南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY



Quantitative Analysis of Microbial Growth Rates Using Machine Learning and Deep Learning

Shengwei Hou
Southern University of Science and Technology
2025/07/04

海洋科学与工程系
Department of Ocean Science and Engineering



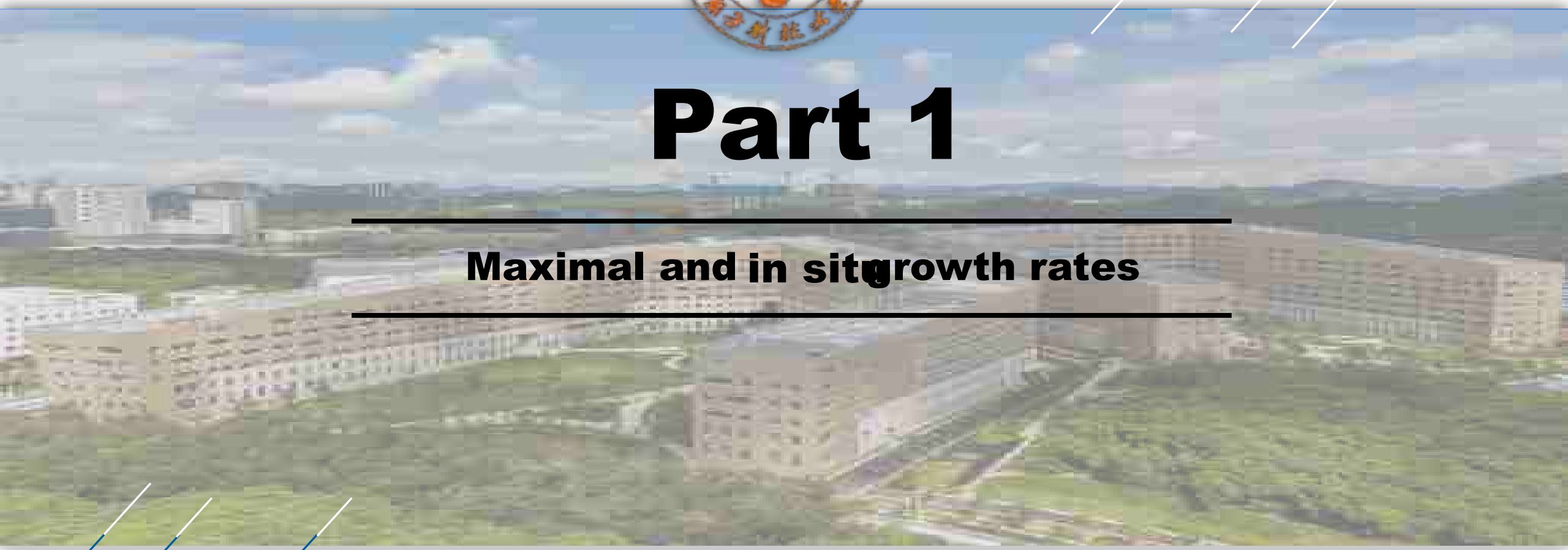
Content

01： Maximal and in situ growth rates

02： Benchmarking growth rate methods

03： Maximal growth rate prediction

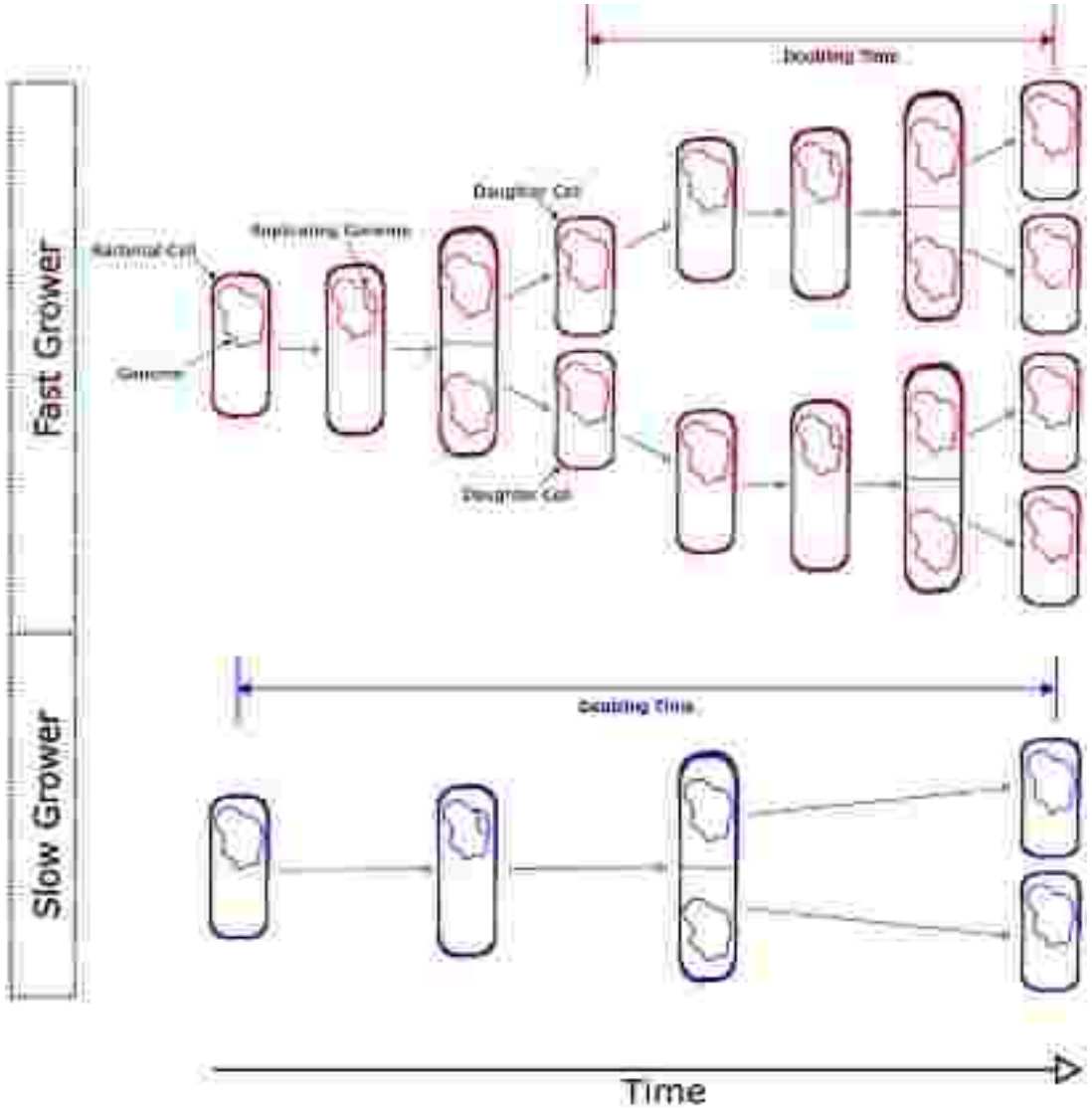
04： Growth rate and OGT prediction



Part 1

Maximal and in situ growth rates

Growth rates



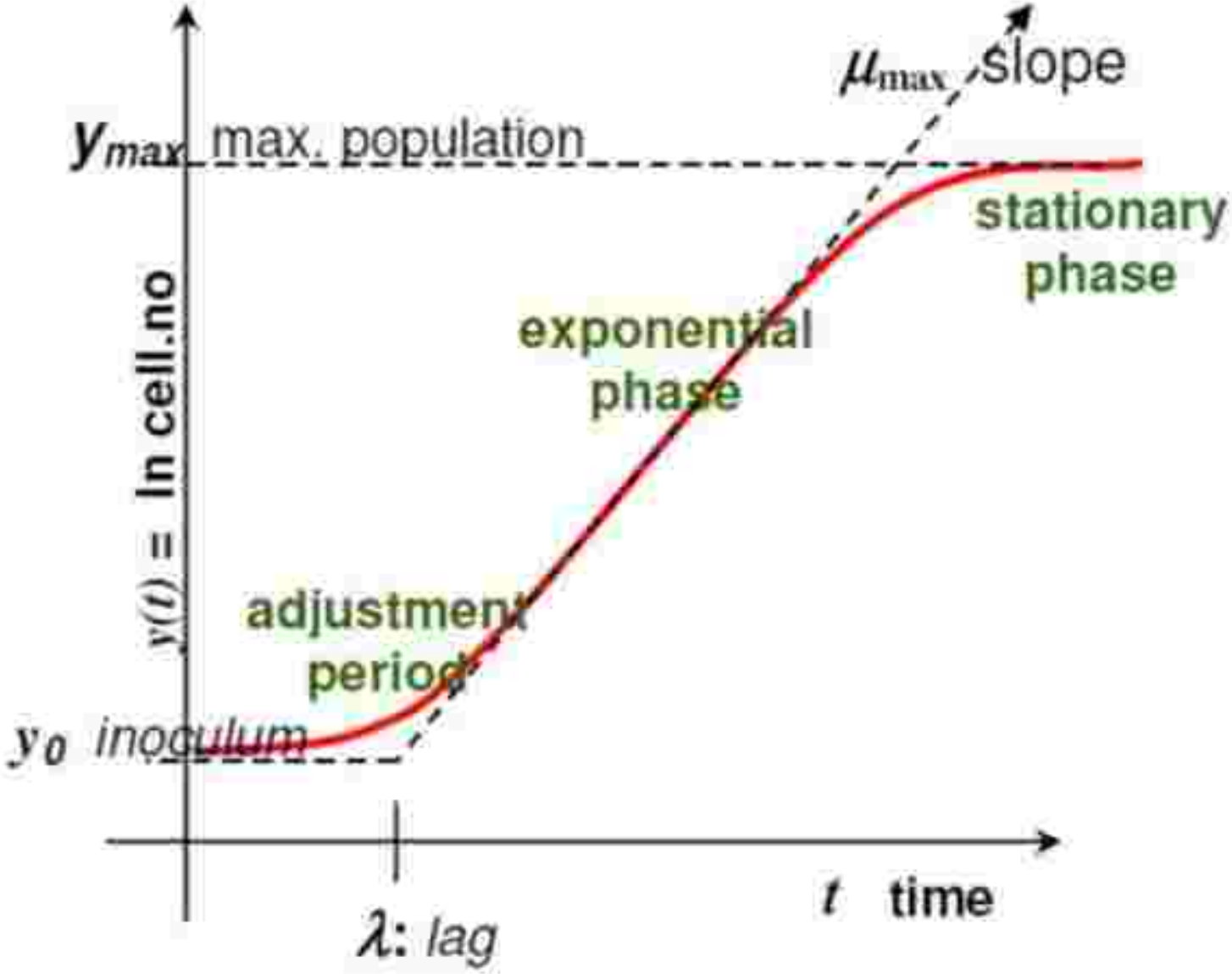
Weissman et al., 2022

- The microbial (maximal) growth rate is a fundamental parameter that varies among different microbes.
- Traditionally, microbial (maximal) growth rates can be measured through laboratory culture experiments if the microbes are culturable.
- However, these methods are time-consuming, labor-intensive, and often limited to a small fraction of the microbial diversity. (excl. extremophiles)





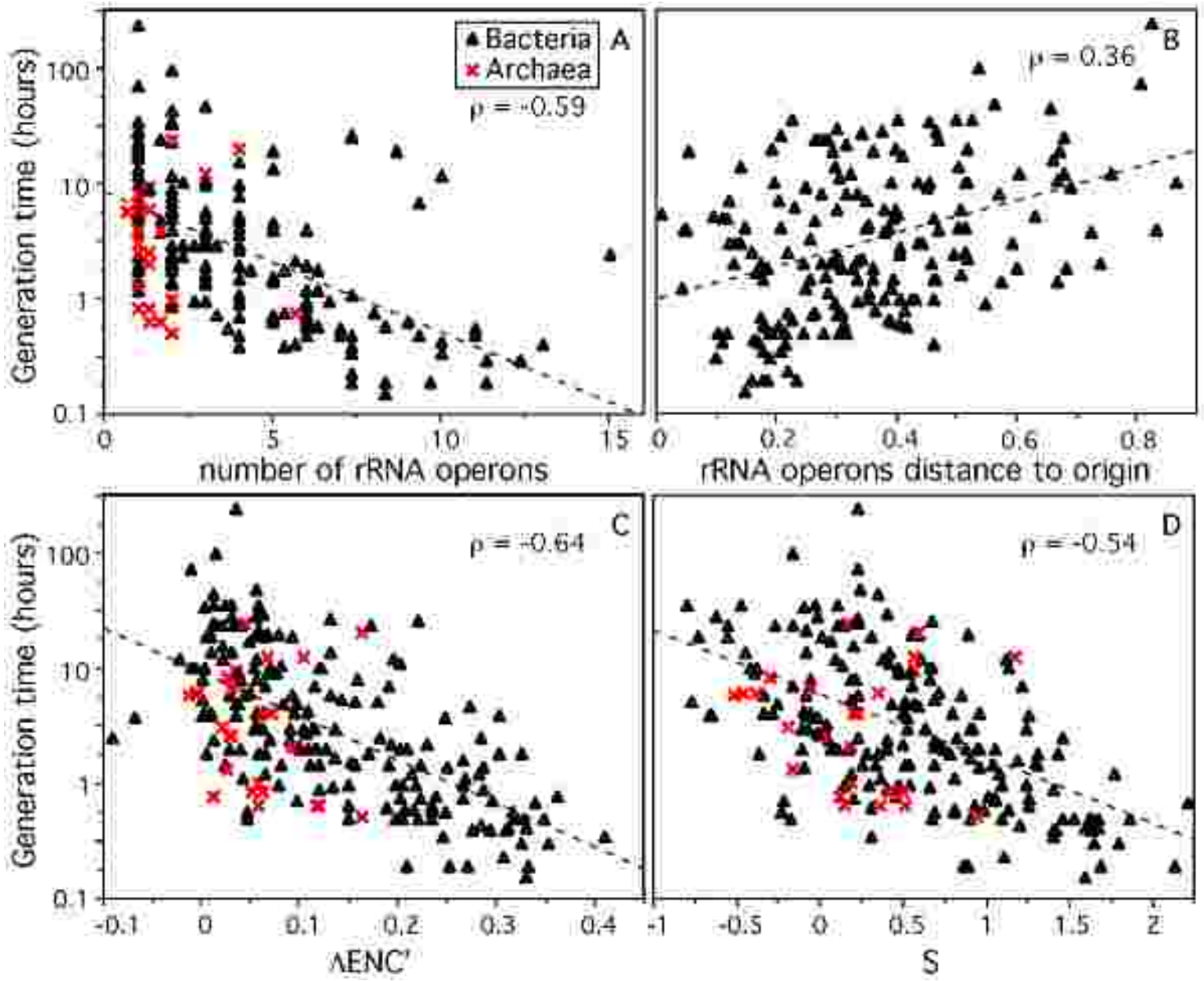
Growth rates



Gaspar et al., 2021 Chem Engineering

- Specific Growth Rate (μ) is the rate at which a microbial population increases per unit of biomass. It's often expressed as the change in biomass concentration (dX) over time (dt), divided by the biomass concentration (X) itself: $\mu = (dX/dt) / X$.
- The Maximum Specific Growth Rate (μ_{max}) is the highest specific growth rate that a microorganism can achieve under optimal conditions, such as an ideal temperature, pH, nutrient availability, and the absence of inhibitors.
- μ_{max} of a microorganism is a key parameter in microbial kinetics.

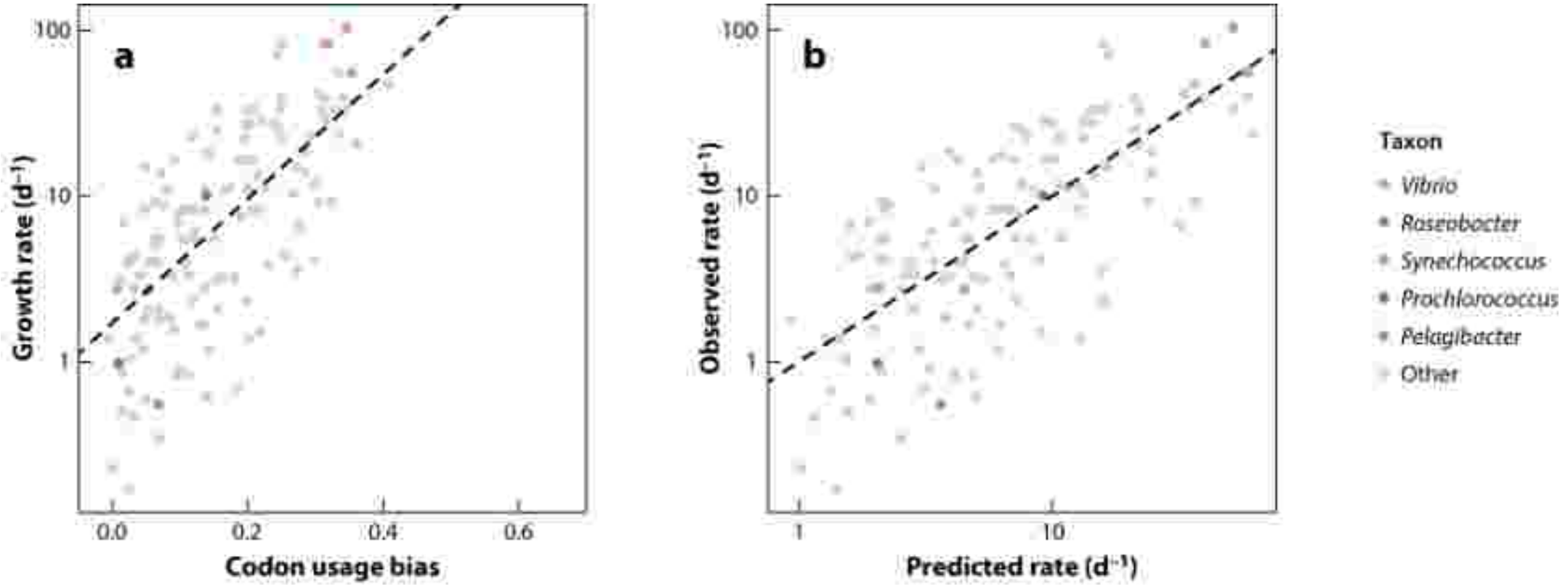
CUB-based growth prediction



- Codon Usage Bias (CUB): fast-growing microbes tend to use the set of synonymous codons corresponding to the most abundant tRNAs (the bias) to synthesize ribosomal proteins. Codon usage bias is thought to result from selection for accurate and fast translation by maximizing the recruitment of the most abundant tRNAs into ribosomes.
- Correlation between d and (A) the number of rRNA operons in the genome, (B) the relative distance from the origin of replication to rRNA genes (excluding species with no retrievable origin), 0.5 corresponds to half the replicon, (C,D) codon usage bias indices ΔENC′ and S.
- Codon usage biases are better correlates of growth rates than any other trait, including rRNA copy number.

Vieira-Silva et al., 2010 PLoS Genetic

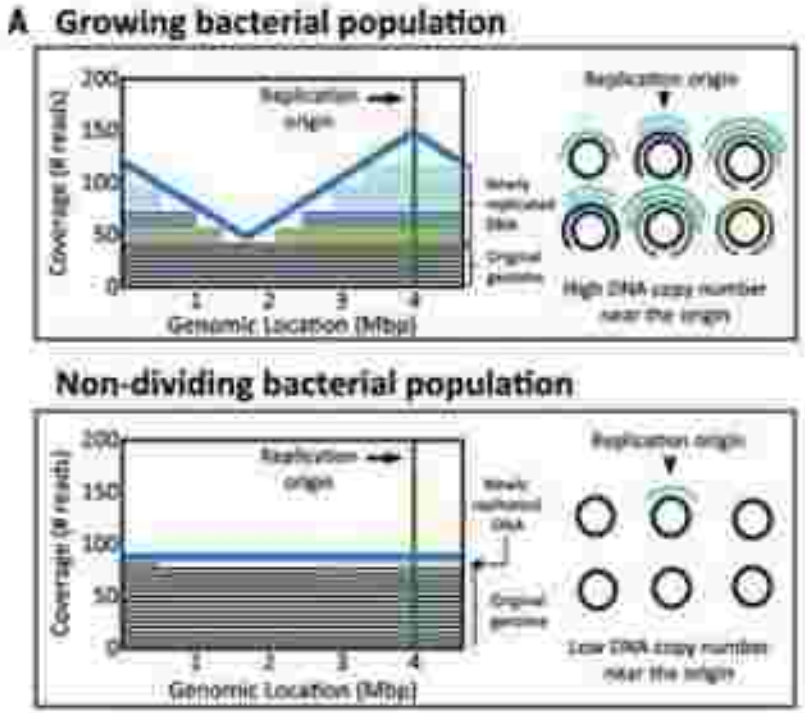
CUB-based growth prediction

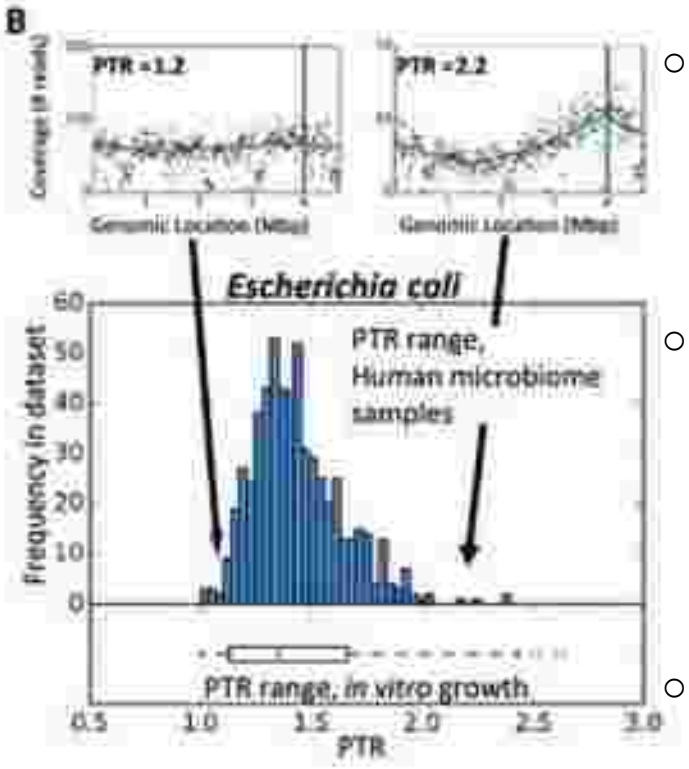


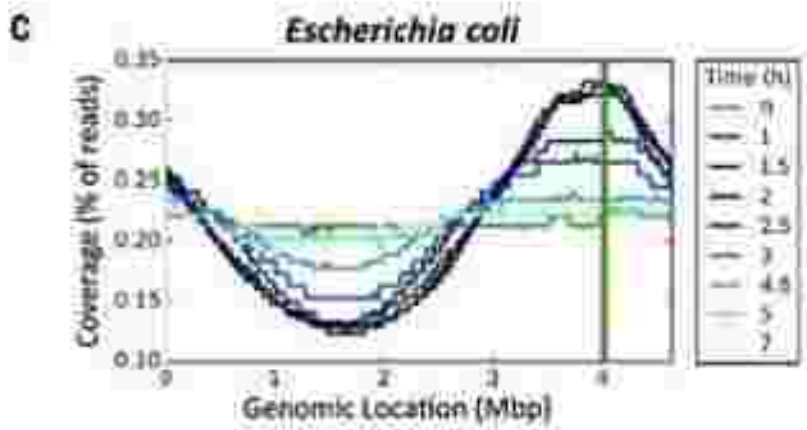
- (a) Maximum growth rate for bacteria as a function of ΔENC′, an index of codon usage bias.
- (b) The predicted maximum growth rate versus the observed maximum rate for select bacteria.
- The growth rate was predicted by a multiregression analysis using two indices of codon usage bias (ΔENC′ and S), the number of rRNA operons, and the distance of the rRNA operons to the origin of replication.

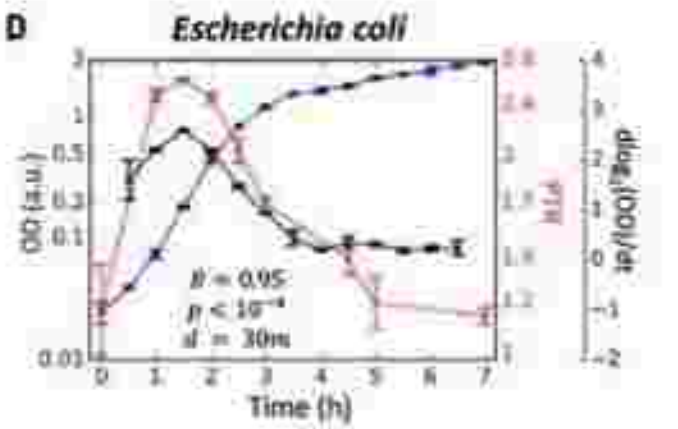
Kirchman et al., 2016 Ann Rev Mar Sci

Peak-to-Trough Ratio (PTR)





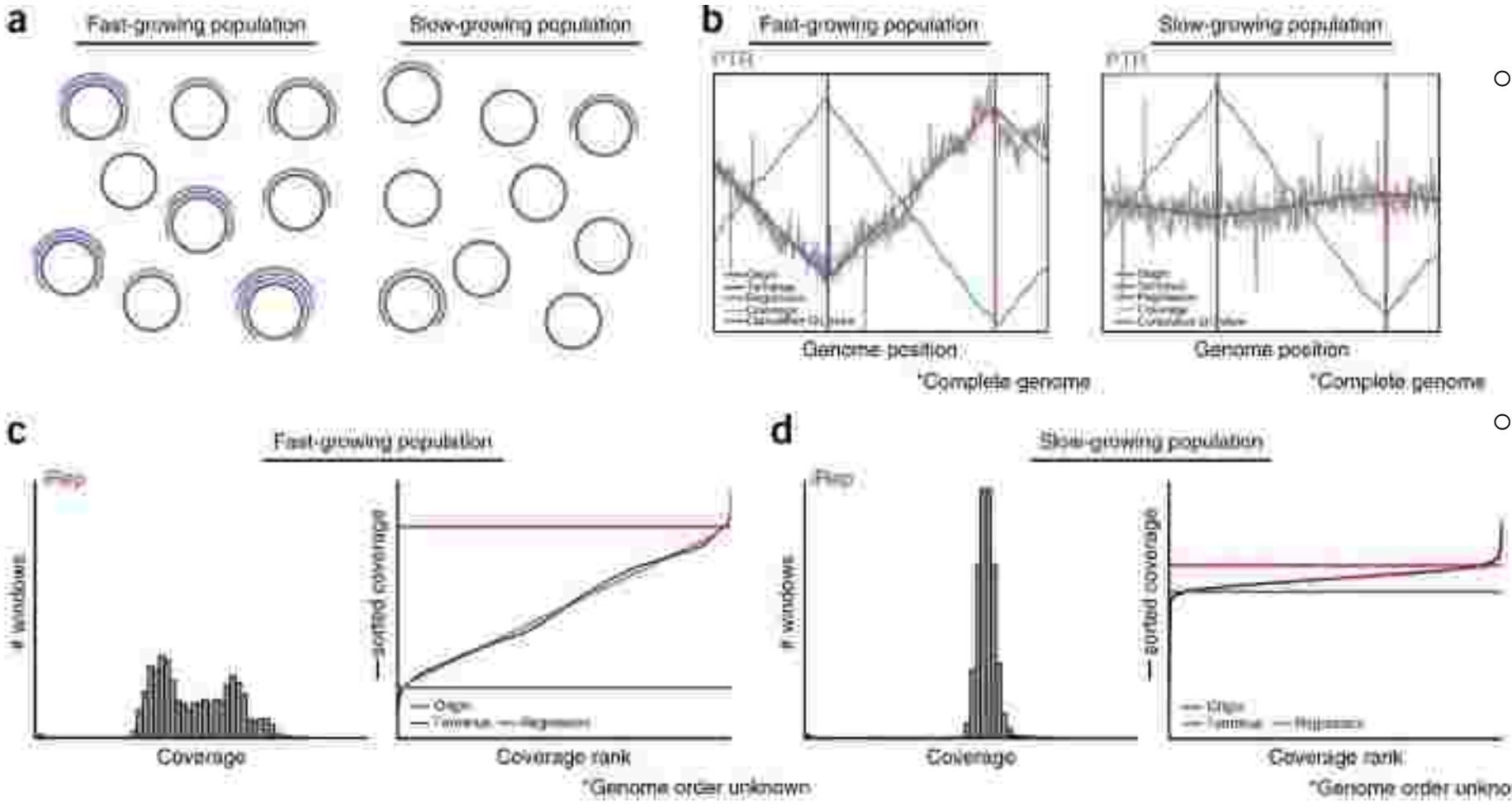




- Peak to Trough Ratio (PTR) refers to the ratio of metagenomic read coverage near the origin (the peak) of genome replication to the terminus (the trough).
- The pattern of metagenomic sequencing read coverage for different microbial genomes contains a single trough and a single peak, the latter coinciding with the bacterial origin of replication.
- Each bacterial cell in a growing population will be at a different stage of DNA replication, generating a coverage pattern that peaks near the known replication origin. This produces a prototypical sequencing coverage pattern with a single peak and a single trough. Bacteria from a non-dividing population have a flat sequencing coverage pattern across the genome.

Korem et al., 2015 Science

iRep to estimate the replication rate



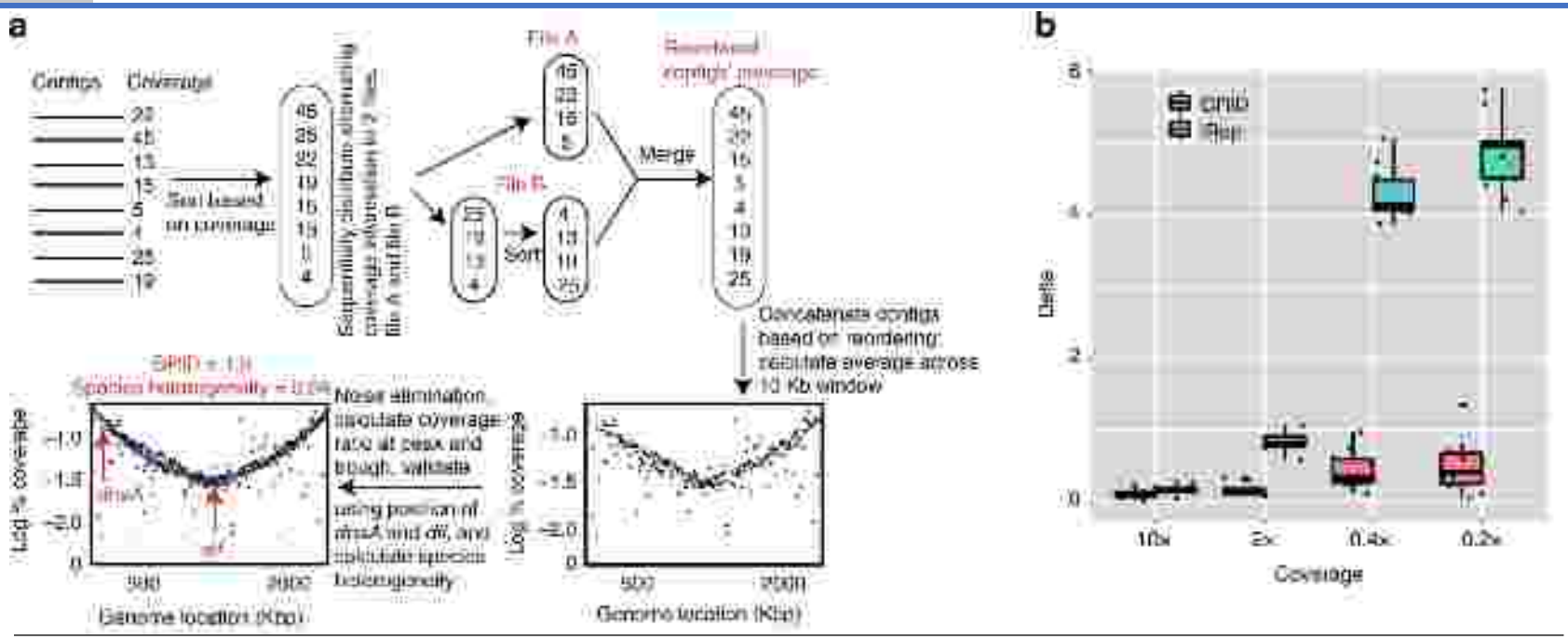
- The iRep algorithm calculates an index of replication (iRep) based on the sequencing coverage trend that results from bi-directional genome replication from a single origin of replication.
- Differences in genome copy number across a population of replicating cells can be determined based on sequencing read coverage over complete genome sequences.

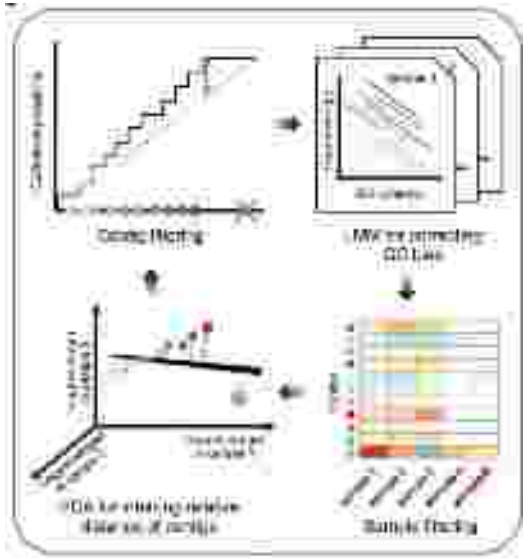
- The ratio between the coverage at the origin ("peak") and terminus ("trough") of replication (PTR) relates to the replication rate of the population. The origin and terminus can be determined based on cumulative GC skew.
- If no complete genome sequence is available, it is possible to calculate the replication rate based on the distribution of coverage values across a draft-quality genome.

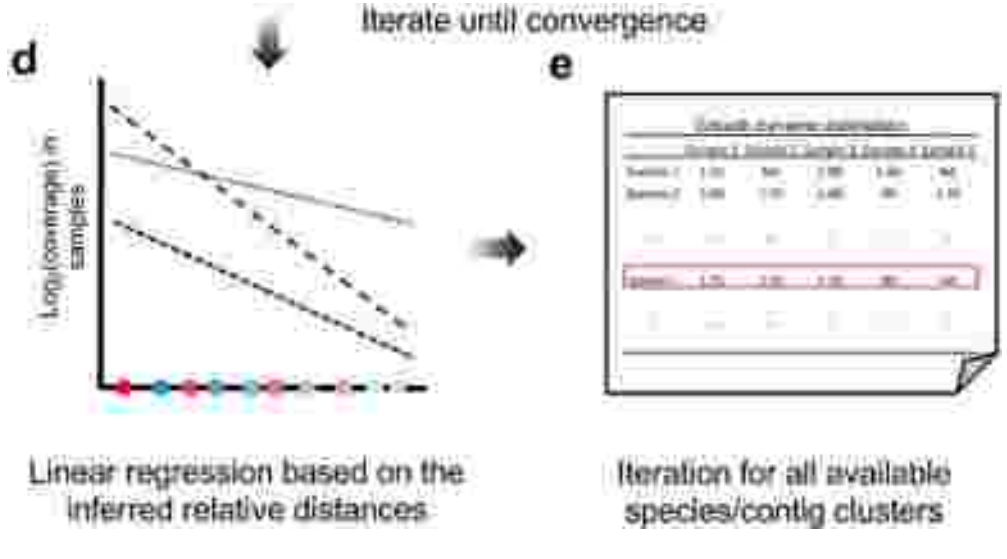
Brown et al., 2016 Nat Biotech

GRiD and DEMIC

海洋科学与工程系
Department of Ocean Science and Engineering

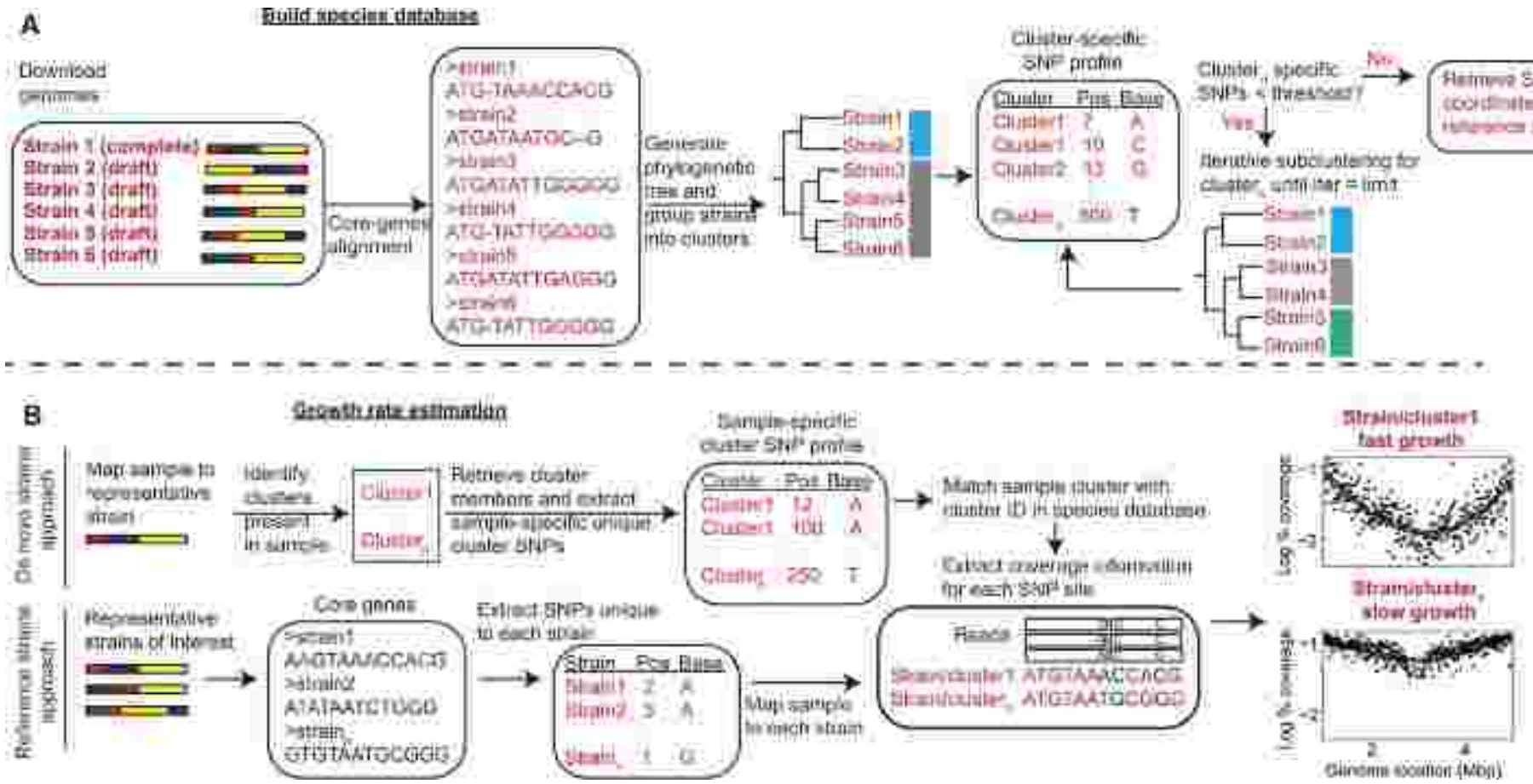






- Growth Rate InDex (GRiD) for estimating *in situ* growth rates of ultra-low coverage (>0.2×) and de novo-assembled metagenomes.
- GRiD then calculates a "refined" growth value with additional statistical filters to reduce noise.
- Dynamic Estimator of Microbial Communities (DEMIC) uses contigs and coverage values to infer relative distances of contigs from replication origin and to accurately compare bacterial growth rates between samples.
- It takes advantage of highly fragmented contigs assembled in multiple metagenomic samples.

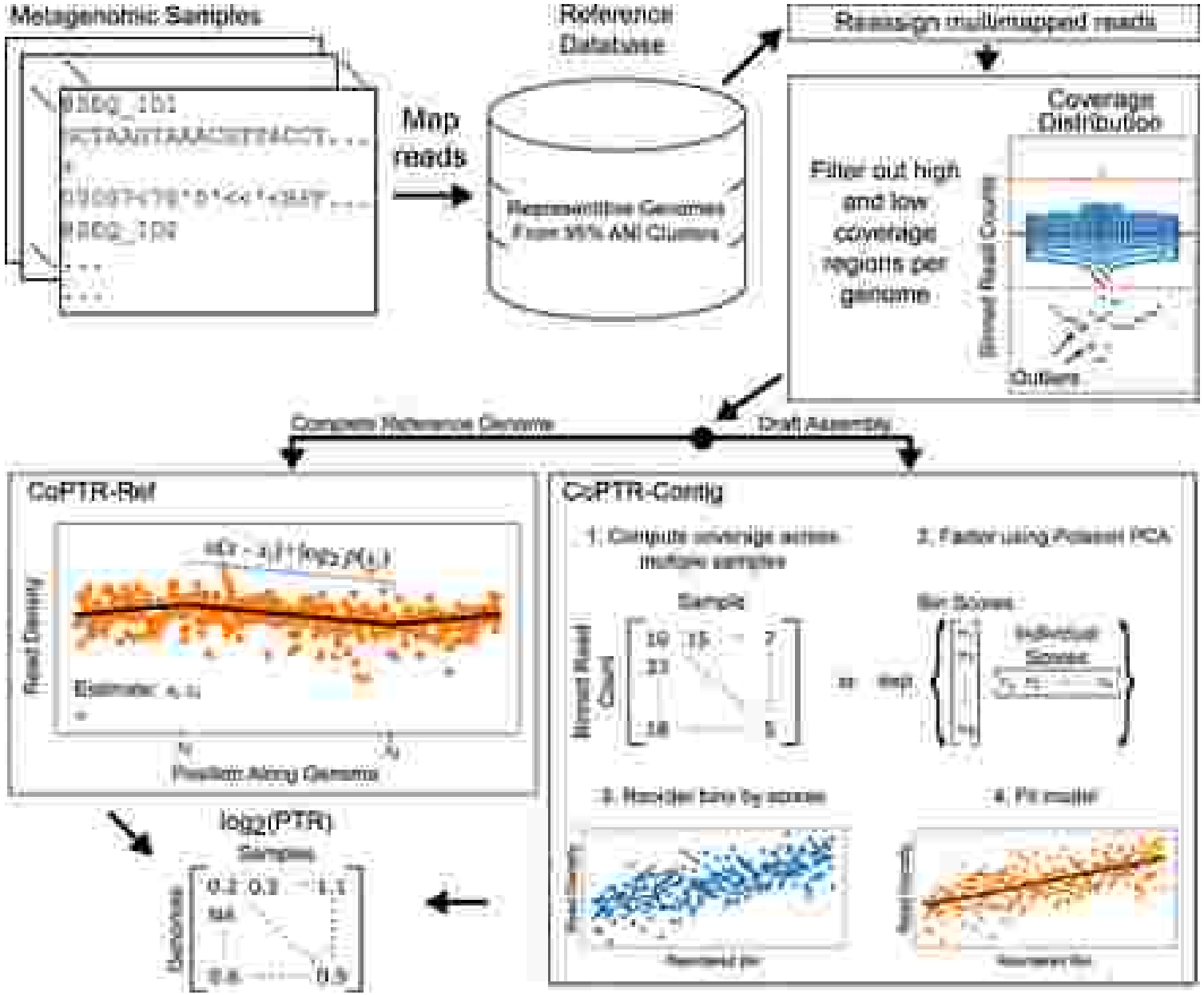
Emiola et al., 2018 Nat Comm; Gao et al., 2018 Nat Met



- Building the species database of member strains available for a species. Core genes are extracted, and multiple sequence alignment is conducted to generate a phylogenetic tree. Strains are grouped into clusters to generates cluster-specific unique SNPs shared within cluster members but absent in strains from other clusters. Clusters having insufficient number of unique SNPs are iteratively subclustered to re-generate a unique SNPs profile, or "SNP coordinate set."

- SMEG maps reads to the representative strain to detect clusters in a sample. SMEG then re-generates a sample-specific, unique SNP profile using only the detected clusters, increasing the number of discriminatory SNPs usable to calculate the growth rate.
- Then, nucleotide coverage is determined at each cluster's unique SNP site across a variable sliding window to calculate the peak-to-trough ratio (PTR).

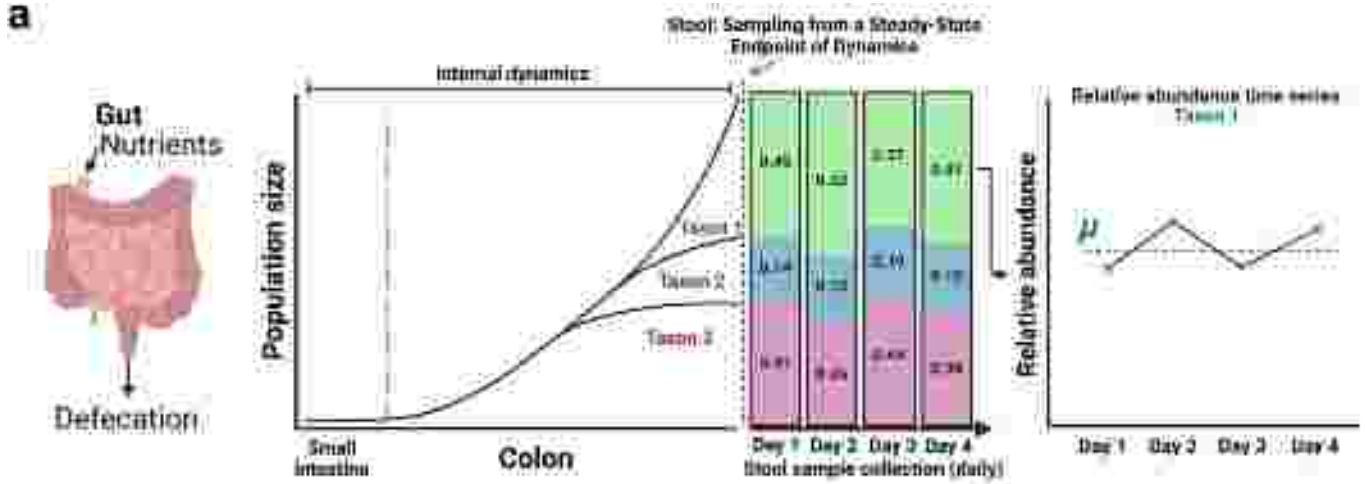
Emiola et al., 2020 Sci Adv

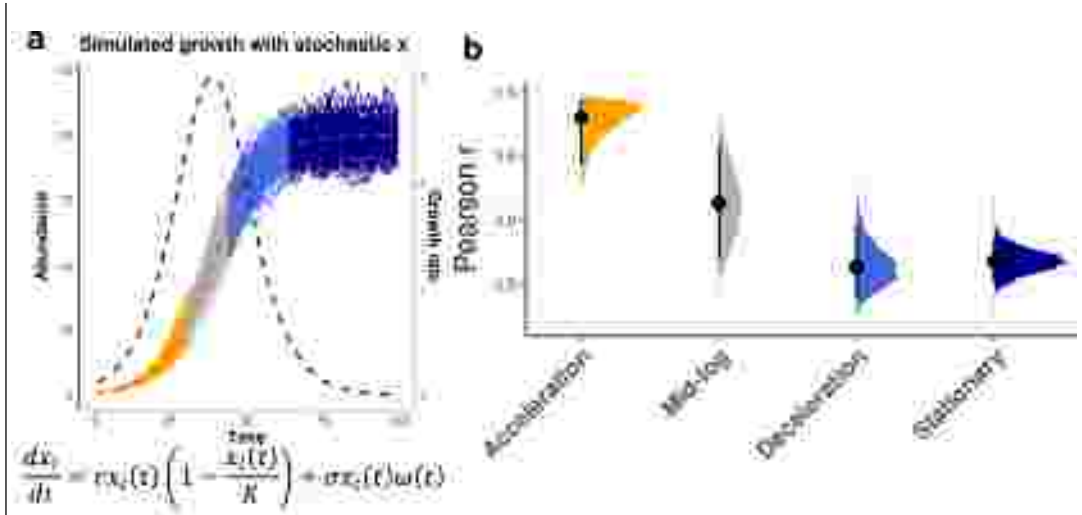


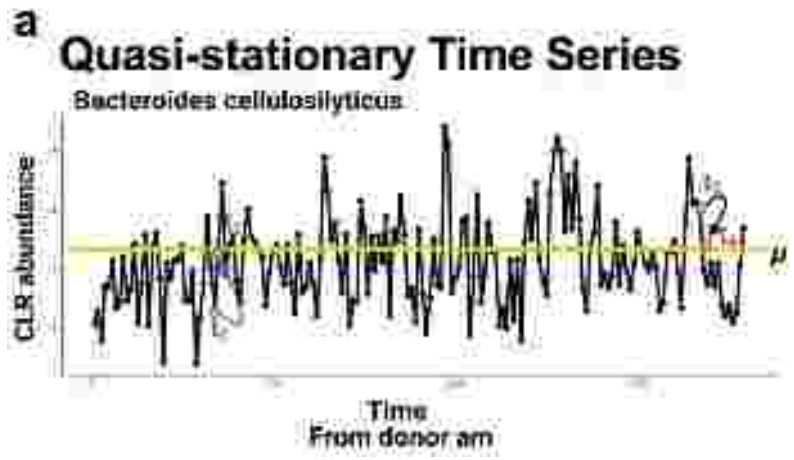
- Sequencing reads from multiple metagenomic samples are mapped to a reference database containing representative strains from complete reference genomes and high-quality assemblies. Multimapped reads are reassigned to a single genome using a probabilistic model.
- Regions of each genome with ultra-high or ultra-low coverage are filtered using filters designed for complete reference genomes or draft assemblies.
- For species with complete reference genomes, PTRs are estimated by maximizing the likelihood of a model describing the density of reads along the genome (CoPTR-Ref). For species with high-quality assemblies, reads are binned across the assembly, bins are reordered based on sequencing coverage across multiple samples using Poisson PCA, and the slope along this order is estimated by maximum likelihood (CoPTR-Contig).
- CoPTR outputs a table of the log2(PTR) per genome in each sample.

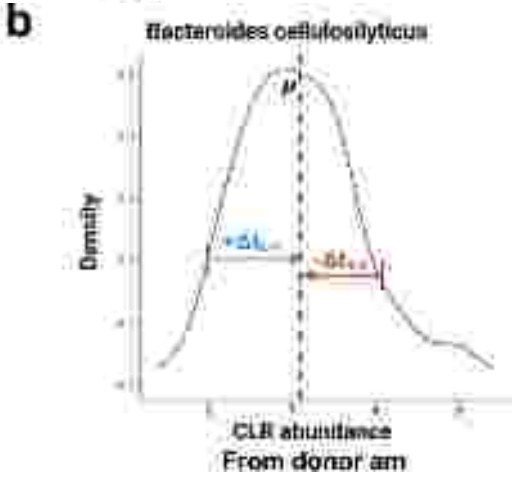
Joseph et al., 2022 Genome Res

Growth phase estimation









- When sampling from a stationary distribution, which is equivalent to 'regression-to-the-mean' as an organism fluctuates around a fixed carrying capacity.
- An alternative stochastic logistic growth model has been proposed, where population dynamics occur internally and the bacterial population sizes measured in a bolus of stool represent a steady-state endpoint of these dynamics.
- This model provides a path toward estimating the growth phases of gut bacterial populations *in situ*.

Lim et al., 2023 Nat Comm

Marine microbial growth rates



Kirchman et al. , 2016



- Community growth rates are measured by a) bacterial production/biomass or b) seawater incubation experiments, usually coupled with FISH or CARD-FISH
- rRNA:rDNA used as an index, doesn't work for oligotrophs
- Computational tools are not benchmarked using natural microbial communities

Lankiewicz et al. , 2016

Summary 1

1. CUB correlated with microbial growth rates, better than any other trait rRNA copy numbers.
2. PTR-based may be used to predict microbial genome replication in situ growth rates.
3. None of these methods has been seriously tested in oligotrophic natur like the marine environments.

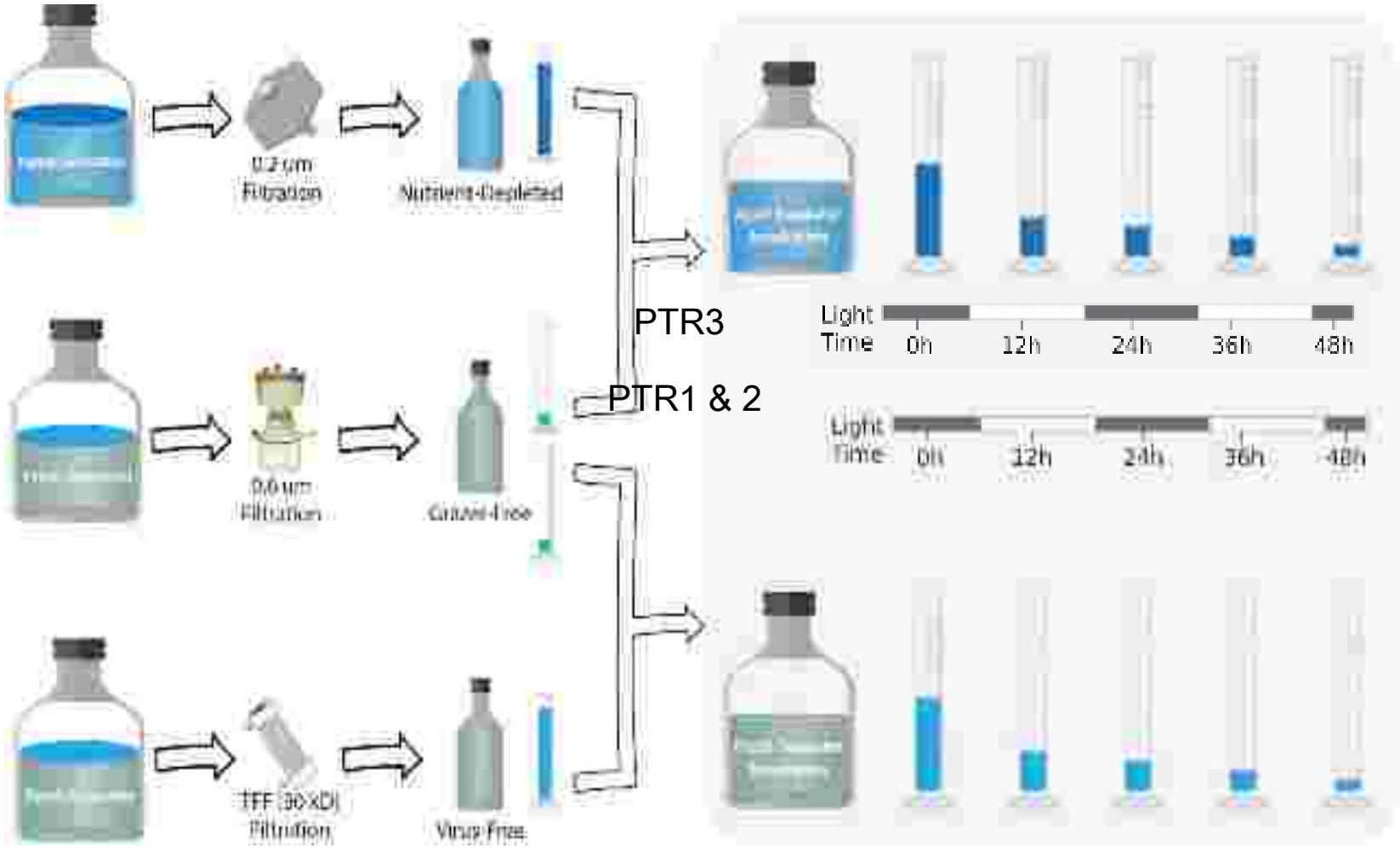


Part 2

Benchmarking growth rate methods

Experimental Design

海洋科学与工程系
Department of Ocean Science and Engineering

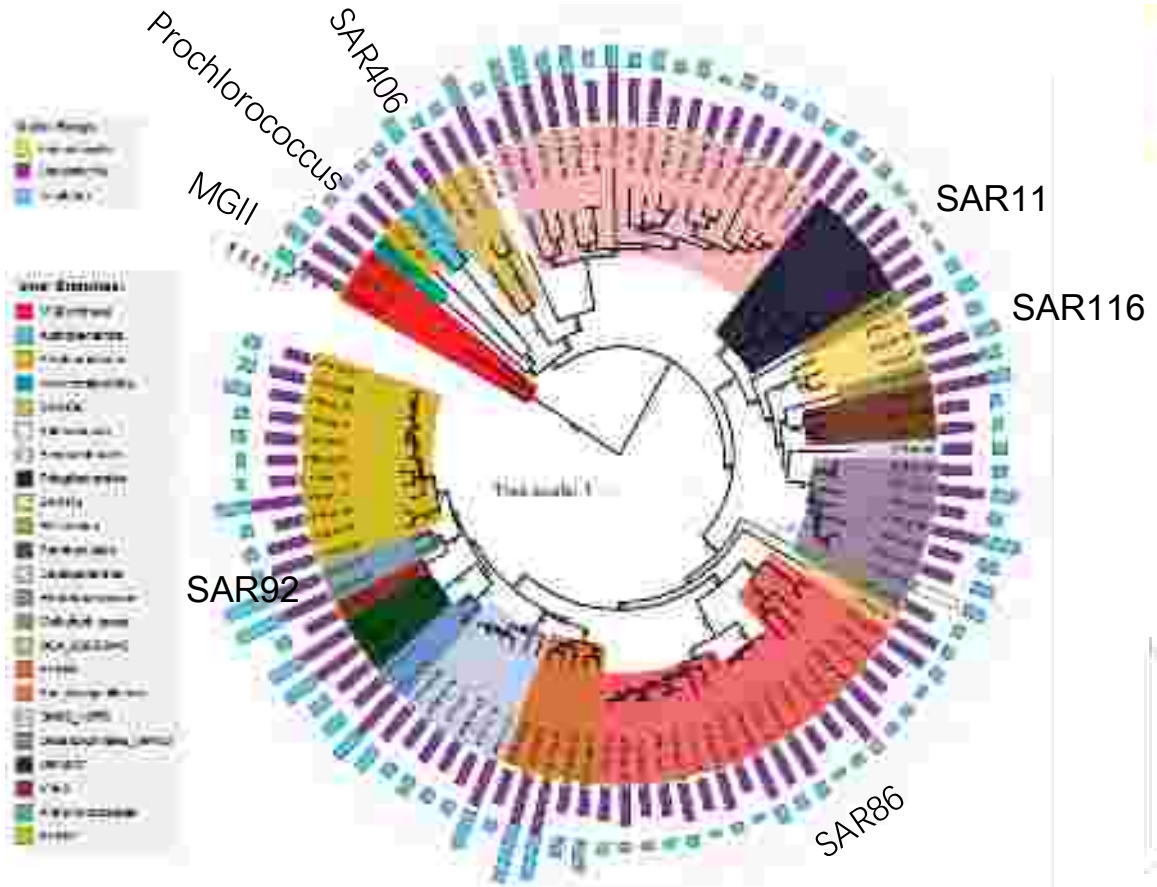


- Four incubation experiments performed in May (PTR1), Sept. (PTR2), and Dec. (PTR3A & PTR3V).
- Different nutrient conditions: aged and fresh seawater incubations (PTR3A & PTR3V)
- Grazers were removed from the Initial community, and 1:10 diluted to reduce virus infection
- Incubated for 48 hours under diurnal cycles (PTR3A & 3V), sampled every 12 hours

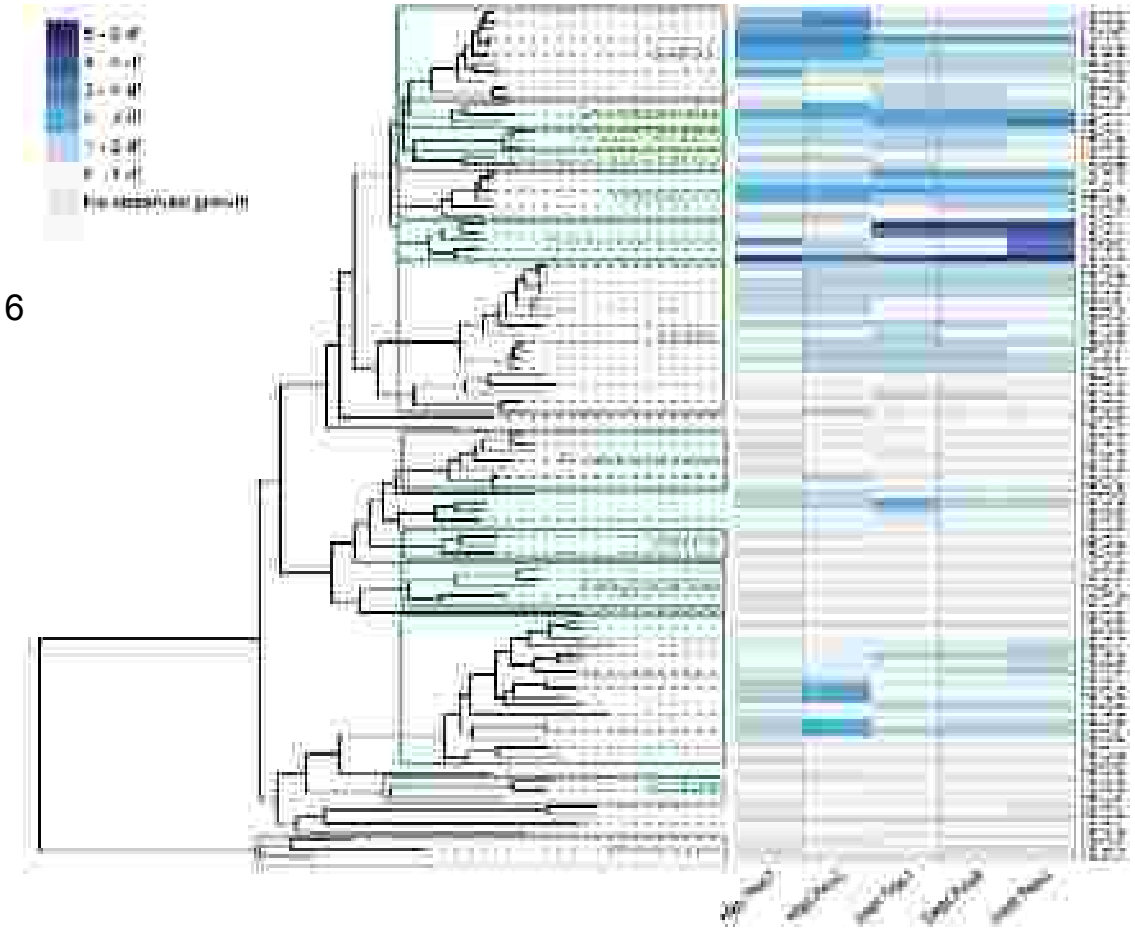
Long & Hou, 2020; Hou, unpublished

MAGs and Growth Rates

海洋科学与工程系
Department of Ocean Science and Engineering



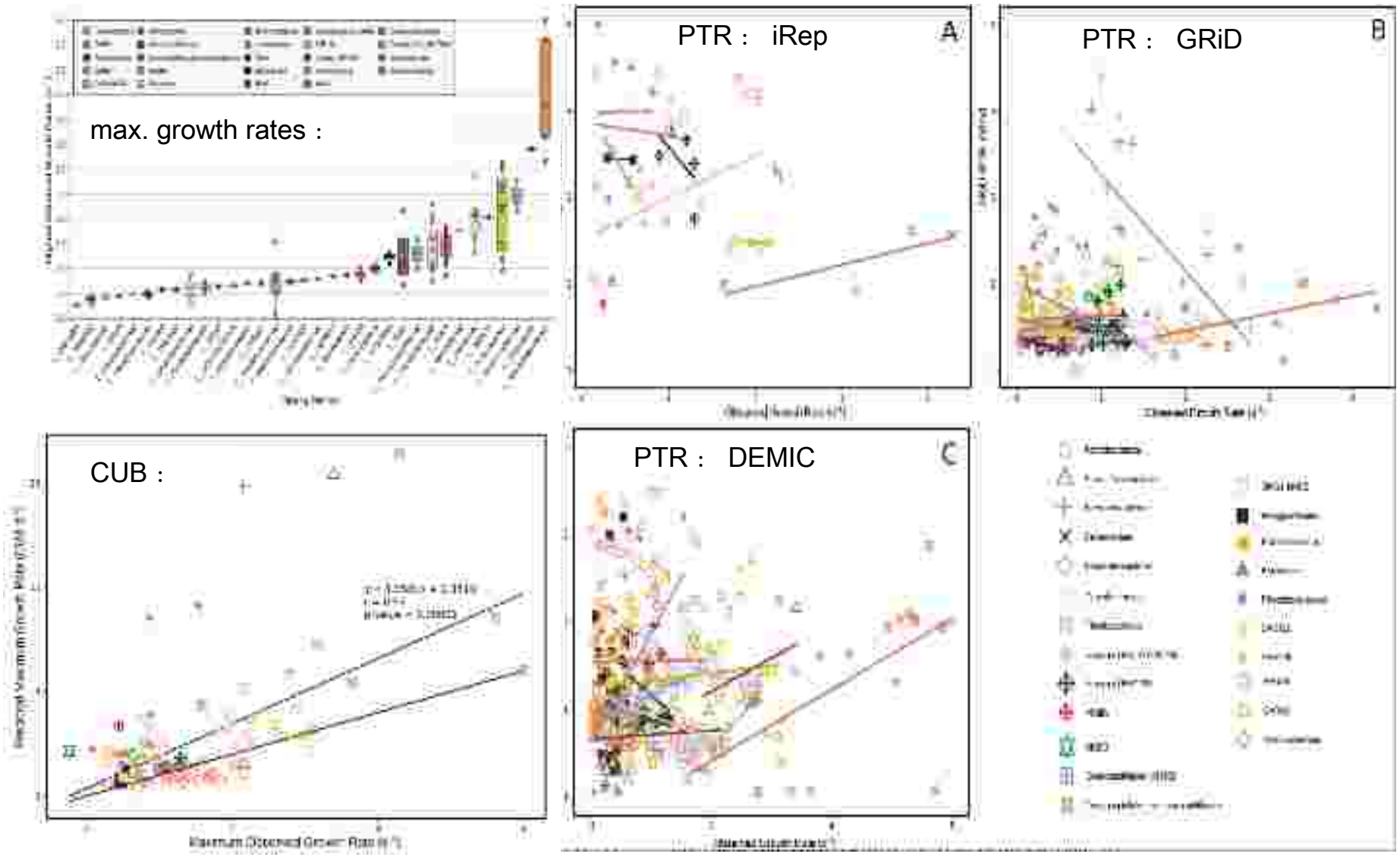
Long & Hou, 2020



- 101 draft MAGs, avg. 70% completion, 3% contamination, covering diverse marine taxa, including MGII, SAR406, SAR86, etc
- The fastest growing organisms were Oceanospirillales Saccharospirillaceae (3.17 – 5.99/day)

Long & Hou, 2020; Hou, unpublished

Benchmarks: PTR and CUB



- CUB works reasonably well in predicting maximum growth rates
- PTR rarely correlates with observed growth rates except for several more clonal fast-growing copiotrophs, such as Oceanospirillales and Sacchar ospirillaceae

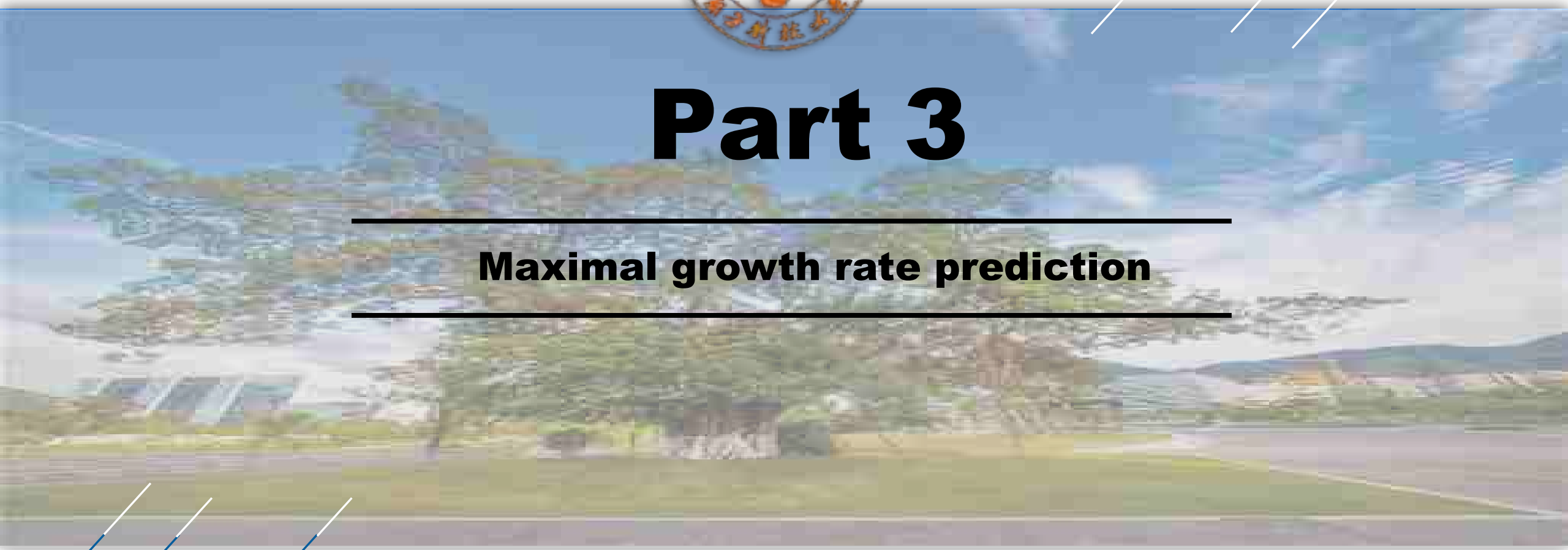
Long & Hou, 2020

Summary 2

1. The CUB-based method can reliably predict maximal growth rates.

2. Current PTR-based growth rate predictors are biased towards fast-grov

   copiotrophic microbes.

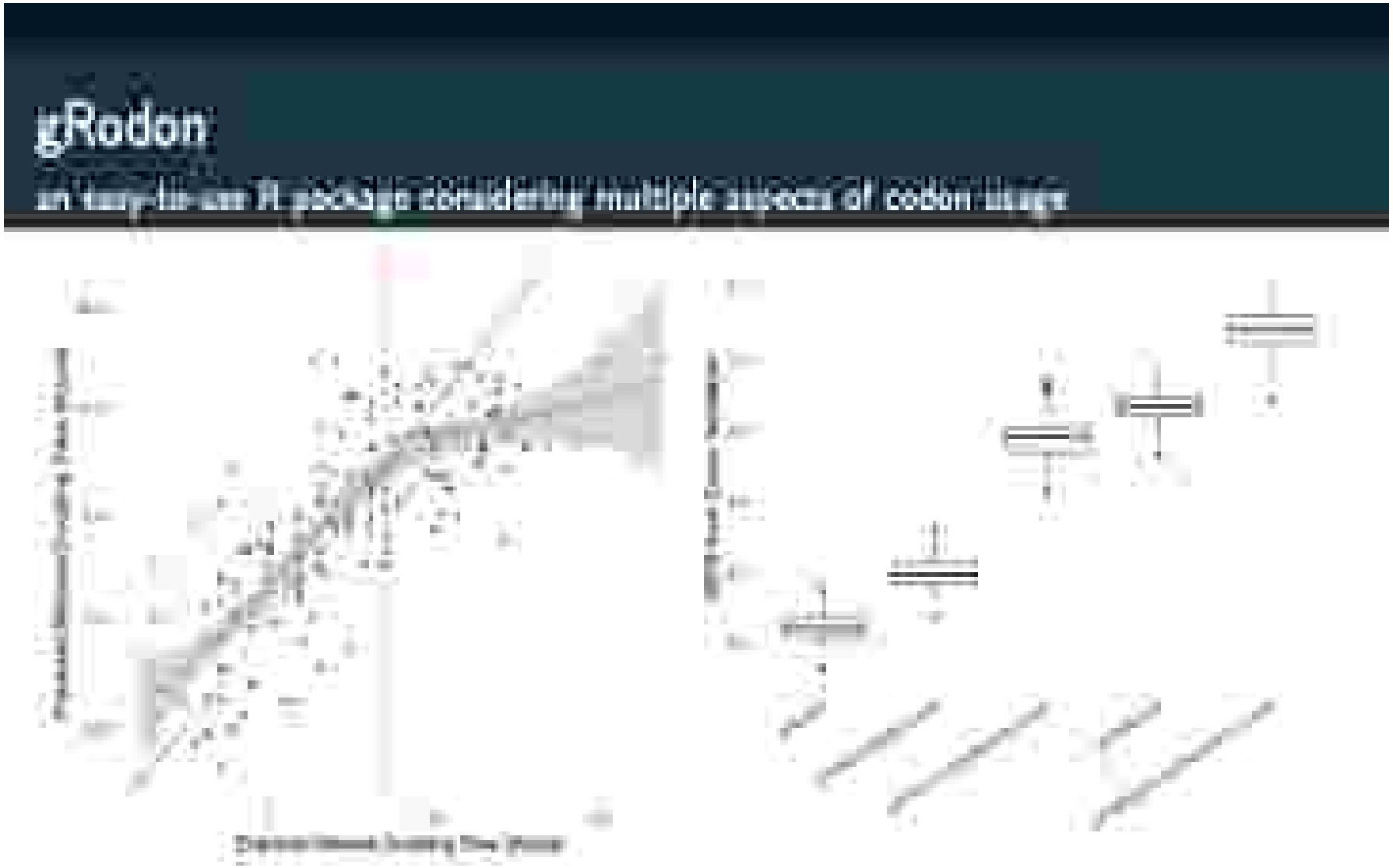


Part 3

Maximal growth rate prediction

gRodon vs growthpred

海洋科学与工程系
Department of Ocean Science and Engineering

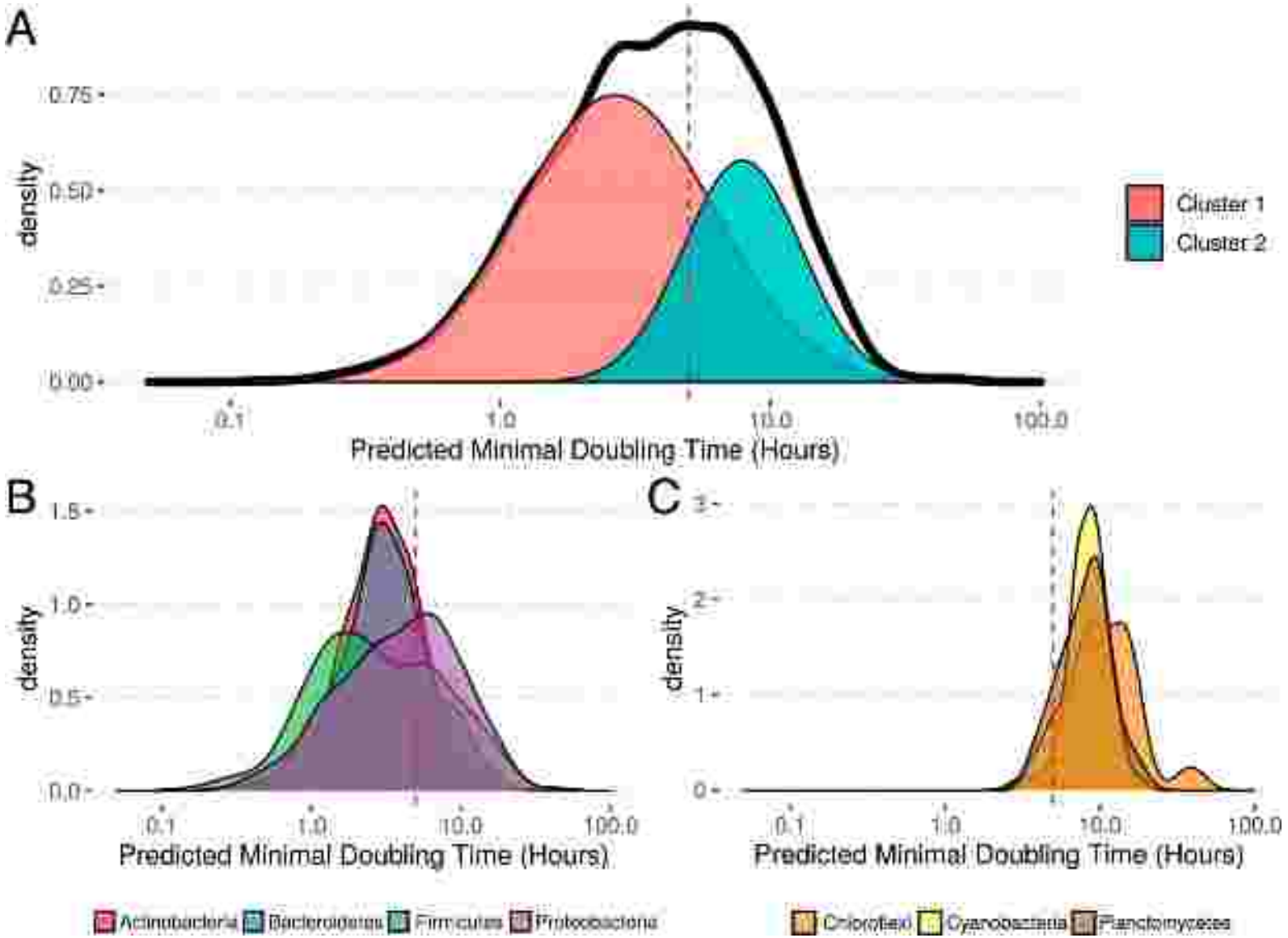


- We extend the work of Vieira-Silva and Rocha (9) by assessing additional dimensions of codon usage: the CUB of highly expressed genes, the "consistency" of bias across highly expressed genes, and the codon pair bias.
- The gRodon model fit the available maximal growth rate data well (adjusted $R^2 = 0.63$), providing a better fit to the data than growthpred at both high and low growth rates.
- gRodon underestimates doubling times when growth is very slow due to a floor on CUB reached in slow-growth regimes.

Weissman & Hou,, 2021 PNAS

Oligotrophs and copiotrophs

海洋科学与工程系
Department of Ocean Science and Engineering

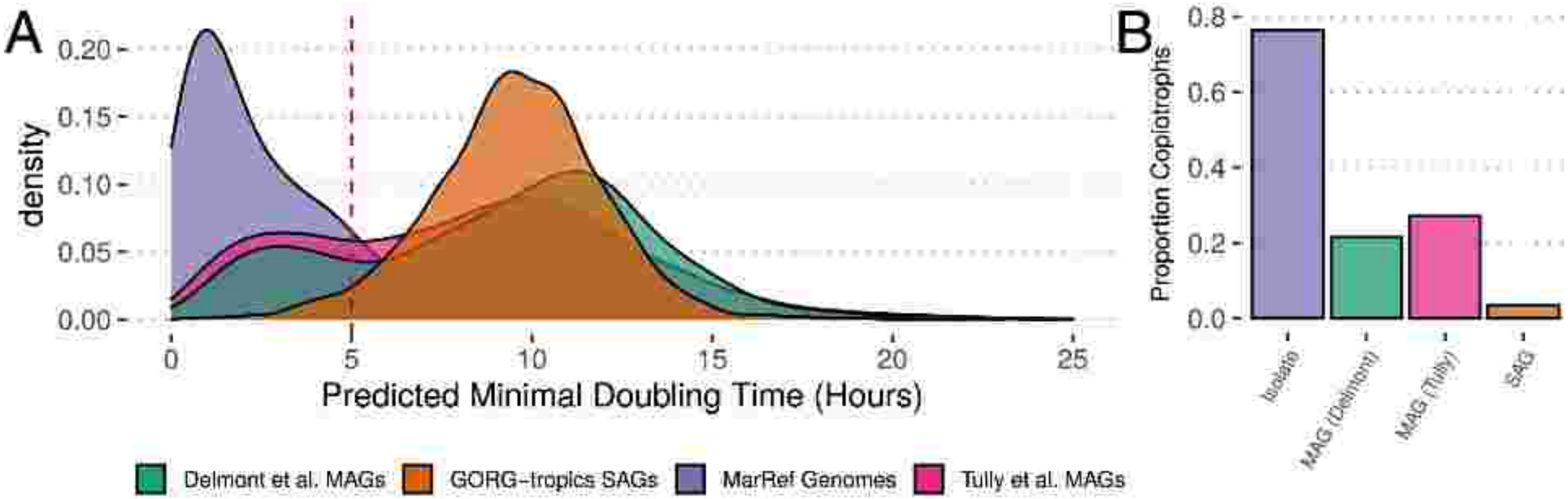


- The distribution of growth rates across RefSeq was roughly bimodal, with the split between peaks corresponding to the 5-h doubling-time cutoff we proposed above for classifying oligotrophs.
- Phyla tended to broadly group together in terms of growth rate, and the 5-h divide separated fast- and slow-growing phyla.
- Using a Gaussian mixture model, we obtained two large clusters of microbes, with mean doubling times of 2.7 and 7.9 h, respectively, roughly corresponding to our proposed copiotroph/oligotroph divide.

Weissman & Hou,, 2021 PNAS

Oligotrophs and copiotrophs

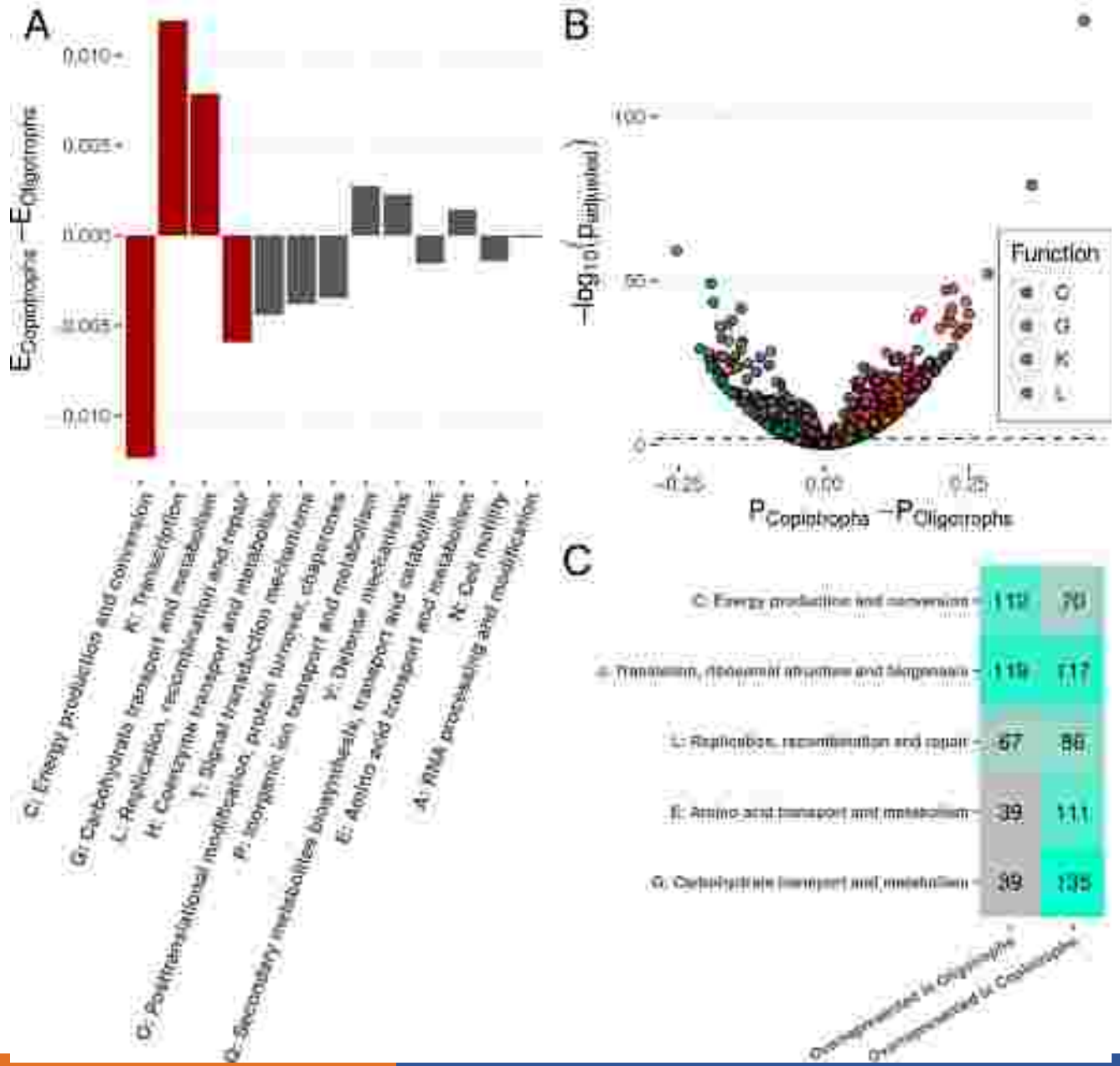
海洋科学与工程系
Department of Ocean Science and Engineering



- Isolate collections showing much shorter predicted doubling times than MAGs and SAGs from the same environments, suggesting that existing culture collections are not only incomplete but also biased.
- SAGs showed a lower overall growth rate than MAGs, with very few doubling times predicted to be under 5 h, likely due in part to how SAGs were sampled (only at the ocean surface rather than at multiple depths).
- MAGs generated by distinct research groups showed surprisingly consistent maximal growth rate distributions.

Weissman & Hou,, 2021 PNAS

Oligotrophs and copiotrophs

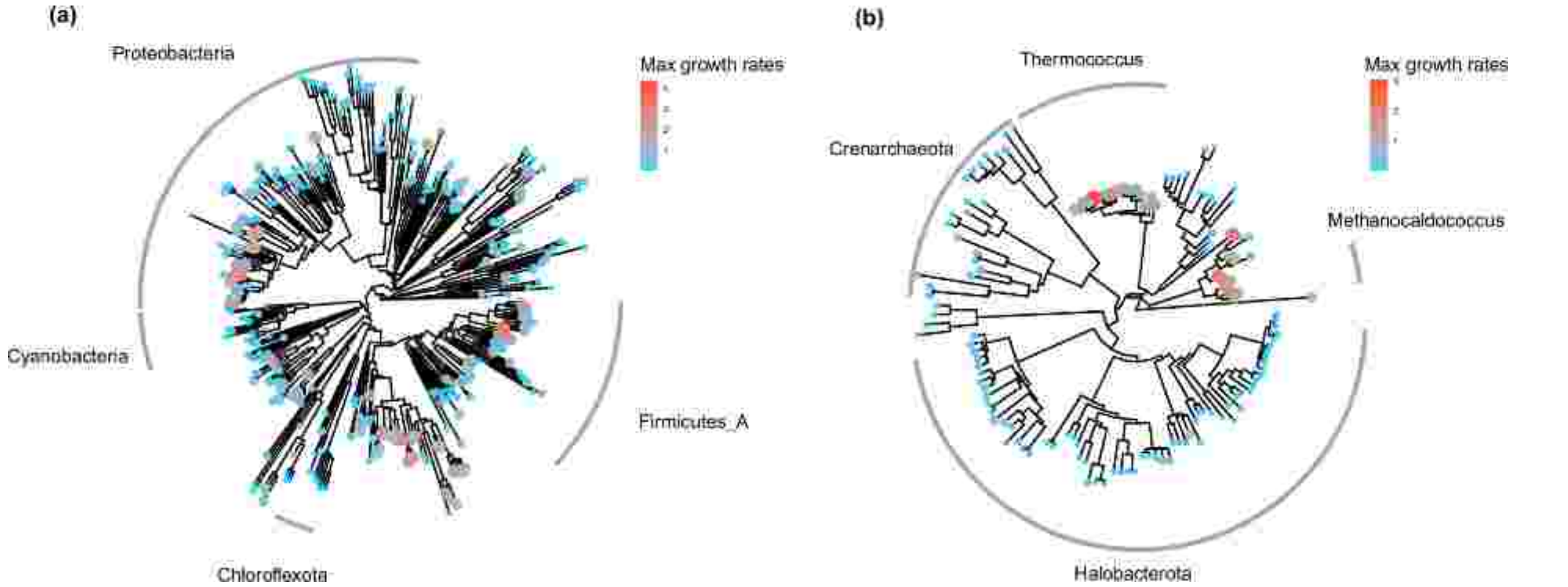


- Gene families involved in transcription and carbohydrate transport and metabolism were strongly overrepresented on copiotroph genomes relative to oligotrophs, corresponding to an overall strategy of rapid acquisition of nutrients and protein production.
- Gene families involved in energy production and conversion and replication, recombination, and repair were overrepresented on oligotroph genomes, corresponding to an overall strategy of energy production and cell maintenance.
- Gene families involved in the transport and metabolism of carbohydrates and amino acids were frequently more prevalent among copiotrophs. Meanwhile, gene families involved in energy production and conversion were more prevalent among oligotrophs.
- Copiotrophs were greatly enriched for gene families involved in breaking down carbohydrates.

Weissman & Hou,, 2021 PNAS

Phydon, with phylogenetic information

海洋科学与工程系
Department of Ocean Science and Engineering

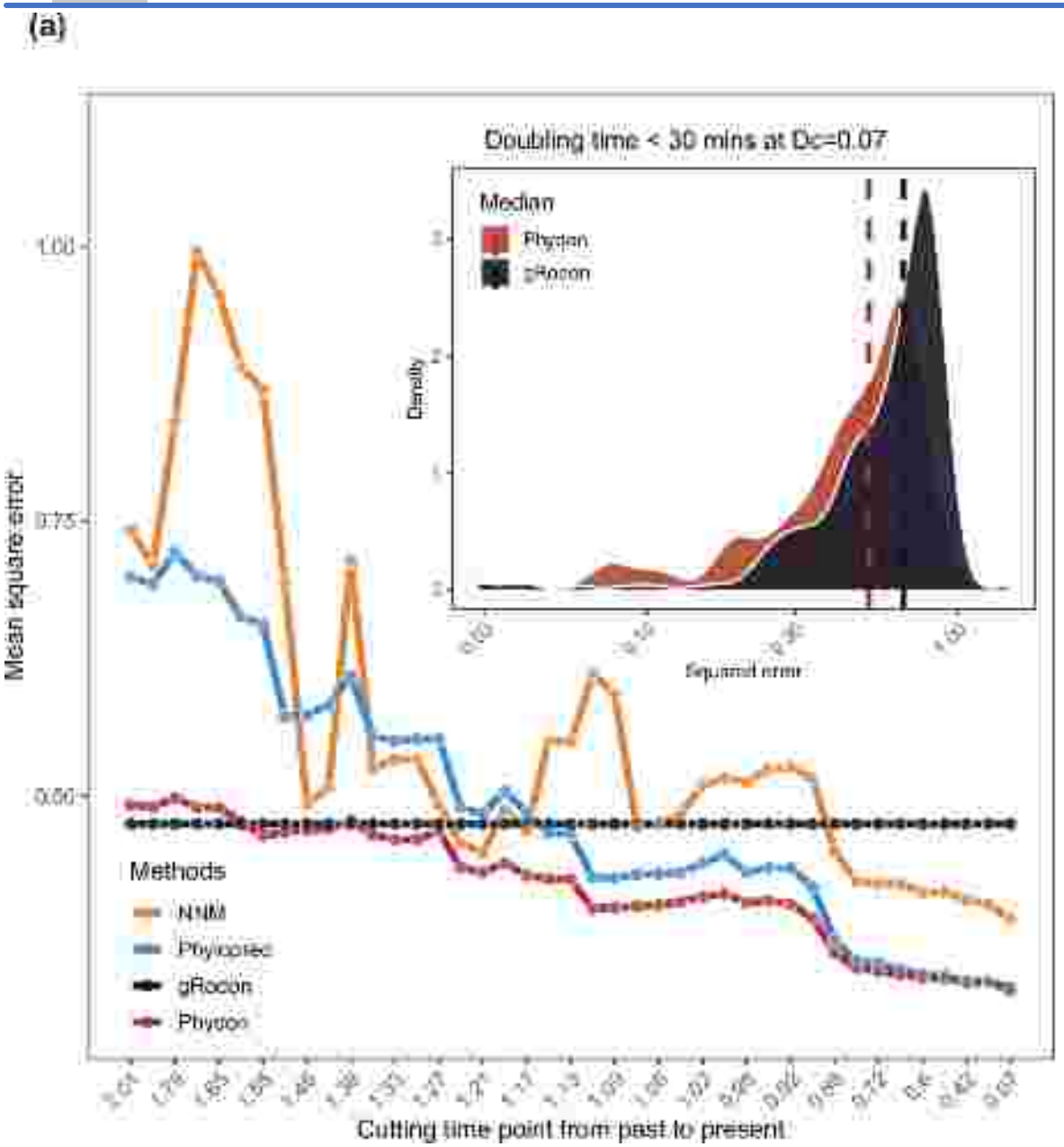


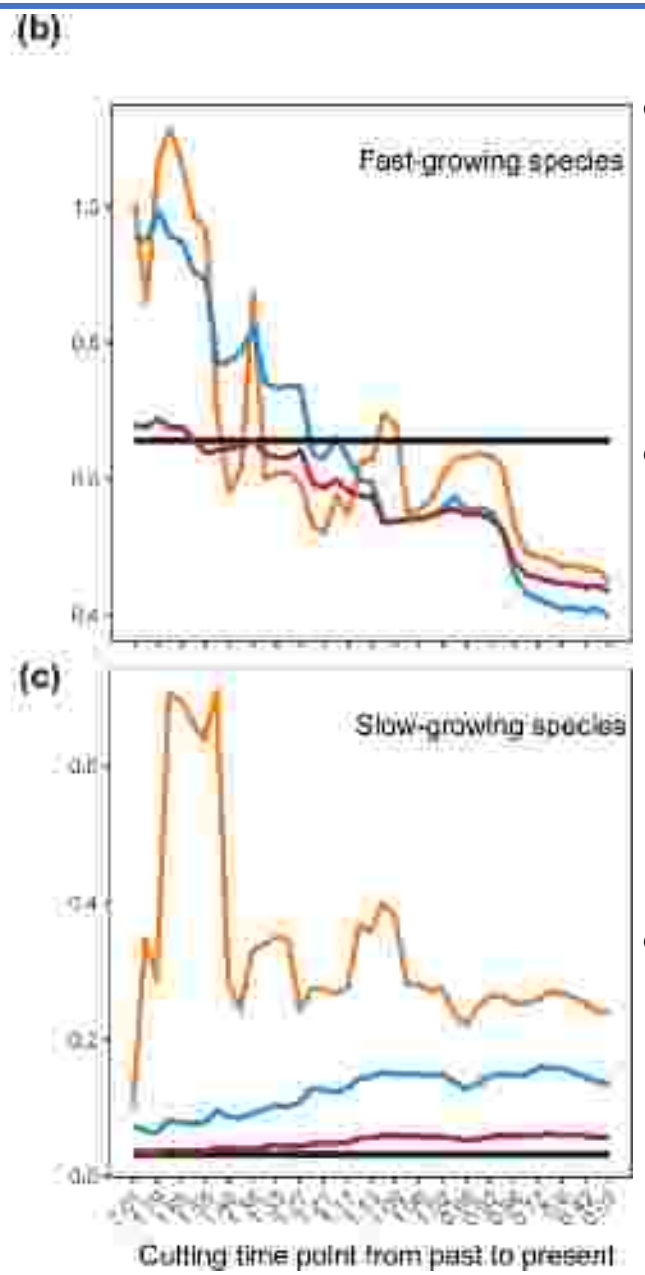
- Phydon, a framework for genome-based maximum growth rate prediction that combines codon statistics and phylogenetic information.
- Phylogenetically conserved patterns in the maximum growth rates of bacteria (a) and archaea (b). Among bacteria, certain groups within Proteobacteria and Firmicutes_A exhibit higher growth rates, while in archaea, the genera Thermococcus and Methanocaldococcus are notable for their rapid growth.

Xu et al., 2025 Nat Comm

Phydon, with phylogenetic information

海洋科学与工程系
Department of Ocean Science and Engineering



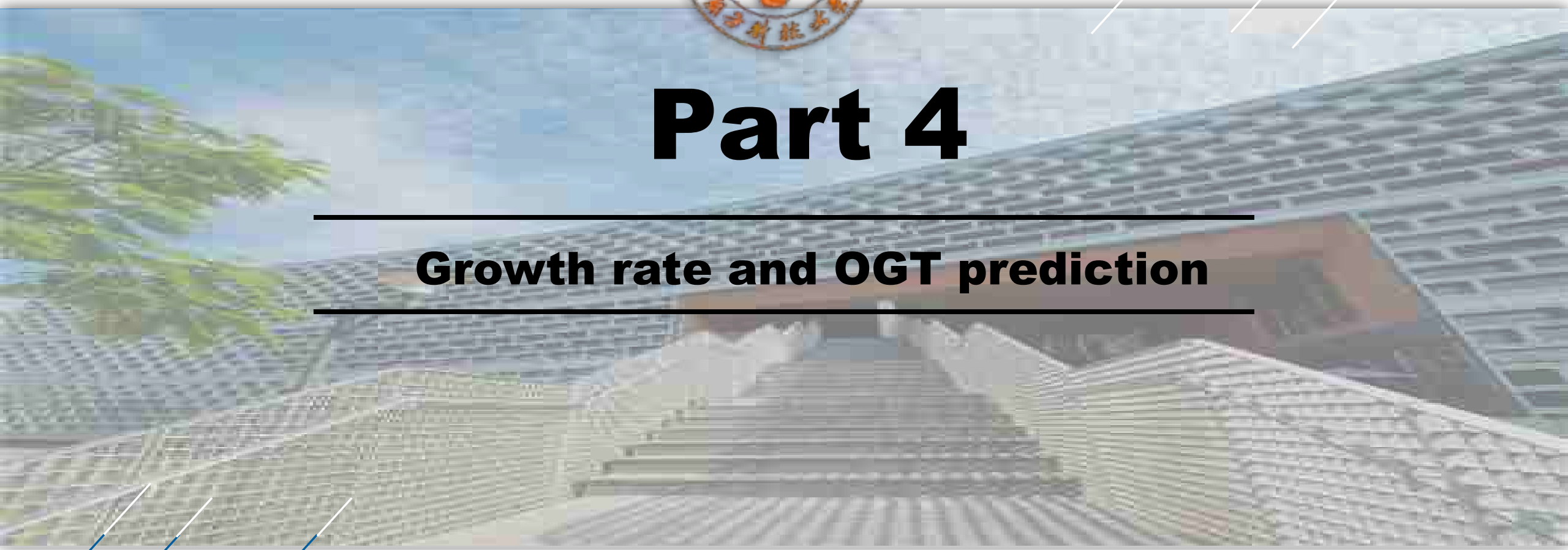


- Phylogenetic prediction methods show increased accuracy as the minimum phylogenetic distance between the training and test sets decreases.
- For fast-growing species, phylogenetic relationships capture selective signals more effectively than codon usage bias, which is inherently limited as a genomic statistic for explaining maximum growth rates.
- However, both CUB- and phylogenetic distance-based methods face challenges in predicting traits for slow-growing species.

Xu et al., 2025 Nat Comm

Summary 3

1. Culture collections are strongly biased toward organisms capable of ra
2. Microorganisms can be naturally grouped into two growth classes, cor
   to oligotrophs and copiotrophs, based on the selective regime an organ
   occupies.
3. The growth rates of slow-growing microbes are challenging to predict
   combining CUB and phylogenetic distance-based methods.

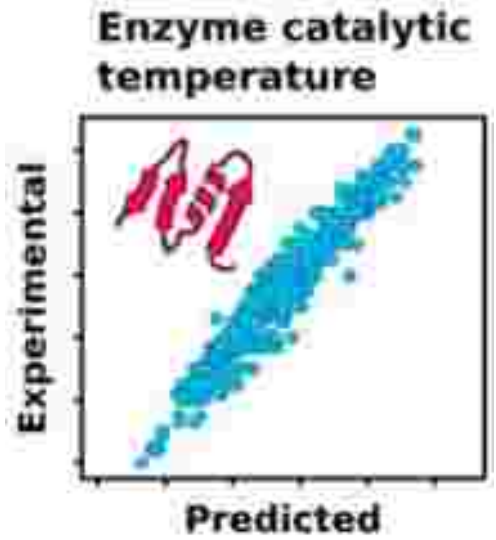


Part 4

Growth rate and OGT prediction

Growth rate and OGT

- To reliably predict microbial growth rates, we have turned to omic approaches to estimate microbial (maximal) growth rates and optimal growth temperatures (OGT).
- As for optimal growth temperature predictions, several methods based on Machine Learning have achieved great performance E.g. Tome, CnnPOGTP. Enzyme catalytic temperature is the most reasonable feature correlating with OGT.



https://doi.org/10.1021/acssynbio.9b00099

Tools name	Methods	Parameters	Output	Performance	Availability	Publishment
not provided	naïve Bayesian inference	40 protein families	OGT range	R^2=0.462	python scripts in supplementary materials	Jensen, et al., 2012
OGT_prediction	multiple linear regression	tRNA, rRNA, ORF and proteome sequences	OGT value	R^2=0.835, RMSE=5.17	github.com/DanielBSauer/OGT_prediction	Sauer and Wang, 2019
Tome	support vector regression	genome-wide 2-mer amino acid composition	OGT value	R^2=0.922	github.com/EngqvistLab/Tome	Li, et al., 2019
CnnPOGTP	convolutional neural network	2-mer	OGT value	R^2=0.880, RMSE=7.54	www.orgene.net/CnnPOGTP	Wang, et al., 2022

https://doi.org/10.1093/bioinformatics/btac289

Growth rate and OGT

海洋科学与工程系
Department of Ocean Science and Engineering

Temperature Data: Two published articles[1,2], BacDive, etc.

Growth rate: EGGO[3], Madin et al.[4]

Genome: NCBI, RefSeq



Data Cleaning

Quality Check

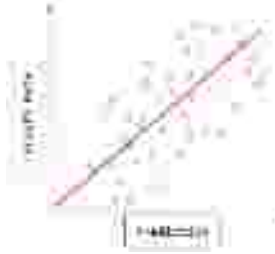
Annotation

Feature Extraction

Final Data

Exploratory Data Analysis

Explanation



Machine Learning

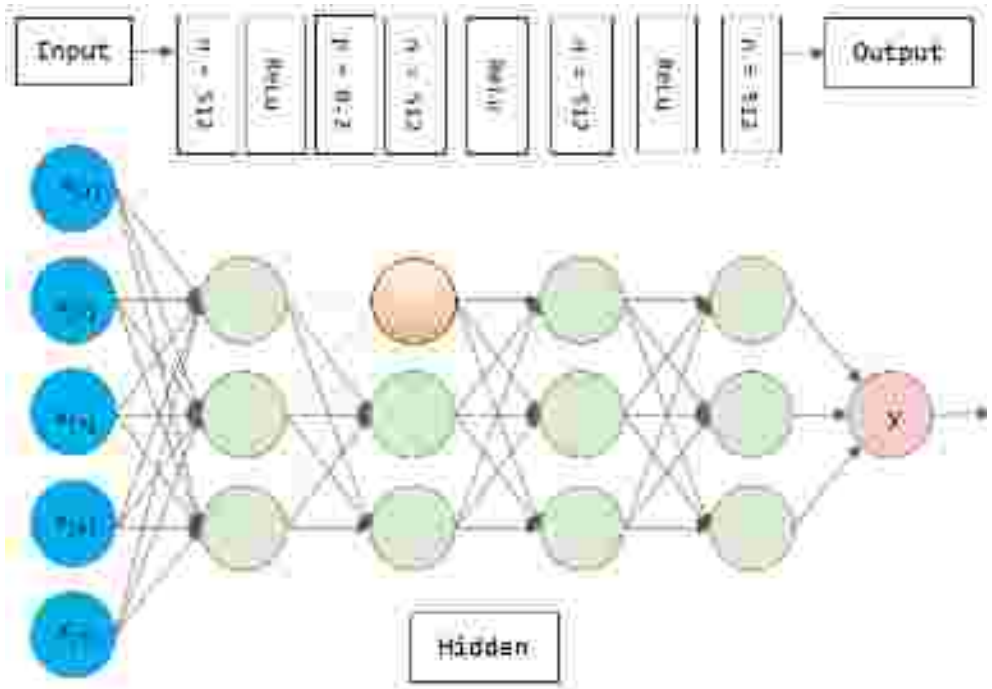
1. https://doi.org/10.1021/acssynbio.9b00099
2. https://doi.org/10.1186/s12866-018-1320-7
3. https://doi.org/10.1073/pnas.2016810118
4. https://doi.org/10.1038/s41597-020-0497-4

Xia et al., unpublished

Methods and models

海洋科学与工程系
Department of Ocean Science and Engineering

MLP construction



XGBoost

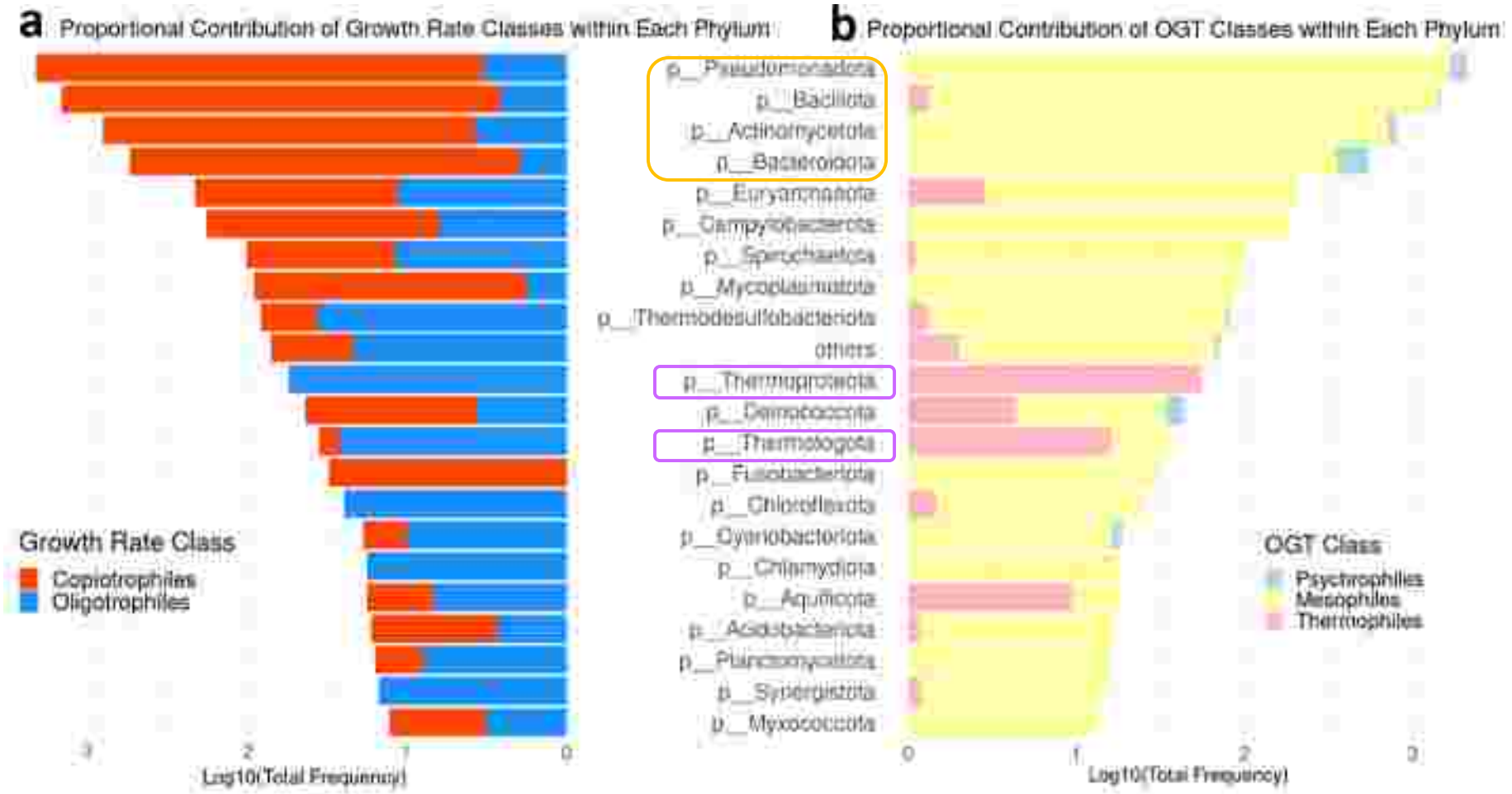
N_estimators = 100
Max_depth = 6
Subsample = 1
Lr = 0.1
Reg_lambda = 0.5

Xia et al., unpublished

Methods and models

海洋科学与工程系
Department of Ocean Science and Engineering

Distribution of maximal growth rate and optimal growth temperature (OGT) within each phylum



These 4 dominant phyla are mainly mesophiles and copiotrophiles.

Thermopoteota and Thermotogota are mainly thermophiles and oligotrophiles.

Psychrophic samples are underrepresented.

Others: the phyla contain < 10 samples.
Maximal growth rate (doubling time, hour): 0-5 (Copiotrophiles), >5 (Oligotrophiles).
OGT (℃): 0-15 (Psychrophiles), 15-60 (Mesophiles), >60 (Thermophiles).

Xia et al., unpublished

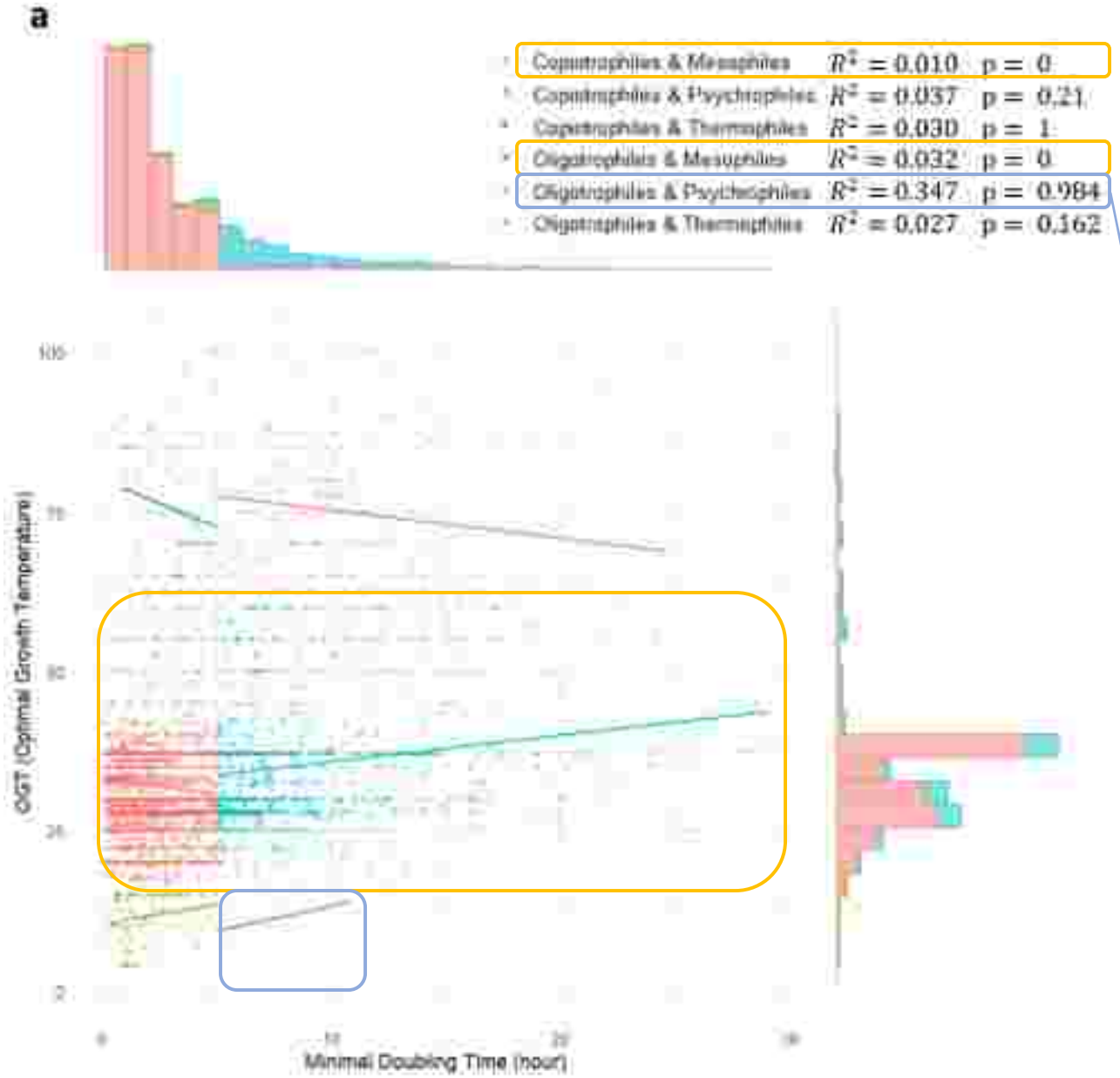
Methods and models

海洋科学与工程系
Department of Ocean Science and Engineering

We combined the maximal growth rate classes (2) and OGT classes (3) to get 6 combinations.

Then, we tested the linear relationship between maximal growth rate vs. OGT.

The significance (adjusted *p*-value) and effect size (R^2) of each combination are shown inset.



Two groups of Mesophiles have significant correlation between max growth rate vs. OGT. But both of them have extremely small effect size. (Mainly because the sample size of mesophiles is very huge)

Oligotrophiles & Psychrophiles group has a medium effect size but the correlation is not significant. (Too small sample size)

Therefore, no significant and great effect size was found in all groups.

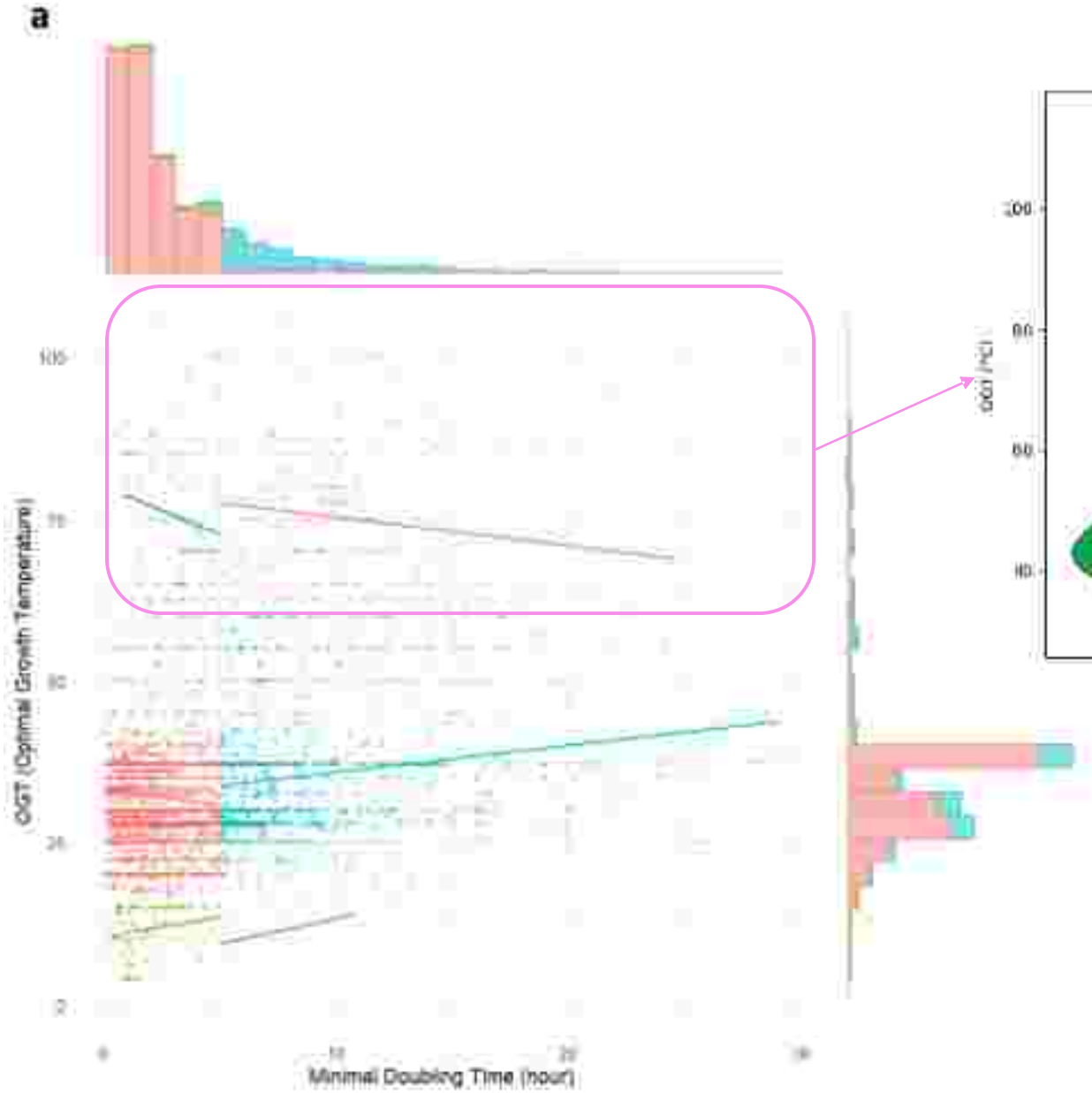
Xia et al., unpublished

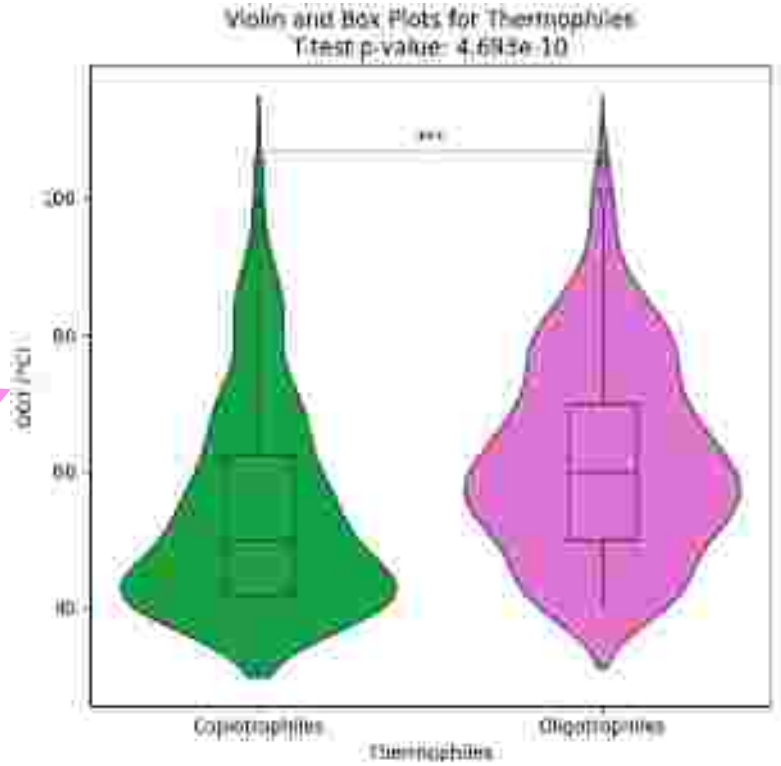
Methods and models

海洋科学与工程系
Department of Ocean Science and Engineering

We combined the maximal growth rate classes (2) and OGT classes (3) to get 6 combinations.

Then, we tested whether the max growth rates and OGTs are different in each group.





The means of OGT in two groups of thermophiles are significantly different.

The mean of OGT is higher in oligotrophiles.

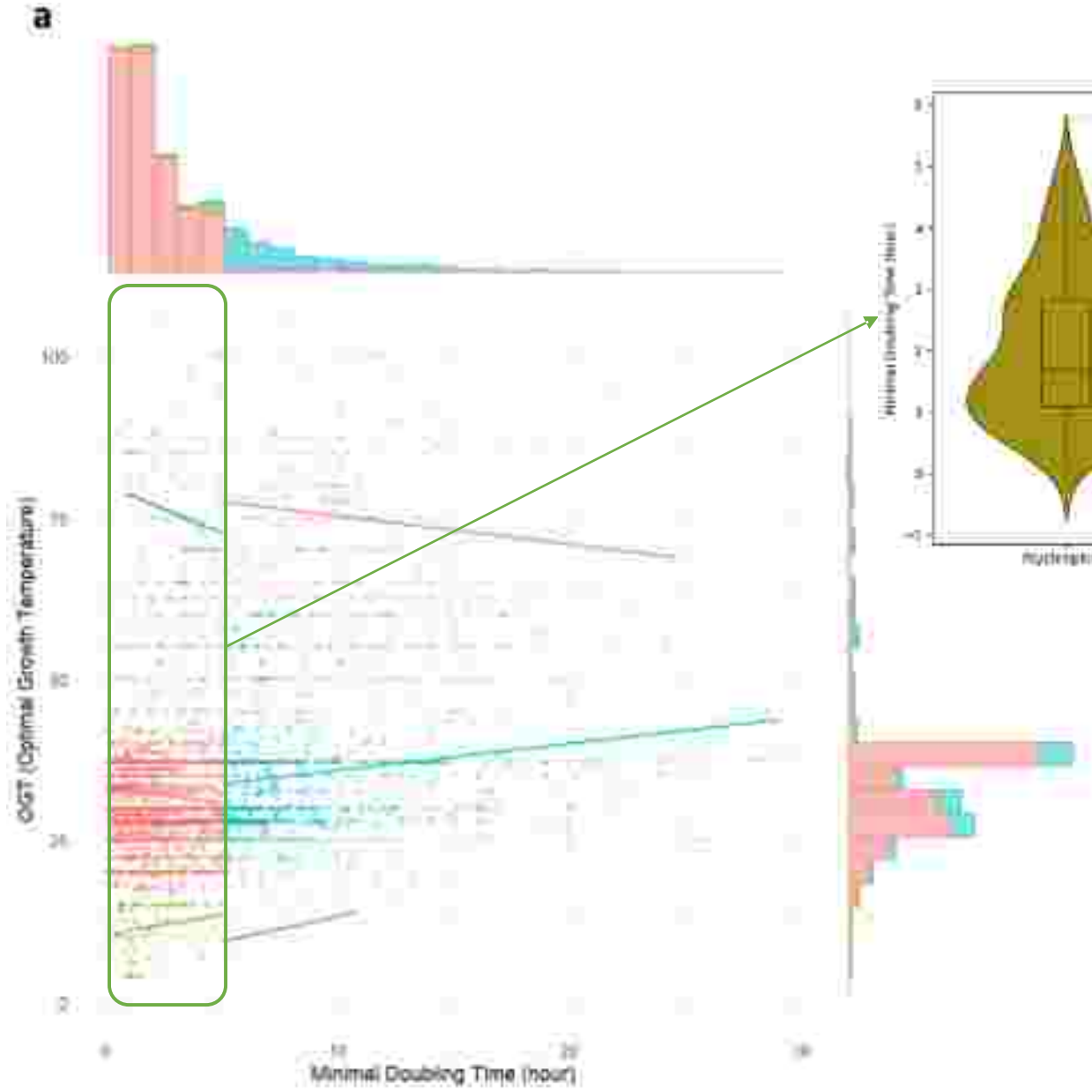
Xia et al., unpublished

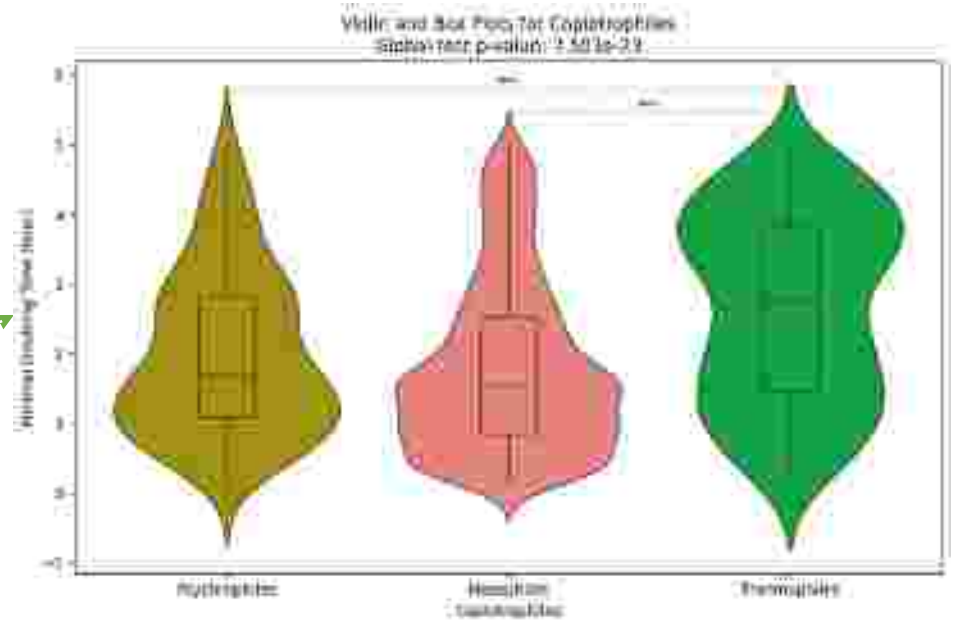
Methods and models

海洋科学与工程系
Department of Ocean Science and Engineering

We combined the maximal growth rate classes (2) and OGT classes (3) to get 6 combinations.

Then, we tested whether the max growth rates and OGTs are different in each group.





The mean of max growth rate in Copiotrophiles & Thermophiles is significantly lower than the other two groups.

While the means of max growth rate between psychrophiles and mesophiles are not significantly different.

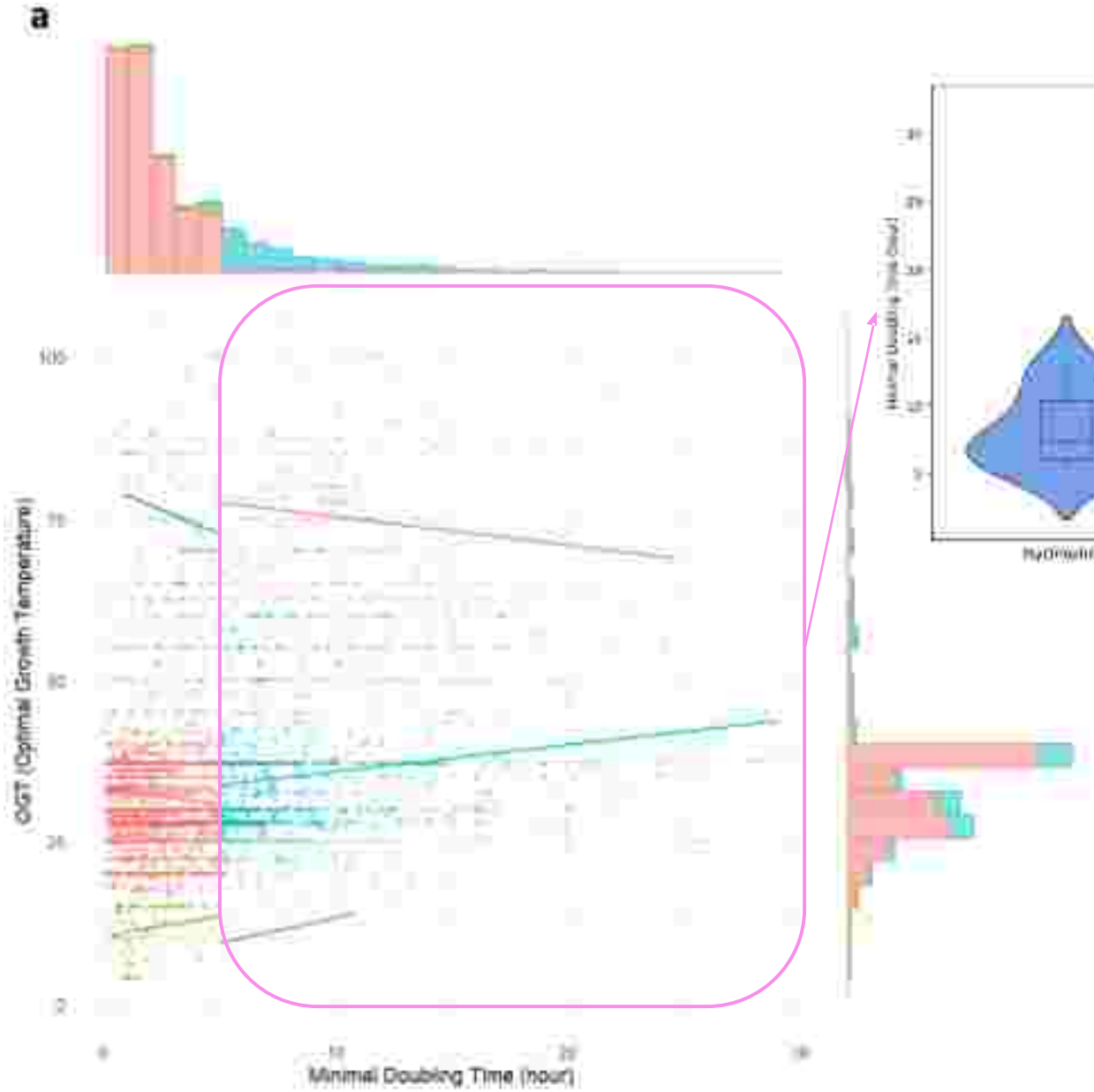
Xia et al., unpublished

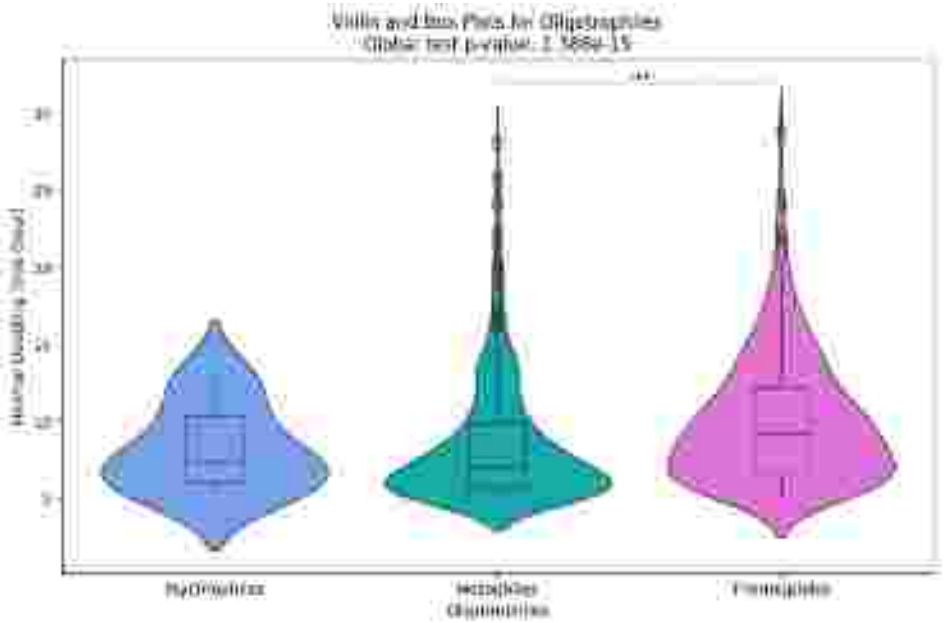
Methods and models

海洋科学与工程系
Department of Ocean Science and Engineering

We combined the maximal growth rate classes (2) and OGT classes (3) to get 6 combinations.

Then, we tested whether the max growth rates and OGTs are different in each group.





The median of max growth rate in Oligotrophiles & Thermophiles is significantly lower than mesophiles.

Therefore, in both copiotrophic and oligotrophic groups, thermophiles have significant lower growth rate.

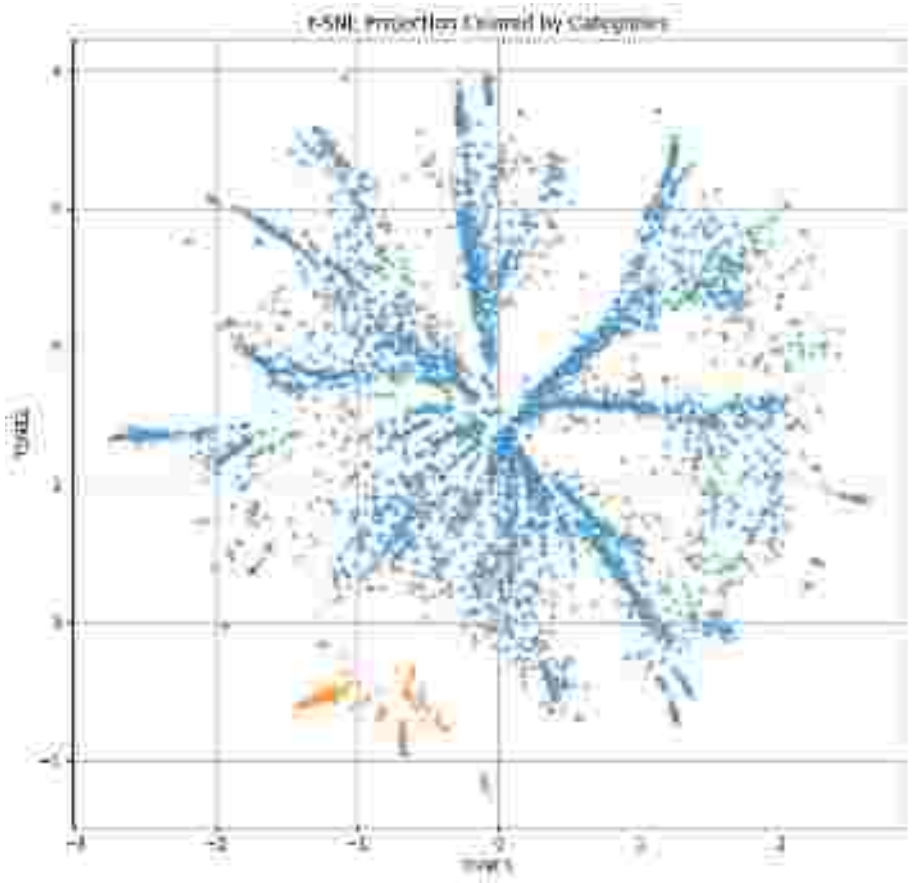
Xia et al., unpublished

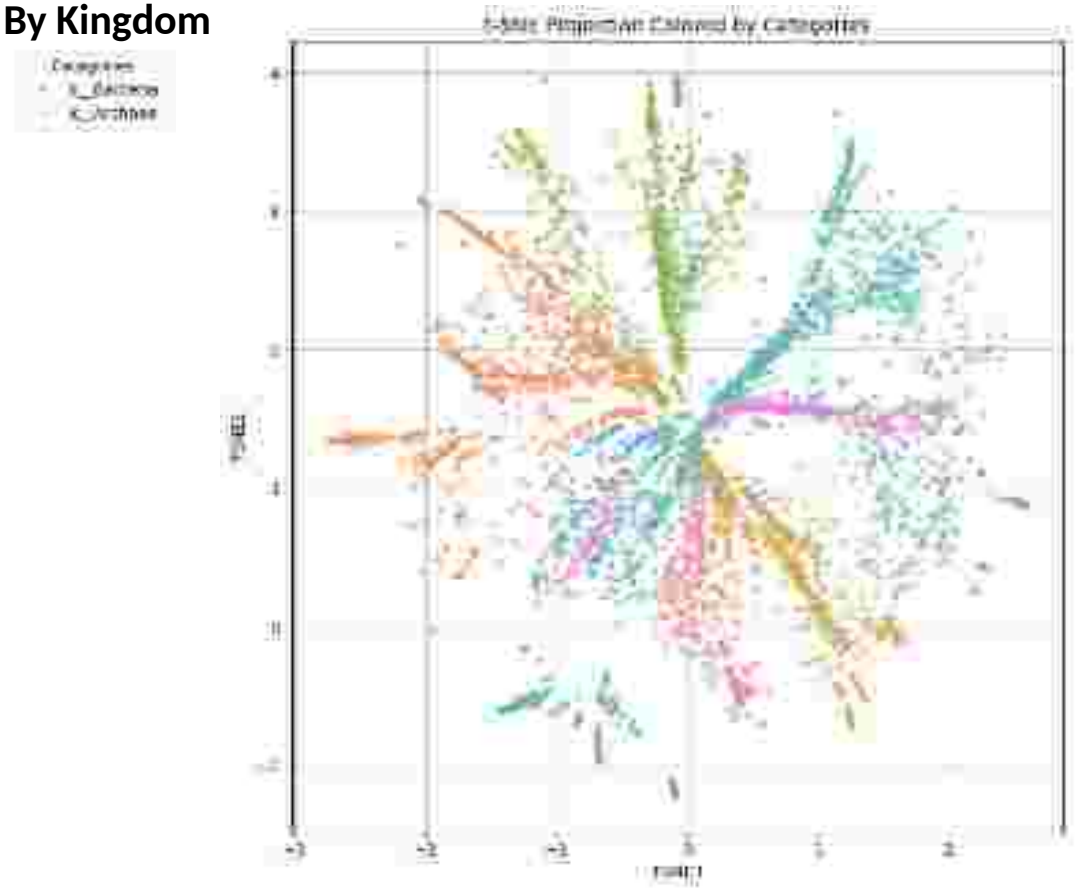
Dimensional reduction

海洋科学与工程系
Department of Ocean Science and Engineering

We employed dimensional reduction algorithm (PCA, tSNE and UMAP) on our dataset. They had similar results, which successfully clustered by different taxonomies. Many features of our data are protein annotation. Practically, they are used in classification.

By Kingdom





By Class

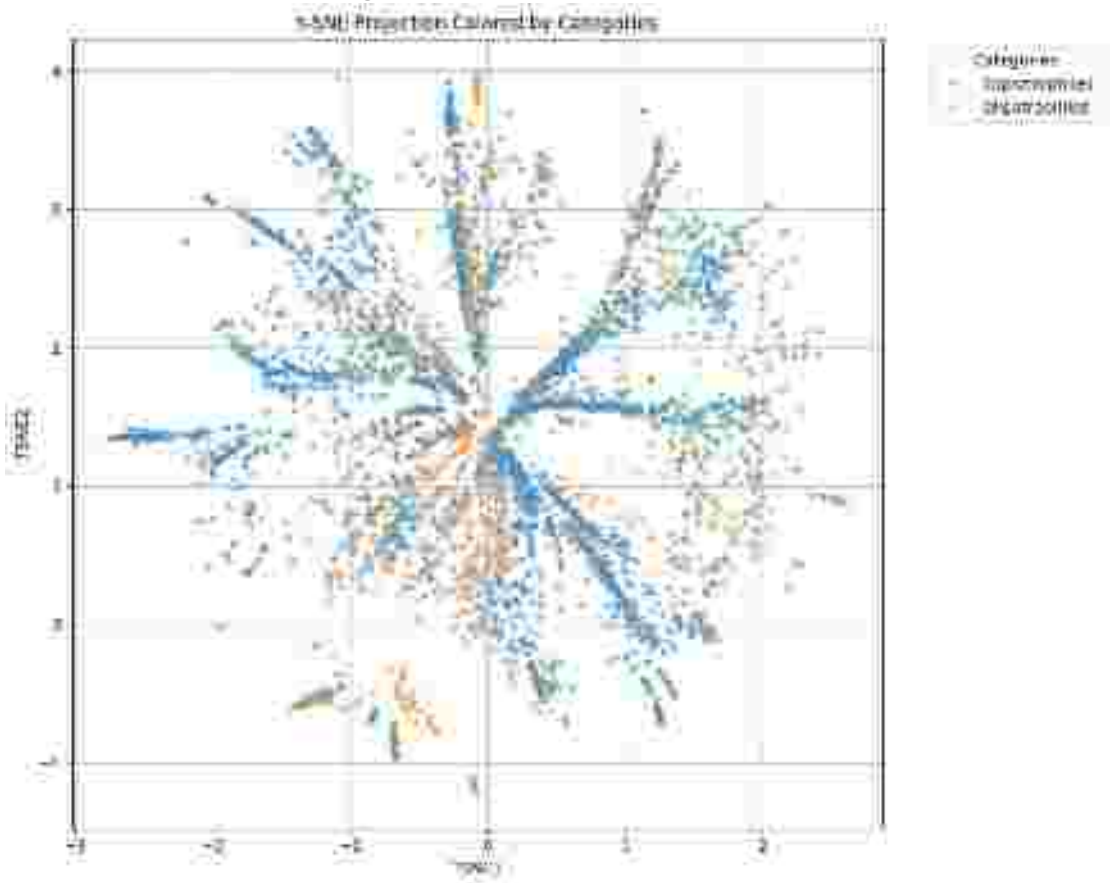


Xia et al., unpublished

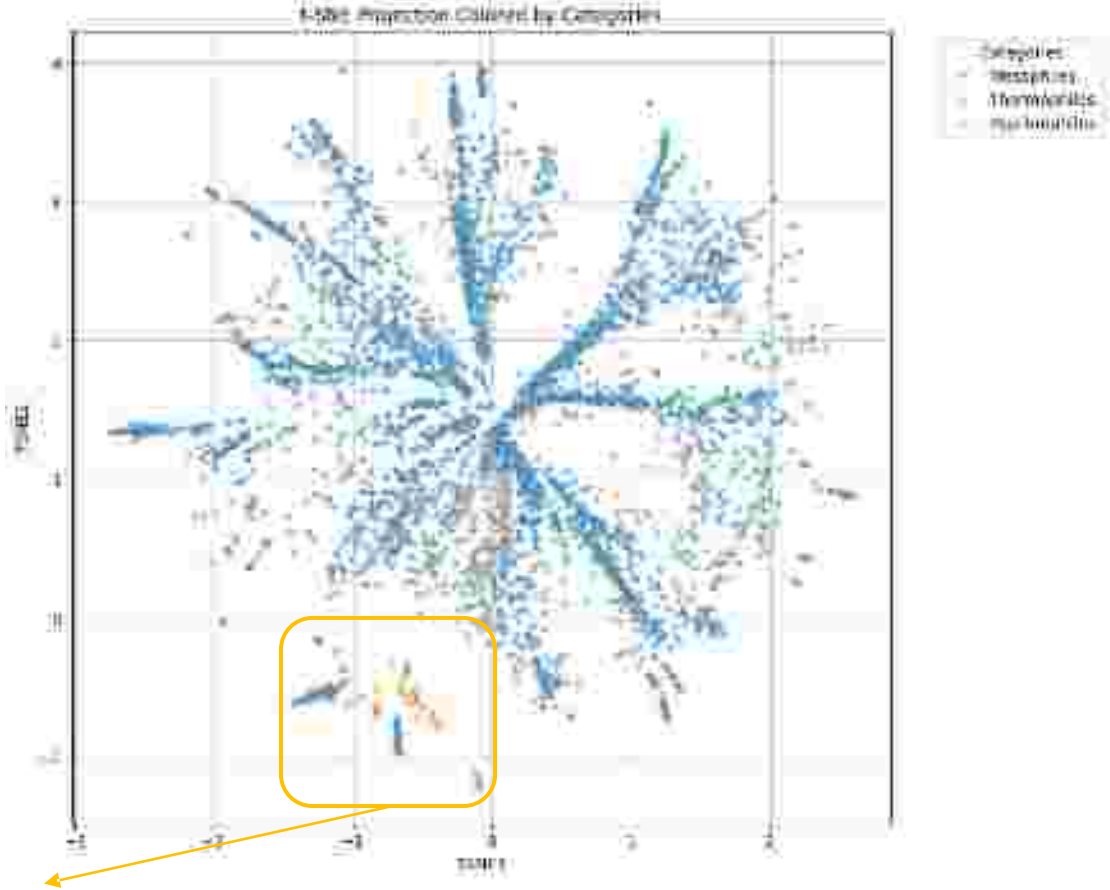
Dimensional reduction

However, it's difficult to cluster max growth rate groups or OGT groups.

Max growth rate class



OGT class



Certain archaeal samples were clustered in Thermophiles group.

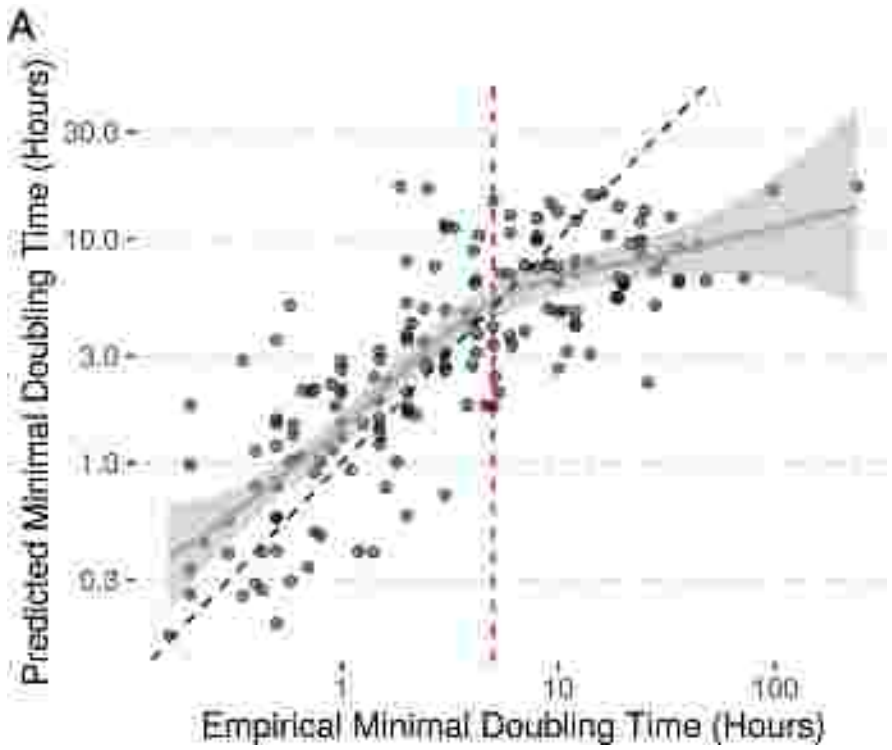
Xia et al., unpublished

Maximal growth rate prediction

海洋科学与工程系
Department of Ocean Science and Engineering

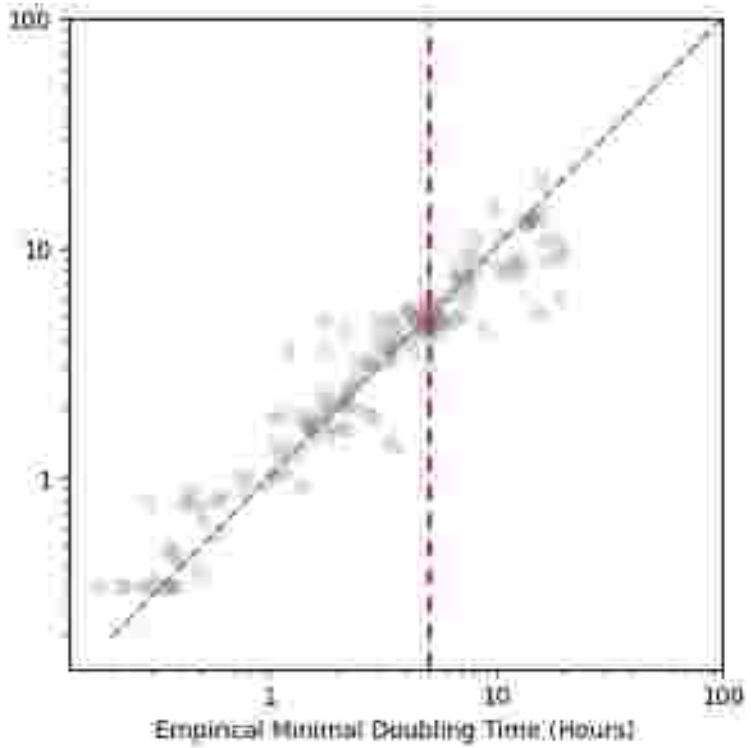
Several machine learning models were trained, and evaluated with the gRodon model. Our XGBoost models performed pretty well.

gRodon (2021), R^2 = 0.63

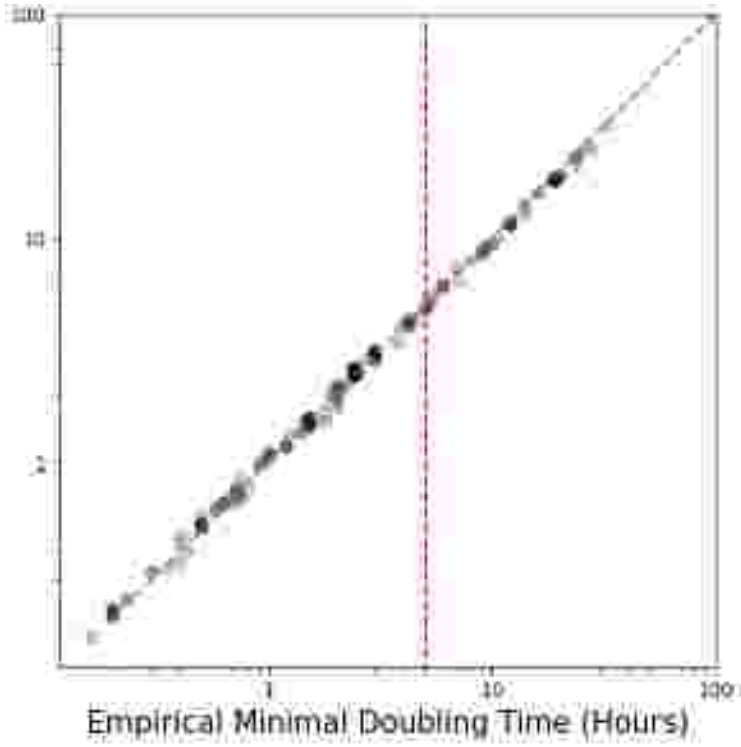


https://doi.org/10.1073/pnas.2016810118

Our best MLP model, R^2 = 0.76



Our best XGBoost model, R^2 = 0.99



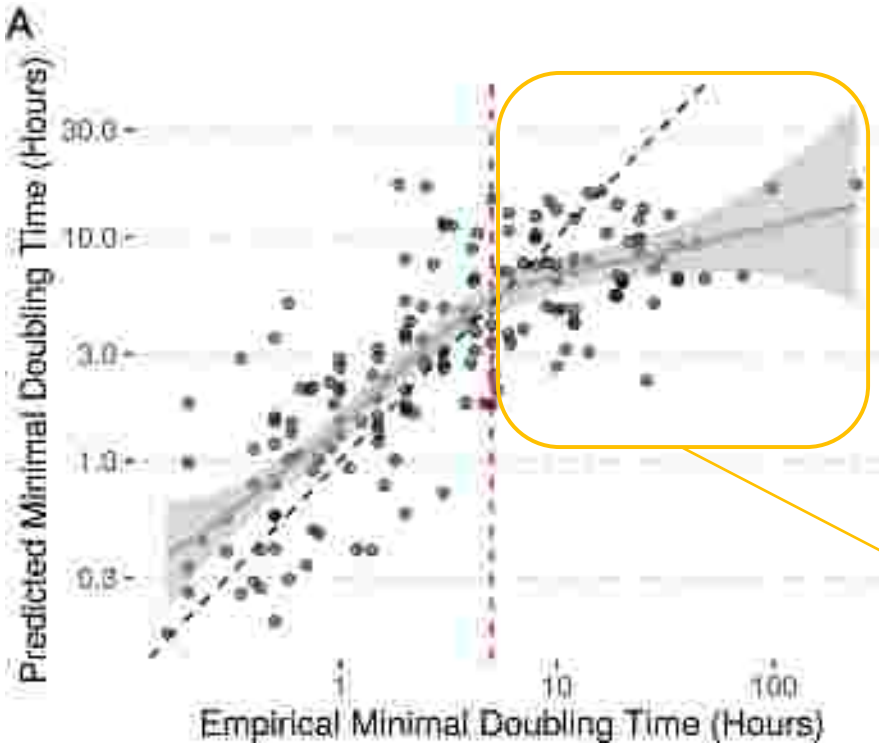
Xia et al., unpublished

Maximal growth rate prediction

海洋科学与工程系
Department of Ocean Science and Engineering

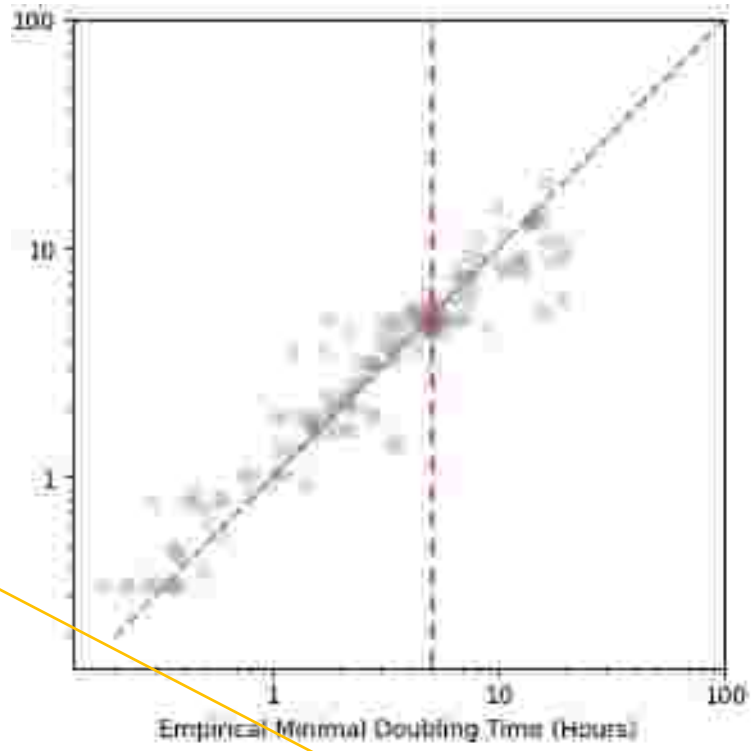
Several machine learning models were trained, and evaluated with the gRodon model. Our XGBoost models performed pretty well.

gRodon (2021), R^2 = 0.63

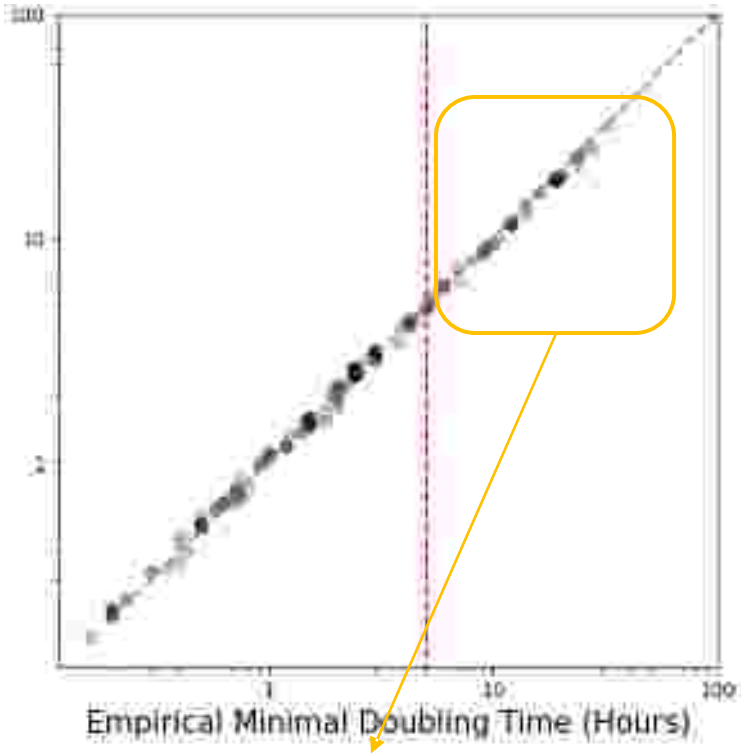


https://doi.org/10.1073/pnas.2016810118

Our best MLP model, R^2 = 0.76



Our best XGBoost model, R^2 = 0.99



Our best model performs better, especially in the oligotrophiles (minimal doubling time > 5 hr).

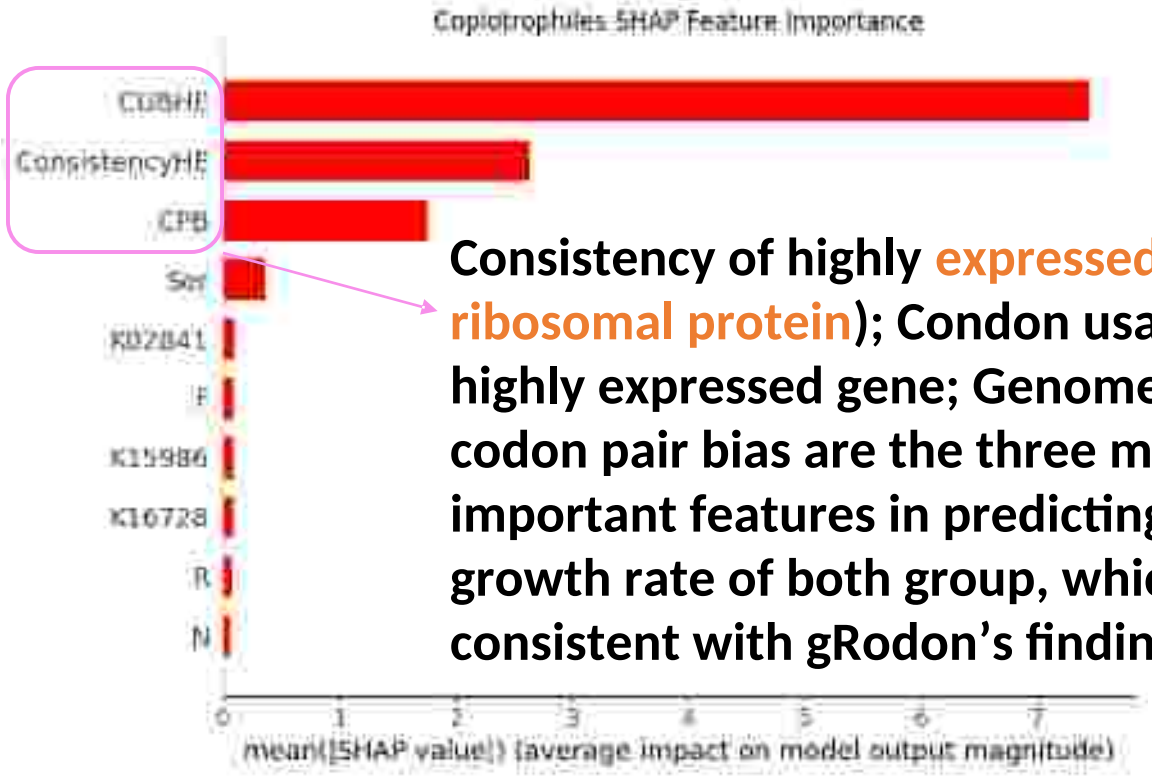
Xia et al., unpublished

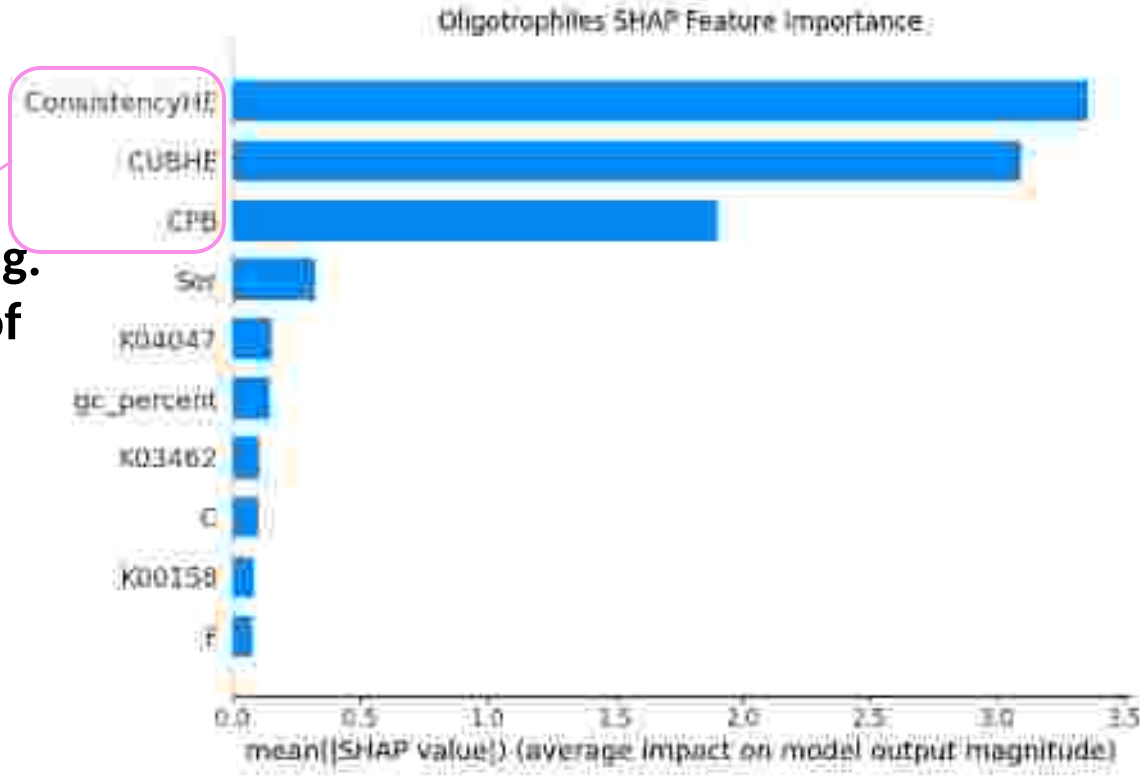
Model interpretabilities

海洋科学与工程系
Department of Ocean Science and Engineering

SHAP values illustrate the importances of features on models' performances.

We compared the important features in copiotrophiles and oligotrophiles based on the best XGBoost model's performance.





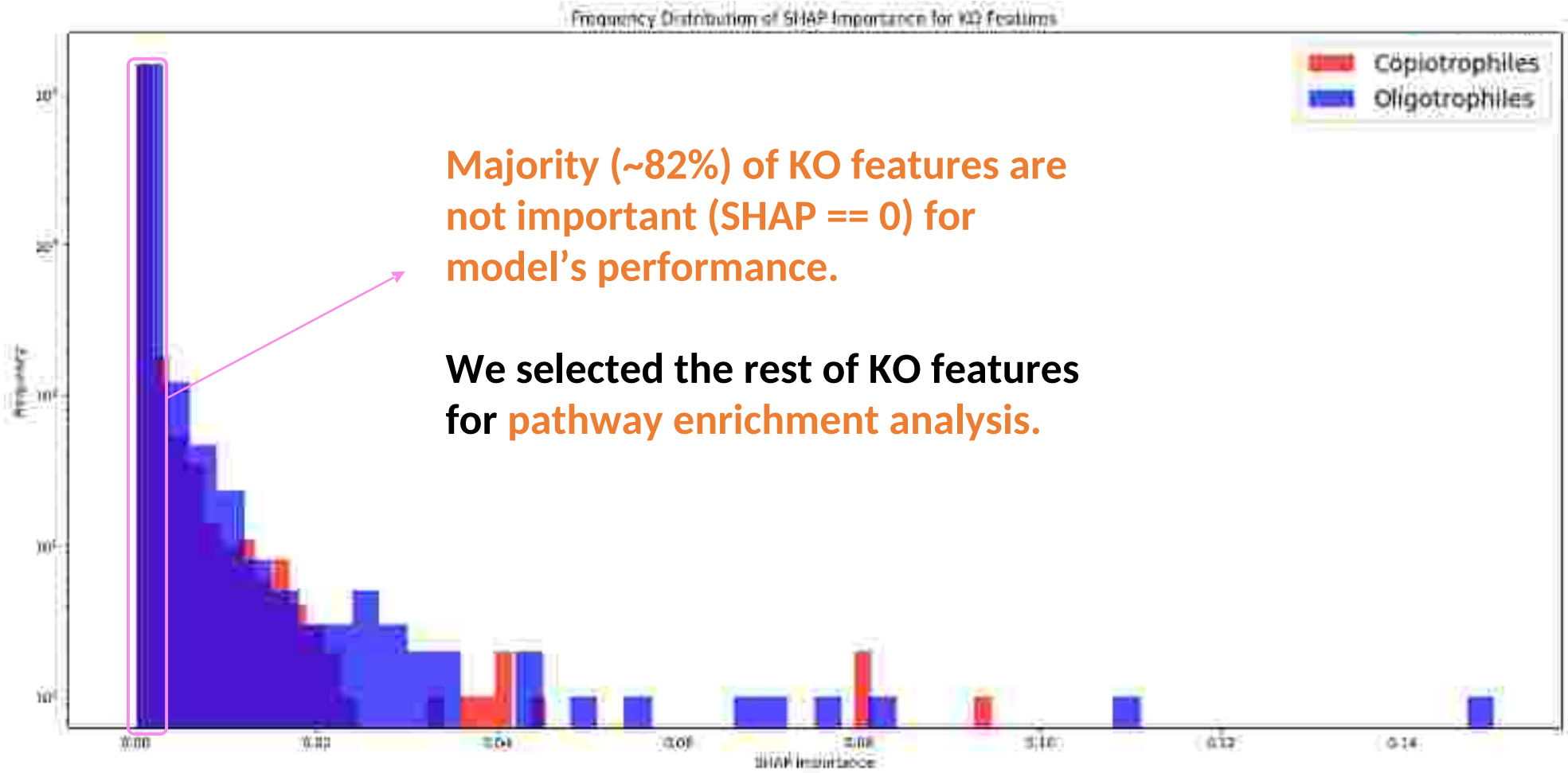
Consistency of highly expressed gene (e.g. ribosomal protein); Condon usage bias of highly expressed gene; Genome-wide codon pair bias are the three most important features in predicting max growth rate of both group, which is consistent with gRodon's findings.

Xia et al., unpublished

Model interpretabilities

海洋科学与工程系
Department of Ocean Science and Engineering

Then we selected important protein annotation features (KO) for biological pathway enrichment analysis.



Majority (~82%) of KO features are not important (SHAP == 0) for model's performance.

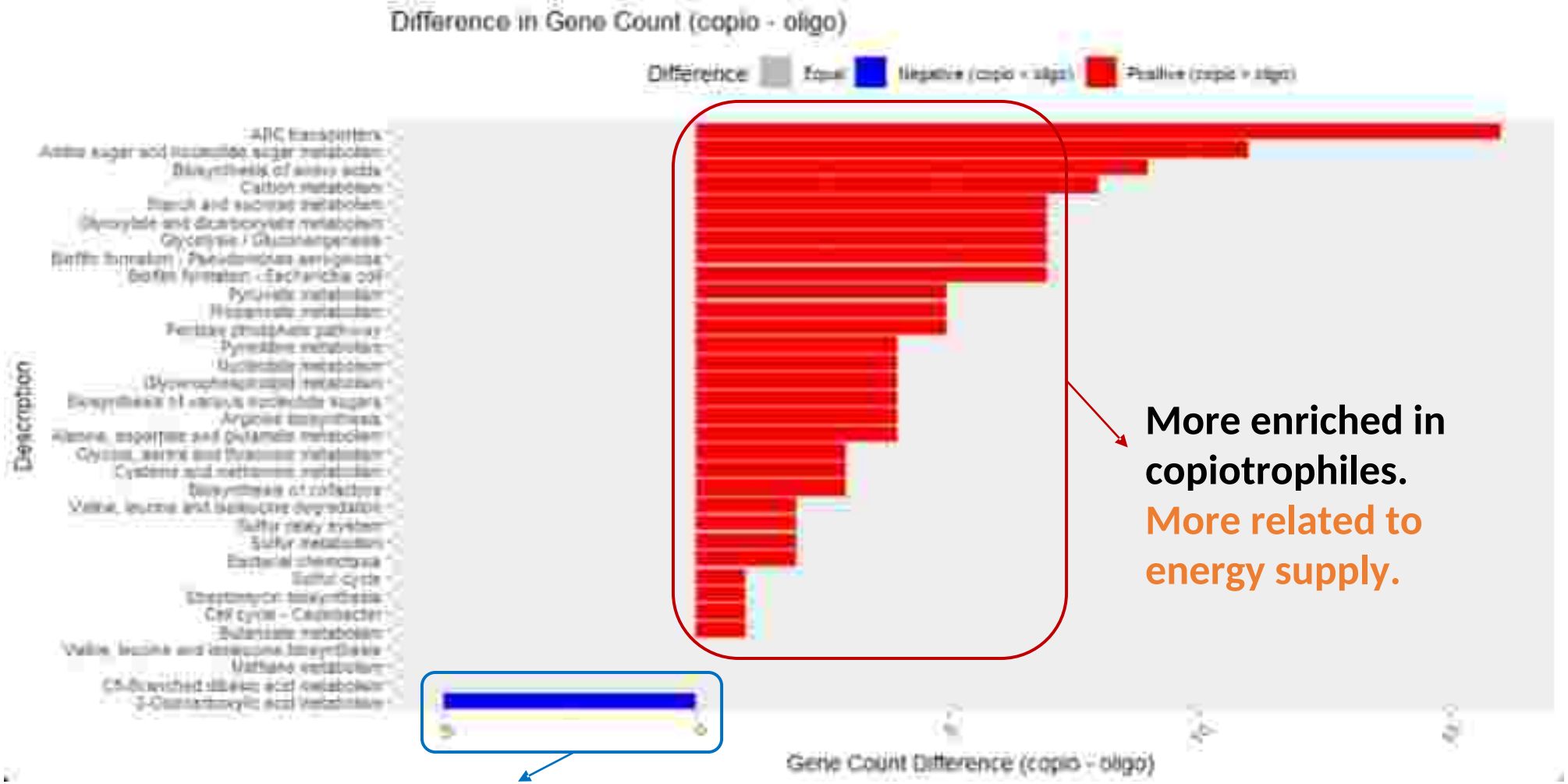
We selected the rest of KO features for pathway enrichment analysis.

Xia et al., unpublished

Model interpretabilities

海洋科学与工程系
Department of Ocean Science and Engineering

Many pathways were enriched in both copiotrophiles and oligotrophiles, although their gene counts in each enriched pathway were different.



More enriched in copiotrophiles.
More related to energy supply.

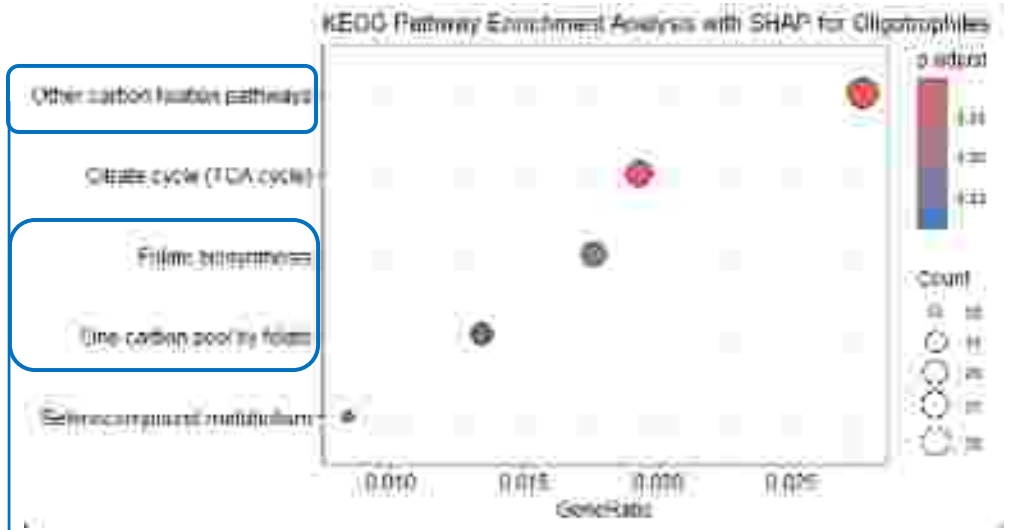
More enriched in oligotrophiles, this pathway is in archaea.

Xia et al., unpublished

Model interpretabilities

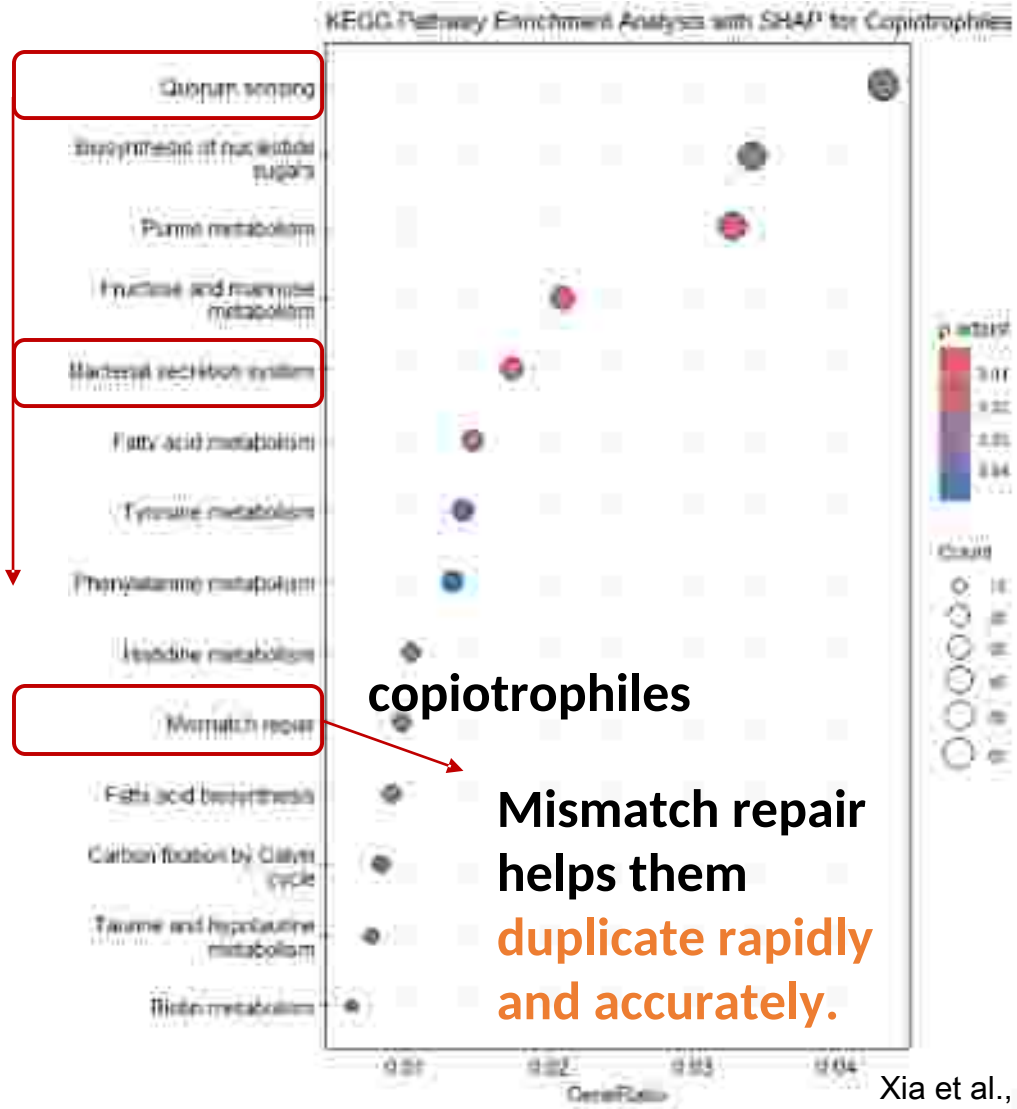
海洋科学与工程系
Department of Ocean Science and Engineering

While some pathways were only enriched in copiotrophiles or oligotrophiles, which were important for max growth rate prediction.



Only enriched in oligotrophiles, indicating their potentials to use multiple (complex) carbon sources.

Only enriched in copiotrophiles, quorum sensing and bacterial secretion system help them to defense other microbes in productive environment.



copiotrophiles

Mismatch repair helps them duplicate rapidly and accurately.

Xia et al., unpublished

Summary 4

1. Machine learning (XGBoost) and deep learning (MLP) models solved the accuracy of CUB-based predictors for slow-growing microorganisms.
2. Different functional genes and metabolic pathways are enriched in oligo copiotrophs, which contribute to the reliable prediction of slow-growing microorganisms.
3. For in situ growth rates and OGT prediction, to be continued ...

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY



Thank you for your attention!

海洋科学与工程系
Department of Ocean Science and Engineering