Transformer-based Prediction of Microbial Growth Rates from Genomic Data

Iyanu Oduwole
University of Tennessee
Knoxville, Tennessee, USA
ioduwole@vols.utk.edu

Mengling He
University of Tennessee
Knoxville, Tennessee, USA
mhe8@vols.utk.edu

Ratish Jha
University of Georgia
Athens, Georgia, USA

Adrienne Hoarfrost
University of Georgia
Athens, Georgia, USA

Andrew D. Steen*
University of Southern California
Los Angeles, California, USA
asteen@usc.edu

Scott Emrich*
University of Tennessee
Knoxville, Tennessee, USA
semrich@utk.edu

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. We further compare *LookingGlass2* to *Evo*, a general-purpose DNA language model with greater context length, and show that microbial-specific pretraining yields superior performance. 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.

CCS Concepts

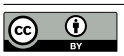
• **Applied computing** → **Computational genomics**; **Bioinformatics**; • **Computing methodologies** → **Information extraction**.

Keywords

deep learning, DNA-based transformer models, microbial growth rate prediction, codon usage bias

ACM Reference Format:
Iyanu Oduwole, Mengling He, Ratish Jha, Adrienne Hoarfrost, Andrew D. Steen, and Scott Emrich. 2025. Transformer-based Prediction of Microbial Growth Rates from Genomic Data. In *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '25), October 11–15, 2025, Philadelphia, PA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3765612.3767253

*Both are corresponding authors of this work



BCB '25, Philadelphia, PA, USA

ACM ISBN 979-8-4007-2200-4/2025/10
https://doi.org/10.1145/3765612.3767253

1 Introduction

Microbial growth rate is a key physiological trait that shapes microbial evolutionary dynamics and ecology [17]. Minimum doubling times vary dramatically—from under 10 minutes in lab-cultured species [9] to several days in oligotrophic marine microbes [2, 13, 14, 16, 37]—even under optimal conditions. This variation reflects underlying genomic differences in DNA replication and biosynthetic capacity.

In large part because of this genetic diversity, an estimated 81% of Earth's microbes remain uncultured [22], limiting our ability to directly measure their traits. Cultivation-based methods are often biased against slow-growing species due to methodological constraints [1, 48], even though these organisms play essential ecological roles. As a result, predicting microbial growth potential from genome sequences has become a central goal in microbial ecology and genomics [41].

Current computational approaches for estimating microbial growth rely primarily on codon usage bias (CUB), particularly in ribosomal and other highly expressed genes [49, 51]. Since CUB correlates with translational efficiency, it serves as a proxy for growth rate [10, 24, 46]. Tools like *gRodon* use these relationships to estimate minimum doubling times. However, these approaches often underperform for slow-growing microbes, which may lack canonical CUB signatures [51]. Moreover, non-genomic factors like temperature can affect growth rates independently, complicating predictions for extremophiles and other niche-adapted microbes [35].

These limitations have motivated the exploration of deep learning methods that can infer growth potential from broader genomic signals. Transformer-based biological language models (BLMs) have shown remarkable ability to capture complex, long-range patterns in nucleotide and protein sequences [3, 26, 31, 39, 42], which makes them well-suited to uncover potential biological signals beyond local codon usage [8]. One such model, *LookingGlass*, was pretrained on diverse prokaryotic genomes and has demonstrated strong performance in downstream tasks such as enzyme function prediction and growth temperature estimation [12, 15]. In parallel, general-purpose DNA language models like *Evo* have emerged, offering

large context windows but without specific adaptation to microbial genomic features.

To our knowledge, this is the first application of a transformer-based biological language model to estimate microbial growth potential from genome sequences. We fine-tune a completely reworked model focused on microbial genomes—*LookingGlass2*—using ribosomal protein-coding genes, other marker features, and the same phenotype dataset as *gRodon* [27], enabling direct performance comparison. We then apply our model to predict minimum doubling times for 25,103 metagenome-assembled genomes (MAGs) from the Genomes from Earth's Microbiomes (GEM) catalog [29]. Our results show that microbial-specific pretraining yields stronger predictive performance while providing new insights into the growth potential of uncultured microbes. Combined this is an substantially improved genome-based framework to study microbial ecology and cultivation strategies.

2 Methods

2.1 Isolates trait dataset preparation

We used the dataset compiled by Madin et al. [27], which includes 581 microbial isolates with empirically measured minimum doubling times under optimal growth conditions. Previous studies have shown that microbial growth rates exhibit low phylogenetic conservation; for example, phylogeny accounts for only 38% of growth rate variation [50]. Similarly, Xu et al. [52] found a weak phylogenetic signal in the data we use.

To further minimize phylogenetic redundancy and improve generalizability, we retained only one representative genome per species to avoid overrepresentation of well-studied species and to ensure each species contributes equally to model training, yielding a curated dataset of 475 genomes. To reduce taxonomic overlap across data splits, we employed stratified sampling by taxonomic class, following the approach of Hoarfrost et al. [12], to maintain taxonomic diversity across splits and prevent model overfitting to particular clades. We split the dataset into 70% for training, 15% for validation, and 15% for testing.

2.2 COGs and Ribosomal sequences annotation

We annotated the isolate genomes using *Prokka* [44], which performs *BLASTp* and *HMMER* searches against the NCBI COG (Clusters of Orthologous Groups) database [11] to assign functional categories to protein-coding genes. To ensure consistency, annotations were cross-validated against the original NCBI COG assignments, and any gene with inconsistent or ambiguous COG annotations was excluded from downstream analyses. 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." 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) were concatenated as "Cellular Processing and Signaling." COG functions related to 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." These annotations were used to fine tune LGM2, testing whether gene subsets from specific functional subsystems are more informative for predicting minimum doubling times.

Ribosomal protein-coding genes were annotated separately using *Barrnap* (https://github.com/tseemann/barrnap).

2.3 GEM dataset preparation

The GEM data set contains 52,515 medium- and high-quality metagenome-assembled genomes (MAGs) derived from 10,450 samples across a wide range of environments [29]. The accompanying metadata includes taxonomic and environmental annotations.

The NCBI RefSeq database is dominated by microbes that have been cultured. Using this observation, we performed two complementary analyses to distinguish between cultured and uncultured MAGs. First, we compared each MAG's taxonomic lineage to the NCBI RefSeq database entries using the Genome Taxonomy Database (GTDB) release v89 [32]. Second, we used RefSeeker [43] to align each MAG to RefSeq isolate genomes. A MAG was classified as cultured if it met either of the following criteria: (i) matched a taxonomic lineage corresponding to a cultured RefSeq genome, or (ii) aligned to a RefSeq isolate with at least 95% average nucleotide identity (ANI), which is the default *RefSeeker* cutoff of for a species-level match. MAGs not meeting either criterion were classified as uncultured.

Ribosomal protein-coding sequences were extracted from each MAG using *Barrnap*. To ensure high-quality input for downstream modeling, we excluded MAGs that contained few ribosomal protein-coding genes (<300 nucleotides after all genes are concatenated as input, which is higher than the 240bp or more default of gRodon). This length filter helped retain MAGs with more complete ribosomal gene sets, while excluding less complete assemblies. Of the 52,515 MAGs, 17,761 MAGs had no detectable ribosomal genes, 9,651 MAGs were excluded based on the <300 nucleotide total filter, resulting in a final dataset of 25,103 MAGs used for analysis.

2.4 gRodon estimation

We estimated the minimum doubling time for genomes in the test set using the *gRodon* R package [51]. *gRodon* predicts doubling time based on codon usage bias. We calibrated *gRodon* using the same ribosomal protein-coding training dataset used by Madin et al. [27], allowing for direct performance comparisons. We also ran the temperature-corrected version of gRodon by incorporating the optimal growth temperature of each organism, when present in the Madin dataset.

2.5 Transformer-based prediction of growth rate

2.5.1 Transformer model architecture. We used the *LookingGlass* v2 model (*LGM2*, [15]) to fine-tune our doubling time prediction models. LGM2 is an updated version of the original LookingGlass model [12], built upon a larger and more diverse training corpus. It is a transformer-based language model adapted from RoBERTa [21], designed to learn biological representations from DNA sequences. The model architecture includes 24 layers with 16 attention heads, a hidden size of 1024, and an intermediate feedforward size of 4096,

with a custom DNA nucleotide tokenizer resulting in a model with approximately 86 million parameters.

2.5.2 Training procedure. LGM2 was pretrained on >11M simulated short-read-like DNA sequences, each averaging 128 nucleotides in length (total training set ca. 1B tokens), derived from genomes sampled uniformly from each taxonomic order in the GTDB database [33]. The pretraining objective was masked language modeling (MLM) [21], where 15% of nucleotides in a sequence are masked at random and predicted using surrounding context.

2.5.3 Transformer fine-tuning for growth rate prediction. We fine-tuned the pretrained LGM2 model to predict the minimum doubling time using a supervised regression approach. For the $LGM2_{rib}$ model variant, the input consisted of nucleotide sequences from all annotated ribosomal genes. These sequences were tokenized using the custom DNA tokenizer provided by *LookingGlass* and segmented into chunks of up to 128 nucleotides in length to match the model's maximum input size.

The training data comprised of all protein-coding genes in each subset concatenated, such that all appropriately annotated sequences from an input genome were a single sequence, and then paired with its empirically measured minimum doubling time. For the ribosomal protein-coding gene subset, all of these sequences combined totaled over 2 million nucleotides; therefore, each genome-based input was on average roughly 4,200 nucleotides. Fine-tuning was performed using the Hugging Face Trainer API with the Adam optimizer and default settings. Training was conducted for 8 epochs using a batch size of 8. The selected hyperparameters included a weight decay of 0.01, a base learning rate of 2e-5, and 500 warmup steps.

Sequence embeddings were extracted from the final transformer layer and used as input for the downstream regression task. To incorporate growth temperatures in the prediction, the final-layer embeddings were concatenated with the corresponding optimal growth temperature value for each organism (where available in the Madin dataset). Model performance on the validation set was monitored at each epoch using Root Mean Squared Error (RMSE) to prevent overfitting. The best-performing model was selected based on the lowest validation RMSE.

2.6 Assessing minimum doubling time prediction performance

All regression frameworks, including *gRodon*, were evaluated using standard metrics including RMSE, Mean Absolute Error (MAE), R^2 and Pearson correlation coefficient.

Microbial isolates were classified into fast and slow growers based on multiple doubling time thresholds: 2.6-hour, 5-hour and 8-hour. Embeddings of the final transformer layers were visualized using tSNE plots.

Classification performance as either fast or slow-growing based on the specific thresholds above was evaluated using AUC-ROC, which measures consistency of the empirical and predicted minimum doubling times with respect to either fast and slow growth, as well as precision, and recall.

2.7 *Evo* minimum doubling time alternative model

To estimate microbial minimum doubling time, we trained a supervised ridge regression model on pooled embeddings from the final layer of the pretrained *Evo* foundation model. After passing through the *Evo* model, each nucleotide was encoded into a 4096-dimensional representation. To obtain a fixed-size representation for each sequence, we applied mean pooling across the sequence dimension, resulting in a single 4096-dimensional vector per sequence, which served as input to the ridge regression model. We selected the optimal L2 regularization parameter from the set {0.01, 0.1, 1, 10, 100} by minimizing the root mean squared error (RMSE) on the validation set.

3 Results

3.1 Ribosomal protein-coding genes encode the strongest signals for growth prediction

To identify which functionally annotated gene categories are most informative for predicting microbial growth rates, we fine-tuned *LookingGlass* v2 (*LGM2*) on distinct subsets of coding sequences grouped by COG (Clusters of Orthologous Groups) functional annotations (Table 1).

Among all tested subsets, the model trained exclusively on ribosomal protein sequences performed best. It achieved the lowest root mean square error (RMSE = 1.013) and mean absolute error (MAE = 0.764), along with the highest coefficient of determination (R^2 = 0.560) on the held-out test set.

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. 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). This suggests that including non-informative or weakly correlated genes may dilute signal and hinder model learning. This aligns with prior studies linking ribosomal gene composition and codon usage to translation efficiency and microbial growth potential.

3.2 *LookingGlass* improves doubling time prediction relative to CUB-based methods

We compared the predictive performance of the fine-tuned *LookingGlass* model trained on ribosomal genes ($LGM2_{rib}$) against two versions of the *gRodon* model using a held-out test dataset (Fig 1).

$LGM2_{rib}$ achieved the strongest Pearson correlation between the predicted and observed minimum doubling times (R = 0.75, $p < 0.005$), significantly outperforming both the original *gRodon* (R = 0.40) and the temperature-corrected *gRodon* ($gRodon_{tmp}$, R = 0.60; Fig. 1). This high correlation indicates that $LGM2_{rib}$ captures growth-related signals more effectively than existing tools.

In terms of absolute prediction accuracy, $LGM2_{rib}$ also achieved the lowest root mean squared error (RMSE = 1.013), confirming superior performance over *gRodon* (RMSE = 1.336) and $gRodon_{tmp}$

Table 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	1.901	1.379	0.071	1.007	0.431
Cellular processing and signaling	1.721	1.312	0.160	0.986	0.457
Information storage and processing	1.610	1.269	0.214	0.994	0.533
Ribosomal sequences	0.993	1.013	0.560	0.764	0.748

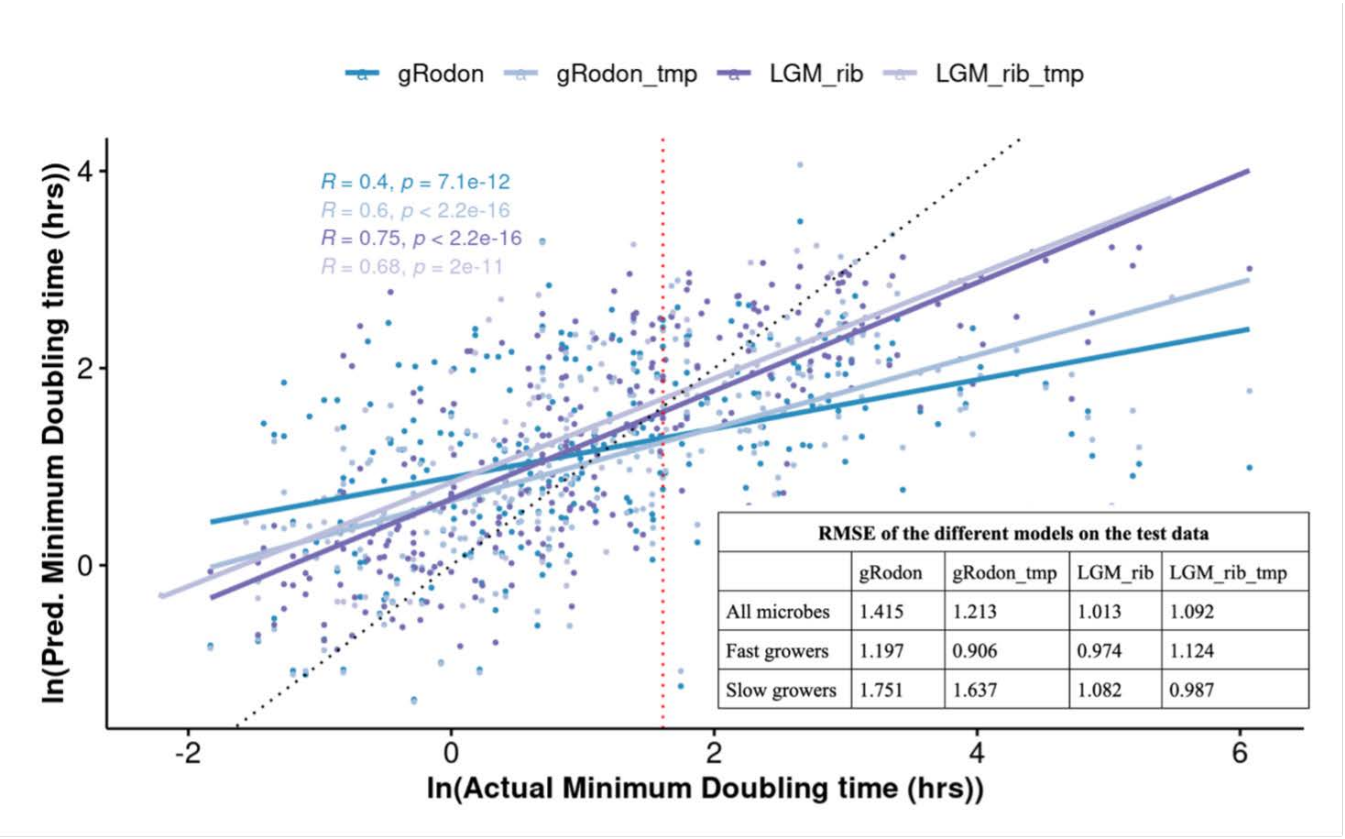


Figure 1: Comparison of predicted versus empirical minimum doubling times using three models: the fine-tuned *LookingGlass* model on ribosomal protein sequences ($LGM2_{rib}$), the original gRodon, and gRodon with temperature correction (gRodontmp). $LGM2_{rib}$ achieved the lowest root mean square error (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 for distinguishing fast- and slow-growing organisms.

(RMSE = 1.162). This suggests that ribosomal sequence embeddings derived from a transformer architecture encode more relevant features for predicting microbial doubling time than codon usage-based models.

To assess performance across microbial growth categories, we further divided test organisms into fast growers and slow growers, using *gRodon*'s default threshold (5-hour minimum doubling). For fast-growing taxa, $gRodon_{tmp}$ slightly outperformed $LGM2_{rib}$ (RMSE = 0.906 vs. 0.974), suggesting that codon usage bias may still offer some advantages for high-growth organisms. However, for slow-growing microbes, $LGM2_{rib}$ significantly outperformed both *gRodon* models (RMSE = 1.082 vs. $gRodon_{tmp}$ = 1.637 and *gRodon* = 1.751).

Our current framework utilizes embeddings from the final layer of the transformer model. An anonymous reviewer informed us that that intermediate layers—particularly the penultimate layer—can sometimes yield better performance. Unfortunately, embeddings from the penultimate layer had comparable to, and slightly worse

performance, relative to the final layer (data not shown), and thus no improvement was observed for this specific application.

3.3 *LookingGlass* works just as well on unseen genomes

One limitation of cross fold validation of this application is the taxonomic diversity in the original set can be highly variable and biased. To address this, we used the training approach of the original LookingGlass model [12]. gRodon, on the other hand, uses 10x fold cross validation using more traditional mathematical modeling performed 100 times. With a deep learning model, this would require 1000 fine tuning runs. "Traditional" 4-fold validation similar to our splits does not do as well, very likely due to taxonomic bias.

Our goal was to develop a powerful new tool for identifying microbes with high or low cultivation potential in metagenomic datasets. To address potential differences in training with respect to this specific goal, we collected 100 microbes from multiple datasets published after 2022 and therefore not in Madin et al. We further double checked to ensure no overlap. Our RMSE (1.35) is still lower than gRodon (1.46) on these genomes unseen by either model, with a similar R (0.35 vs 0.37). This is roughly a 8% improvement.

3.4 Generic DNA embeddings from *Evo* underperform relative to *LookingGlass*

To assess whether *LookingGlass*'s performance stems from its microbial-specific pretraining or simply reflects the general power of transformer-based DNA models, we evaluated a second deep learning model, *Evo*, which was pretrained on over 300 billion nucleotides from diverse prokaryotic and phage genomes. In contrast to *LGM2*, which was trained on shorter Illumina-sized sequences (roughly 100 nucleotides), *Evo*'s architecture is optimized for capturing long-range genomic dependencies. Specifically, it uses a StripedHyena architecture [34] with 29 Hyena layers interleaved with three multi-head attention layers, enabling substantially longer context windows.

Evo employs a simple 4-token vocabulary (one per nucleotide base) and has been successfully applied to both generative and predictive genomic tasks [30]. We used the original *Evo* model for this comparison, rather than the recently released *Evo* − 2, to ensure consistency with *LGM2* in training regime and model complexity.

We fine-tuned *Evo* embeddings on the same downstream task—predicting minimum doubling time from ribosomal protein-coding genes (Table 2)—and compared its performance to both $LGM2_{rib}$ and *gRodon*. Evo outperformed both the original and temperature-corrected versions of *gRodon* (RMSE = 1.276, R^2 = 0.427; and RMSE = 1.203, R^2 = 0.510 with and without temperature correction, respectively), but still lagged behind $LGM2_{rib}$ (RMSE = 1.013, R^2 = 0.560).

These results suggest that microbial domain-specific pretraining confers advantages beyond what is achievable through increased model capacity or context length alone. While general-purpose DNA models like *Evo* provide strong baselines, fine-tuning a transformer already attuned to microbial sequence patterns results in more accurate predictions of microbial doubling time.

		Evo embed	Evo embed+tmp
RMSE	All microbes	1.276	1.203
	Fast growers	1.177	1.092
	Slow growers	1.543	1.499
R^2		0.427(1.466e-04)	0.511(3.292e-06)

Table 2: Root mean square error (RMSE) and coefficient of determination (R^2) for models predicting minimum doubling time from ribosomal protein-coding sequences for the *Evo* alternative deep learning model. Like LGM_{rib} and *gRodon*, *Evo* also performs better for fast growing microbes.

3.5 *LookingGlass* model embeddings differentiate fast growers from slow growers

We also assessed how different minimum doubling time thresholds affected the classification of these microbes as either fast or slow growers using *LGM2*-derived sequence embeddings. Specifically, we compared three thresholds: 5-hour (see above), 2.6-hour (the empirical median of the training set), and 8-hour (the 75th percentile of the doubling time distribution), among others.

Of all thresholds tested, the 5-hour cutoff yielded the best overall classification performance, yielding the highest area under the receiver operating characteristic curve (AUC-ROC = 0.87), compared to 2.6-hour (AUC-ROC = 0.78) and 8-hour (AUC-ROC = 0.81). Performance plateaued for thresholds near 5 hours, suggesting this cutoff robustly separates fast- and slow-growing microbes.

Consistent with prior observations, our model more accurately identified fast growers than slow growers (Fig. 2a), further 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 (Fig. 2b–d), 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.

3.6 Cultured microbes are biased toward faster predicted growth across ecosystems

We applied the fine-tuned $LGM2_{rib}$ model to predict minimum doubling times for 25,103 metagenome-assembled genomes (MAGs) from the GEM (Genomes from Earth's Microbiomes) dataset. Only MAGs with at least 300 ribosomal coding nucleotides were included to ensure input quality.

Across most ecosystems, cultured MAGs skewed toward faster growth (<5 hours; Fig. 3). Environments such as city subways, ruminant hosts, plant-associated habitats, lab enrichments, and algae-associated systems were particularly enriched in cultured microbes predicted to have short doubling times (Fig. 3a), reinforcing known methological cultivation biases toward fast growers.

An exception was observed in human-associated environments, where cultured microbes spanned a broader range of predicted doubling times (Fig. 3a), 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.

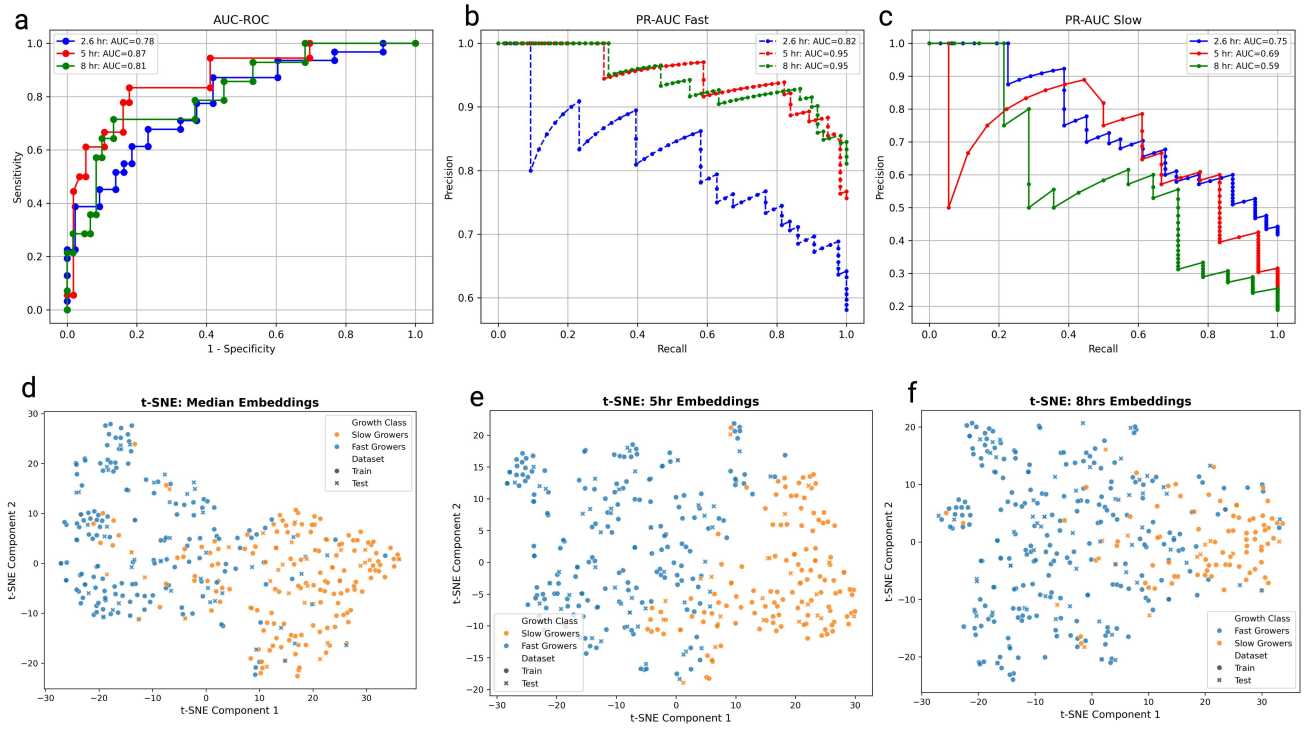


Figure 2: A 5-hour doubling time threshold improves microbial growth classification and cluster separation in *LookingGlass*-derived embeddings. (a-c) Classification performance using three doubling time thresholds (2.6, 5, and 8-hour). The fine-tuned LookingGlass model, when combined with predicted minimum doubling times, achieves the highest AUC-ROC at the 5-hour threshold. (d–f) t-SNE projections of LookingGlass-derived sequence embeddings show clustering of fast (blue) and slow (orange) growers under different classification thresholds: (e) 5 hours, (d) 2.6 hours (empirical median), and (f) 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.

Proteobacteria, Firmicutes, and Deinococcota—phyla with generally short doubling times—are largely cultured. In contrast, within Cyanobacteria, only fast-growing representatives appear to be cultured, while slow-growing members remain uncultured. Other phyla, such as Bacteroidota, Fusobacteriota, Actinobacteriota, and Campylobacteriota, include both fast- and slow-growing cultured strains (Fig. 4b). These findings demonstrate that genomic growth potential, as captured by *LGM2*, helps explain persistent cultivation biases across ecosystems and microbial phyla.

4 Discussion

To our knowledge, this is the first successful application of a transformer-based biological language model to predict microbial growth rates directly from genomic sequences – marking a significant step forward in linking genotype to phenotype in uncultured microbes. The fine-tuned *LGM2* model performed best when trained on ribosomal protein-coding sequences (Fig. 1), consistent with prior work showing that codon usage bias (CUB) in ribosomal and other constitutively expressed genes (e.g., translation factors) is strongly associated with translational efficiency and rapid growth [10, 45, 49]. Moreover, the conservation and ubiquity of ribosomal genes across microbial taxa [20, 40] make them well-suited for sequence-based prediction across isolates and metagenomes.

Compared to *gRodon* [51], a widely used tool based on codon usage features derived from highly expressed genes that predicts minimum doubling time, $LGM2_{rib}$ demonstrated substantially higher prediction accuracy 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.

While Evo, a general-purpose DNA language model, offers a promising deep learning baseline, it consistently underperformed compared to $LGM2_{rib}$ across all metrics. This suggests that microbe-specific pretraining—especially on short-read Illumina-scale sequences—captures growth-related signals more effectively than general-purpose, long-context DNA models. This result reinforces the importance of domain-adaptive language modeling in biological prediction tasks.

Like other sequence-based models, $LGM2_{rib}$ showed reduced accuracy for extremely slow-growing microbes (Fig. 2). This is likely due to their under-representation in the training data. Evolutionary dynamics—such as weak purifying selection, small effective

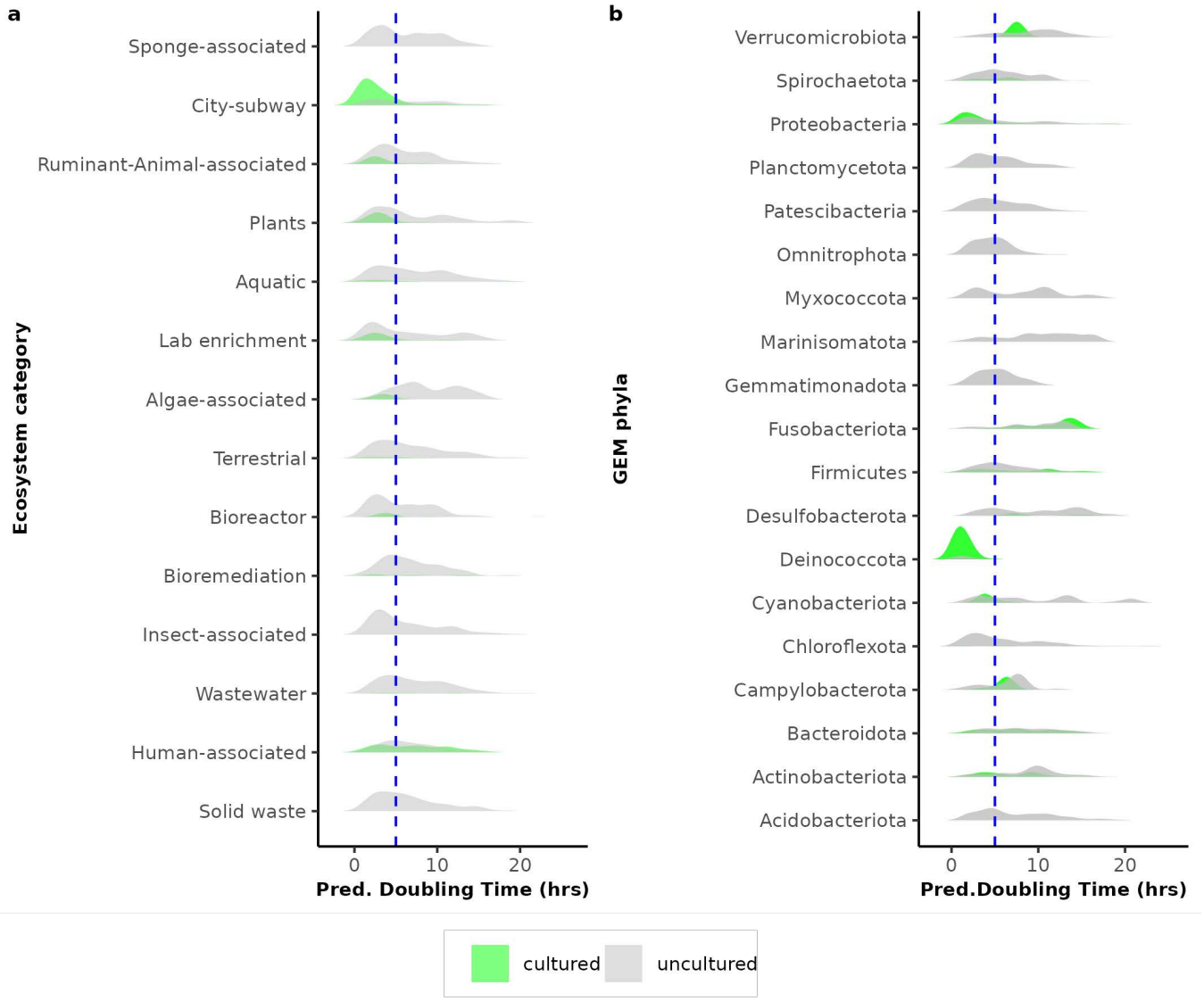


Figure 3: LGM predicts minimum doubling time of microbes inferred to be cultured (green) and uncultured (gray) in GEM dataset across a) ecosystem and b) GEM phyla. The vertical dashed blue line represents a doubling time of 5 hrs, distinguishing fast growers from slow growers. Only phyla with over 50 representatives and ecosystems with at least 75 representatives are shown.

population sizes, and increased genetic drift—may further reduce the strength of codon optimization signals [6, 25]. Nonetheless, $LGM2_{rib}$ demonstrated a clear improvement over prior approaches in predicting doubling times for these organisms, offering a new avenue for identifying ecologically important taxa often overlooked due to cultivation bias. This improvement is particularly relevant for efforts to cultivate slow-growing microbe or to model microbial contributions to ecosystem function.

Our findings are consistent with prior observations that codon optimization signals plateau around minimum doubling times of 5–6 hours, likely reflecting biological constraints on translational efficiency [51]. Classification based on a 5-hour threshold yielded the best performance (Fig. 3), outperforming other thresholds considered. This likely reflects biologically meaningful sequence patterns that differentiate slower-growing taxa [49].

When applied to over 25,000 metagenome-assembled genomes (MAGs) from the Genomes from Earth's Microbiomes catalog, $LGM2_{rib}$ revealed striking differences in predicted doubling times between cultured and uncultured microbes across diverse environments (Fig. 3). Cultured microbes were typically predicted to grow faster, particularly in nutrient-rich or anthropogenic environments such

as subways. These trends mirror longstanding cultivation biases that favor copiotrophic microbes—those adapted to nutrient-rich environments—while often excluding oligotrophic taxa that grow slowly under nutrient-limited conditions [19, 47]. Human-associated microbes were an exception with successful cultivation spanning a broad range of growth rates. This likely reflects decades of targeted cultivation, sequencing, and functional profiling of pathogenic and commensal taxa in efforts such as the Human Microbiome Project [23].

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 [38]. 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 [4, 7].

A key limitation of our approach is the scarcity of microbial genomes with experimentally verified minimum doubling times, especially among slow growers. This restricts both the diversity of training data and the generalizability of the model, particularly for underrepresented slow growers. Continued expansion of cultivation efforts [5, 18, 28, 36] and phenotypic databases [27], coupled with standardized trait curation, will be essential for developing more generalizable and ecologically informed prediction models—especially for poorly characterized and slow-growing microbes.

5 Conclusion

Our results demonstrate that domain-informed transformer models like *LGM2* can bridge gaps in microbial trait prediction, enabling better ecological inference from genomic data—even in the absence of cultivation or metadata. By demonstrating the utility of a domain-adapted transformer model for a key microbial life history trait, this work provides a foundation for more accurate phenotype prediction in environmental and clinical contexts. This and future improvements 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.

Acknowledgments

ADS and IO gratefully acknowledge funding from NSF OCE-2145434. chatGPT version 4.0 was used to clarify some sections and to score each section from 1 to 10 to suggest areas to improve as an experiment by SJE in advance of teaching DSE 511 (Computational Methods for Data Science) at UTK, acceptable as such by current AI ACM policy.

References

[1] Karine Alain and Joël Querellou. 2009. Cultivating the uncultured: Limits, advances and future challenges. *Extremophiles* 13, 4 (2009), 583–594. doi:10.1007/s00792-009-0261-3

[2] Farooq Azam and Francesca Malfatti. 2007. Microbial structuring of marine ecosystems. *Nature Reviews Microbiology* 5, 10 (2007), 782–791.

[3] Ashley Babjac, Owen Queen, Shawn-Patrick Barhorst, Kambiz Kalhor, Andrew D. Steen, and Scott J. Emrich. 2023. Adapting protein language models for explainable fine-grained evolutionary pattern discovery. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2609–2616.

[4] Negash Kabtimer Bereded, 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 (2022), 87. doi:10.1186/s12866-022-02496-z

[5] Karissa L. Cross, 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 (2019), 1314–1321.

[6] Les Dethlefsen and Thomas M. Schmidt. 2007. Performance of the translational apparatus varies with the ecological strategies of bacteria. *Journal of Bacteriology* 189, 8 (2007), 3237–3245. doi:10.1128/JB.01686-06

[7] Li Dong, 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 (2018), 1603.

[8] Edo Dotan, Gal Jaschek, Tal Pupko, and Yonatan Belinkov. 2024. Effect of tokenization on transformers for biological sequences. *Bioinformatics* 40, 4 (2024).

[9] R. G. Eagon. 1962. *Pseudomonas Natriegens*, a marine bacterium with a generation time of less than 10 minutes. *Journal of Bacteriology* 83, 4 (1962), 736–737. doi:10.1128/jb.83.4.736-737.1962

[10] Idan Frumkin, 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 (2018). doi:10.1073/pnas.1719375115

[11] Michael Y Galperin, Yuri I Wolf, Kira S Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V Koonin. 2020. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research* 49, D1 (11 2020), D274–D281. doi:10.1093/nar/gkaa1018

[12] A. Hoarfrost, 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 (2022), 1–12.

[13] Hiroyuki Imachi, 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. (2025). biorxiv:: 2025–02

[14] Hiroyuki Imachi, 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 (2020), 519–525.

[15] Ratish Jha and Adrienne Hoarfrost. 2025. LookingGlass-2 (Revision 270a2dc). doi:10.57967/hf/5761

[16] Christian Jogler, 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 (2011), 5826–5829. doi:10.1128/AEM.05132-11

[17] Arthur L. Koch. 2001. Oligotrophs versus copiotrophs. *BioEssays* 23, 7 (2001), 657–661. doi:10.1002/bies.1091

[18] William H. Lewis, Guillaume Tahon, Patricia Geesink, Diana Z. Sousa, and Thijs J. G. Ettema. 2021. Innovations to culturing the uncultured microbial majority. *Nature Reviews Microbiology* 19, 4 (2021), 225–240.

[19] Junhui Li, 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 (2019), 2162–2172.

[20] Kyungtaek Lim, Yoshikazu Furuta, and Ichizo Kobayashi. 2012. Large variations in bacterial ribosomal RNA genes. *Molecular Biology and Evolution* 29, 10 (2012), 2937–2948.

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, et al. 2019. RoBERTa: A Robustly optimized BERT pretraining approach. (2019). doi:10.48550/arXiv.1907.11692

[22] Karen G. Lloyd, Andrew D. Steen, Joshua Ladau, Junqi Yin, and Lonnie Crosby. 2018. Phylogenetically novel uncultured microbial cells dominate earth microbiomes. *MSystems* 3, 5 (2018), e00055–18.

[23] Jason Lloyd-Price, Anup Mahurkar, Gholamali Rahnavard, Jonathan Crabtree, Joshua Orvis, A Brantley Hall, Arthur Brady, Heather H Creasy, Carrie McCracken, Michelle G Giglio, Daniel McDonald, Eric A Franzosa, Rob Knight, Owen White, and Curtis Huttenhower. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* 550, 7674 (2017), 61–66. doi:10.1038/nature23889

[24] Andrew M. Long, Shengwei Hou, J. Cesar Ignacio-Espinoza, and Jed A. Fuhrman. 2021. Benchmarking microbial growth rate predictions from metagenomes. *The ISME Journal* 15, 1 (2021), 183–195.

[25] Michael Lynch. 2010. Evolution of the mutation rate. *Trends in Genetics* 26, 8 (2010), 345–352.

[26] Ali Madani, Eric R. Greene Ben Krause, 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 (2023), 1099–1106.
[27] Joshua S. Madin, Daniel A. Nielsen, Maria Brbic, Ross Corkrey, David Danko, Kyle Edwards, Martin K. M. Engqvist, et al. 2020. A synthesis of Bacterial and Archaeal phenotypic trait data. *Scientific Data* 7, 1 (2020), 170.
[28] Da-Shuai Mu, Yang Ouyang, Guan-Jun Chen, and Zong-Jun Du. 2021. Strategies for culturing active/dormant marine microbes. *Marine Life Science and Technology* 3, 2 (2021), 121–131. doi:10.1007/s42995-020-00053-z
[29] Stephen Nayfach, 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 (2021), 499–509.
[30] Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, et al. 2024. Sequence modeling and design from molecular to genome scale with Evo. *Science* 386, 6723 (2024), eado9336.
[31] Carlos Outeiral and Charlotte M. Deane. 2024. Codon language embeddings provide strong signals for use in protein engineering. *Nature Machine Intelligence* 6, 2 (2024), 170–179.
[32] Donovan H. Parks, Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J. Mussig, and Philip Hugenholtz. 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology* 38, 9 (2020), 1079–1086. doi:10.1038/s41587-020-0501-8
[33] Donovan H. Parks, 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 (2018), 996–1004.
[34] Michael Poli, Jue Wang, Stefano Massaroli, Jeffrey Quesnelle, Ryan Carlow, Eric Nguyen, and Armin Thomas. 2023. *StripedHyena: Moving Beyond Transformers with Hybrid Signal Processing Models.* doi:10.57967/hf/1595
[35] P. Buford Price and Todd Sowers. 2004. Temperature dependence of metabolic rates for microbial growth, maintenance, and survival. *Proceedings of the National Academy of Sciences* 101, 13 (2004), 4631–4636. doi:10.1073/pnas.0400522101
[36] Andre A. Pulschen, 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 (2017), 1346.
[37] Michael S. Rappé, Stephanie A. Connon, Kevin L. Vergin, and Stephen J. Giovannoni. 2002. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* 418, 6898 (2002), 630–633.
[38] Alexandra-M. Reimers, 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 (2017). doi:10.1073/pnas.1617508114
[39] Alexander Rives, 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 (2021). doi:10.1073/pnas.2016239118 Article e2016239118.
[40] Monica Santamaria, 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 (2012), 682–695.
[41] David B. Sauer and Da-Neng Wang. 2019. Predicting the optimal growth temperatures of Prokaryotes using only genome derived features. *Bioinformatics* 35, 18 (2019), 3224–3231.
[42] Robert Schmirler, Michael Heinzinger, and Burkhard Rost. 2024. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications* 15, 1 (2024), 7407.
[43] Oliver Schwengers, Torsten Hain, Trinad Chakraborty, and Alexander Goesmann. 2019. ReferenceSeeker: Rapid determination of appropriate reference genomes. (2019). biorxiv:: 863621
[44] Torsten Seemann. 2014. Prokka: Rapid Prokaryotic genome annotation. *Bioinformatics* 30, 14 (2014), 2068–69.
[45] Premal Shah 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 (2011), 10231–10236. doi:10.1073/pnas.1016719108
[46] Paul M. Sharp, 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 (2005), 1141–1153.
[47] Andrew D. Steen, 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 (2019), 3126–3130.
[48] Sonia R. Vartoukian, Richard M. Palmer, and William G. Wade. 2010. Strategies for culture of 'unculturable' bacteria. *FEMS Microbiology Letters* 309, 1 (08 2010), 1–7. doi:10.1111/j.1574-6968.2010.02000.x
[49] Sara Vieira-Silva and Eduardo P. C. Rocha. 2010. The systemic imprint of growth and its uses in ecological (Meta) genomics. *PLoS Genetics* 6, 1 (2010), e1000808.
[50] Jeth Walkup, 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 (2023), 73.
[51] J. L. Weissman, 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 (2021). doi:10.1073/pnas.2016810118
[52] Liang Xu, Emily Zakem, and J. L. Weissman. 2024. Improved maximum growth rate prediction from microbial genomes by integrating phylogenetic information. (2024). biorxiv:: 2024–2010

A 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)

B Code availability

The codes are available at https://github.com/UTbioinf/microbial-growth-rate-transformer/