophia Katz, And Ruth Hershberg. 2018. “The Codon Usage Of Lowly Expressed Genes Is Subject To Natural Selection.” *Genome Biology And Evolution* 10(5): 1237–46.

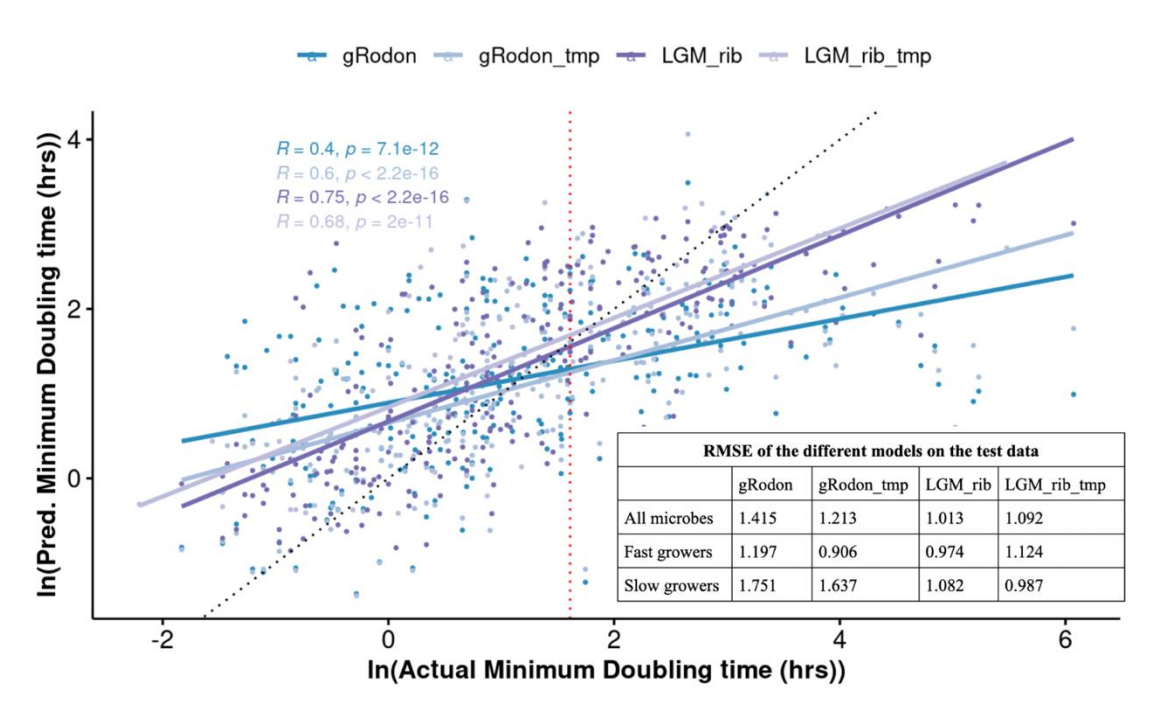
Appendix IV

Table

Table 4.1: Performance of LookingGlass-derived models finetuned on different annotated gene subsets

Trained category	Test loss	Test RMSE	Test R^2	Test MAE	Pearson's R
Coding (ORFs) sequences	1.786	1.336	-0.027	1.082	0.002
Metabolism related sequences	1.901	1.379	0.071	1.007	0.431
Cellular processing and signaling sequences	1.721	1.312	0.160	0.986	0.457
Information storage and processing sequences	1.610	1.269	0.214	0.994	0.533
Ribosomal sequences	0.993	**1.013**	**0.560**	0.764	0.748

Figures



RMSE of the different models on the test data

	gRodon	gRodon_tmp	LGM_rib	LGM_rib_tmp
All microbes	1.415	1.213	1.013	1.092
Fast growers	1.197	0.906	0.974	1.124
Slow growers	1.751	1.637	1.082	0.987

Figure 4.1: Comparison of predicted versus empirical minimum doubling times using three models: the fine-tuned LookingGlass model on ribosomal protein sequences (*LGM_rib*), the original *gRodon*, and *gRodon* with temperature correction (*gRodon_tmp*). *LGM_rib* achieved the lowest root mean square error (RMSE) and the highest correlation with observed doubling times. The black dashed line represents the identity line, indicating perfect agreement between predicted and actual values. All models tend to underestimate doubling times for slow-growing microbes, as shown by the clustering of points below the identity line at higher doubling times. The red dashed line marks the 5-hour threshold used by *gRodon* to distinguish fast- and slow-growing organisms.

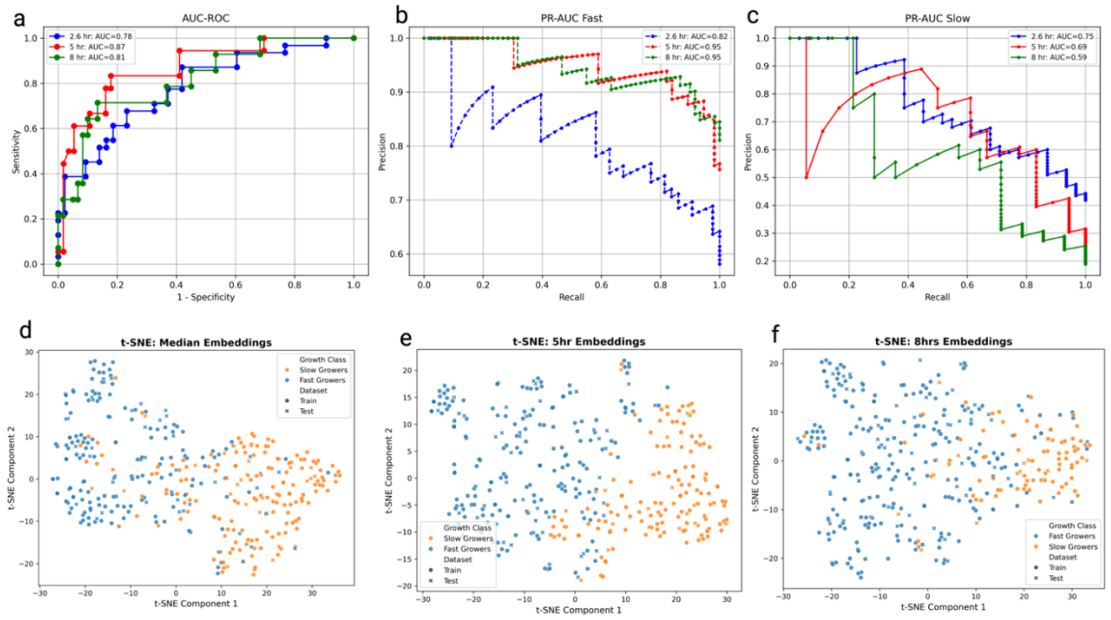


Figure 4.2: A 5-hour doubling time threshold improves microbial growth classification and cluster separation in LGM_rib embeddings. Classification performance using three doubling time thresholds (2.6, 5, and 8 hours): a) AUC_ROC b) Precision_Recall for identifying fast growers c) Precision_Recall for identifying slow growers. The LGM_rib model achieves the highest AUC-ROC at the 5-hour threshold. (d–f) t-SNE projections of LGM_rib embeddings show clustering of fast (blue) and slow (orange) growers under different classification thresholds: (b) 5 hours, (c) 2.6 hours (empirical median), and (d) 8 hours (75th percentile). The 5-hour threshold results in the clearest separation between fast and slow growers, with more distinct clustering of fast-growing microbes. Points from training and test datasets are marked separately.

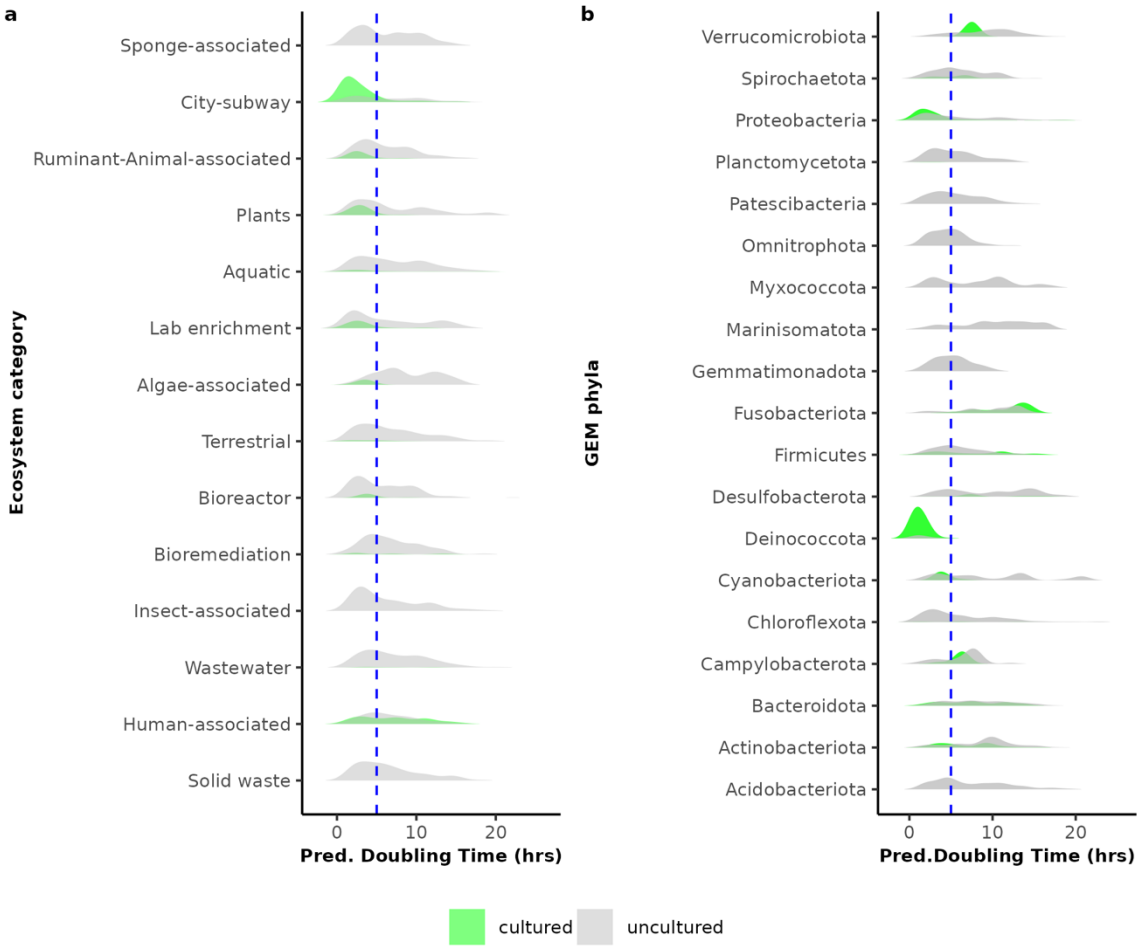
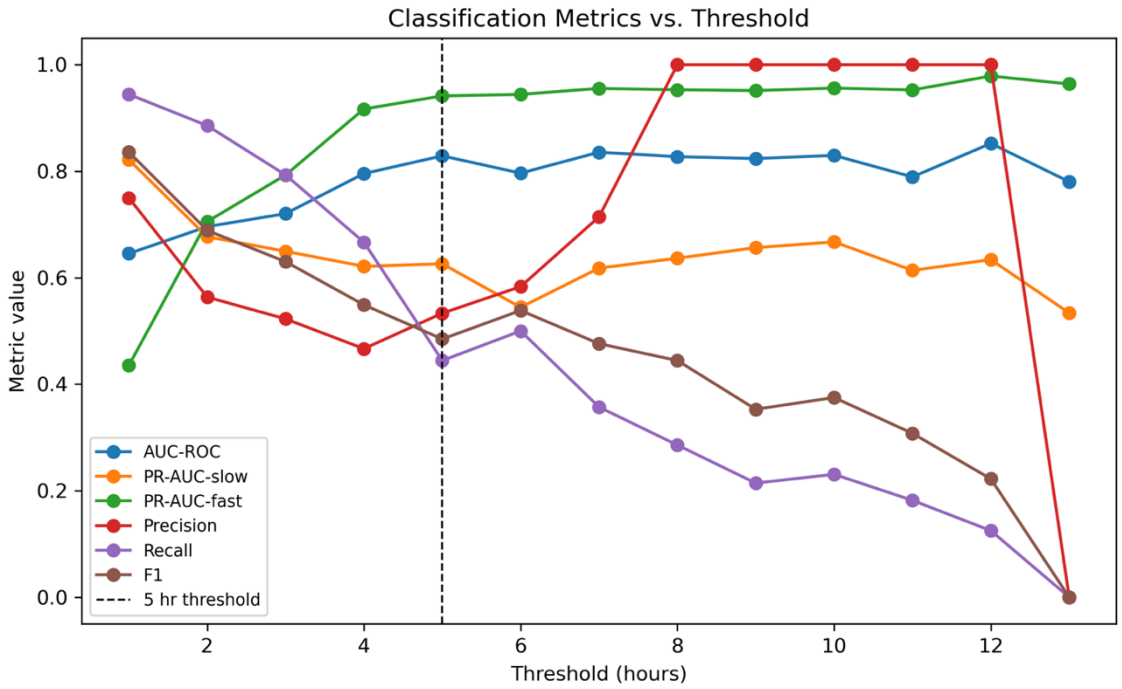
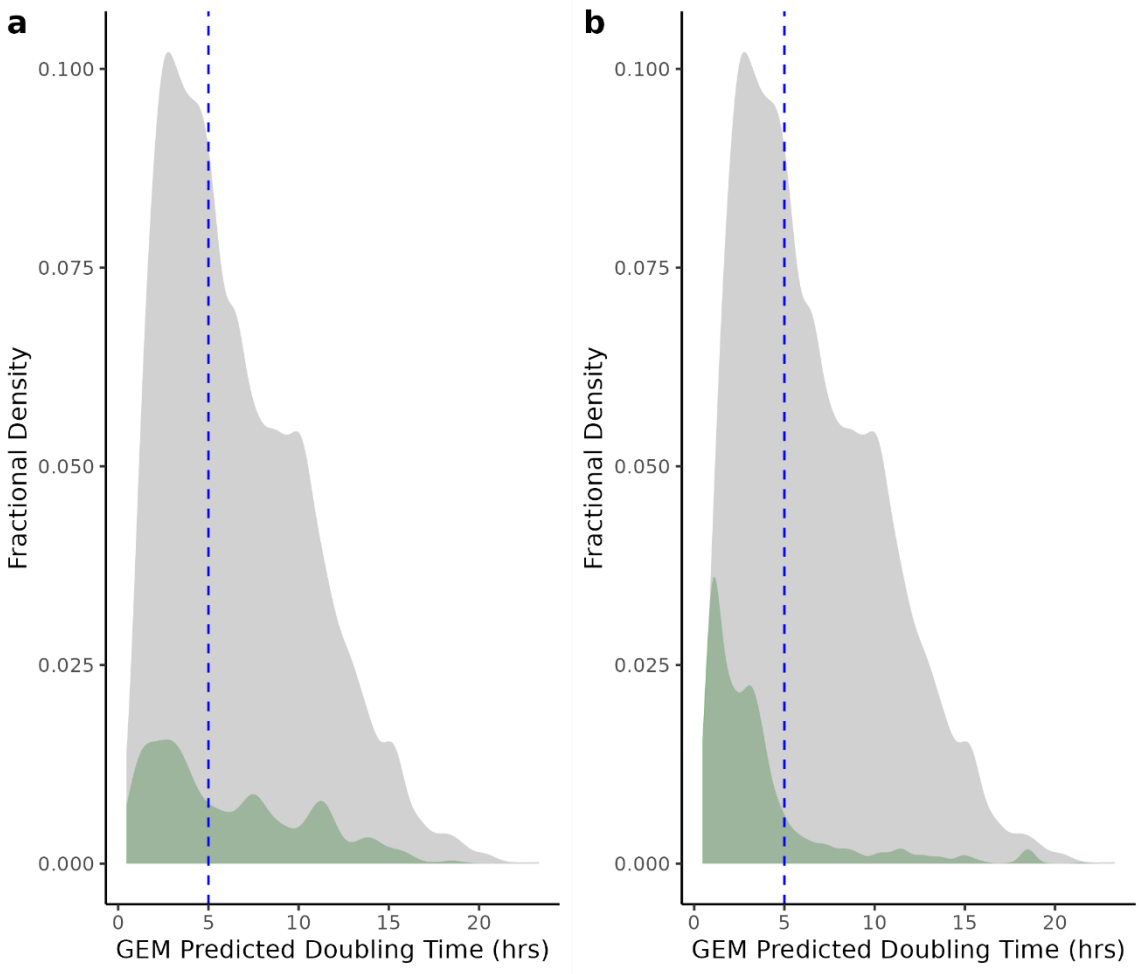


Figure 4.3: LGM_rib predicts minimum doubling time of microbes inferred to be cultured (green) and uncultured (gray) in GEM dataset across a) ecosystem and b) GEM phyla. The vertical dashed blue line represents a doubling time of 5 hours, distinguishing fast growers from slow growers. Only phyla with over 50 representatives and ecosystem with 75 representatives are shown in the plot.

Supplemental figures



Supplemental Figure 4.1: Classification metrics for differentiating fast from slow growers using different empirical minimum doubling time thresholds. The vertical dashed line at 5hrs indicates the gRodon threshold. The metrics were computed using logistic regression trained on sequence embeddings from the Lookingglass models. At 5hrs, the AUC and PR-AUC already reached the plateau indicating a pragmatic cutoff for fast and slow growers.



Supplemental Figure 4.2: Cultured microbes are biased towards fast growers except in human associated environments. Predicted minimum doubling time of a) all GEM microbes b) Excluding human related environments. The dashed line indicates 5hrs threshold which differentiates fast growers from the slow growers.

CHAPTER V

CONCLUSION

This dissertation demonstrates that bioinformatics and machine learning techniques effectively reveal genomic and phenotypic characteristics of uncultured microbes, providing critical insights beyond the scope of traditional culture-based methods. Through a combination of metagenomic assembled genome (MAG) analyses, computational predictions, and advanced machine learning models, it becomes clear that uncultured microbes represent distinct biological entities characterized by high genomic novelty and unique physiological traits such as slow growth and specific metabolic adaptations. The results further highlight the importance of utilizing genomic and computational approaches to uncover microbial adaptation and ecological functions in extreme and understudied environments, particularly permafrost. This work reveals the relationship between thermophilicity and microbial growth strategies in permafrost versus active layers, and shows thermophilic traits as an adaptation to slow growth in permafrost, highlighting distinct evolutionary trajectories driven by environmental constraints

Furthermore, computational predictions of microbial growth rate using transformer-based deep learning models, such as LookingGlass, outperform traditional methods like gRodon, illustrating the significant potential of machine learning to predict microbial growth rates accurately from genomic data. This model particularly shows enhanced

prediction of growth rate of slow growing microbes, a challenge for many existing growth rate prediction tools.

Future Direction

Future research could expand upon the approaches presented here by integrating larger and more diverse genomic datasets and novel codon features to improve predictive accuracy and generalizability of the deep learning models. Also, future work could incorporate attention -based interpretability methods to identify specific genomic features contributing to the predicted growth rates. I anticipate that applying deep learning models to microbial ecology will play a central role in linking microbial genotypes to ecological phenotypes, aiding both cultivation efforts and ecosystem modeling

Advances in single-cell genomics and co-culturing techniques may provide valuable complementary experimental validation for novel genomic factors linked to microbial culturability in this study. Additionally, future laboratory study investigating the ecological and evolutionary implications of thermophiles in cold environments could yield deeper insights into their niche adaptation strategies. Ultimately, combining multi-omics data with advanced machine learning and ecological modeling frameworks will continue to enhance our understanding of microbial diversity, ecology, and cultivability in Earth's most extreme and understudied environments.

VITA

Iyanu M. Oduwole was born in Ijebu Igbo, Ogun State, Nigeria. He earned his bachelor's degree from Olabisi Onabanjo University, Ago Iwoye, where he graduated as the top student in the Faculty of Basic Medical Sciences and the second overall best graduating student in the university. For his undergraduate research, he investigated the enzyme kinetics of alpha-amylases derived from *Aureobasidium pullulans* (black yeast). He went on to pursue a master's degree at the American University in Cairo, Egypt, where he graduated with a perfect GPA. During his time there, he worked on novel DNA ligases from the Red Sea brine pools. It was through this project that Iyanu developed a strong interest in bioinformatics, which led him to pursue a Ph.D. in the field.

Iyanu joined the Genome Science and Technology (GST) program at the Bredesen Center in January 2021 and began his doctoral research in Dr. Andrew Steen's laboratory, focusing on metagenomics, microbial ecology and metadata analysis. In 2023, he developed a keen interest in machine learning, attending his first summer school in the field and enrolling in related coursework. When Dr. Steen transitioned to the University of Southern California in 2024, Iyanu continued his dissertation research under the joint supervision of Dr. Andrew Steen and Dr. Scott Emrich. At Dr. Emrich's lab, he completed the final chapter of his dissertation, combining his expertise in bioinformatics and machine learning to investigate microbial ecology and growth prediction.